



Universidad Autónoma de Querétaro
Facultad de Lenguas y Letras
Doctorado en Lingüística

Uso de palabras clave para la detección automática del discurso sexista en prensa y redes sociales

Opción de titulación
Tesis

Que como parte de los requisitos para obtener el Grado de
Doctorado en Lingüística

Presenta:
Héctor Castro Mosqueda

Dirigido por:
Dr. Ignacio Rodríguez Sánchez

Co-dirigido por:
Dr. Antonio Rico-Sulayes

Dr. Ignacio Rodríguez Sánchez
Presidente

Firma

Dr. Antonio Rico-Sulayes
Secretario

Firma

Dra. Eva Patricia Velásquez Upegui
Vocal

Firma

Dra. Mónica Sanaphre Villanueva
Suplente

Firma

Dra. Ester Bautista Botello
Suplente

Firma

Dra. Adelina Velázquez Herrera
Directora Facultad de Lenguas y Letras

Dra. Ma. Guadalupe Flavia Loarca Piña
Directora de Investigación y Posgrado

Centro Universitario
Querétaro, Qro.
Enero de 2023
México



Dirección General de Bibliotecas y Servicios Digitales de
Información



Uso de palabras clave para la detección automática del
discurso sexista en prensa y redes sociales

por

Héctor Castro Mosqueda

se distribuye bajo una [Licencia Creative Commons
Atribución-NoComercial-SinDerivadas 4.0 Internacional](#).

Clave RI: LLDCC-267072-0223-123

Dedicatoria

Al culminar este proyecto, solo tengo una palabra en mente: ¡Gracias!

Agradezco a Dios por bendecirme con la vida de personas maravillosas que me han apoyado en este camino. Todo el arduo trabajo que he realizado a lo largo de varios años solo ha sido posible gracias al apoyo incondicional de mi esposa, Rosa María, quien me ha brindado su apoyo en todo momento, y a mis hijos, Héctor y María José, cuya paciencia ha sido puesta a prueba en incontables ocasiones.

Además, dedico este trabajo a mis padres, quienes a través de su ejemplo me han enseñado lo que significa tener la fortaleza para enfrentar los desafíos de la vida.

Agradecimientos

Quiero expresar mi más profunda gratitud a aquellos que hicieron posible la culminación de mi proyecto de investigación.

En primer lugar, quiero agradecer al **Dr. Ignacio Rodríguez Sánchez** por su confianza en mí y su paciencia a lo largo de todo el proceso. Su visión y guía me ayudaron a encontrar mi área de investigación y a superar momentos de incertidumbre. Espero haber cumplido con sus expectativas.

También deseo agradecer al **Dr. Antonio Rico Sulayes** por la oportunidad que me brindó y por compartir conmigo sus conocimientos. Trabajar a su lado me permitió tener una perspectiva más amplia y expandir mis áreas de oportunidad.

A la **Dra. Eva Patricia Velásquez Upegui**, quiero agradecerle por sus valiosos consejos y por ayudarme a superar momentos difíciles. Gracias a su apoyo pude ser resiliente y permanecer en el programa de doctorado.

A la **Dra. Mónica Sanaphre Villanueva** por su calidez y por hacerme sentir parte del programa y de la UAQ.

A la **Dra. Ester Bautista Botello** le agradezco sus valiosos comentarios que me permitieron tener mayor certeza sobre la relevancia de mi trabajo de investigación.

Finalmente, agradezco a la Universidad Autónoma de Querétaro por brindarme la oportunidad de enriquecerme en conocimiento y ser parte de su comunidad académica.

Resumen

Esta tesis emplea herramientas de análisis léxico, métodos de procesamiento cuantitativo y procedimientos propios del procesamiento del lenguaje natural para analizar muestras de lenguaje con el fin de identificar, analizar y clasificar la presencia, a veces sutil y a veces flagrante, del sexismo lingüístico en nuestra vida cotidiana.

Esta tesis inicia realizando un análisis de colocaciones para investigar cómo se representan lingüísticamente a las mujeres y a los hombres en la prensa en línea en español. Posteriormente, esta investigación analiza cómo la extracción de palabras clave, una técnica propia de la Lingüística de Corpus, puede emplearse en la obtención de rasgos para mejorar la precisión de los algoritmos en tareas de Clasificación Automática de Texto en el campo del Procesamiento del Lenguaje Natural; en particular, esta tesis se ocupa de extraer palabras clave de un corpus obtenido de las redes sociales para clasificar el lenguaje misógino.

En estas investigaciones, comenzamos basándonos en un corpus de cuarta generación y eventualmente en textos de comunicación mediada por ordenador. Concretamente, el corpus NOW, que contiene 7,200 millones de palabras, fue consultado en un inicio para identificar representaciones lingüísticas de género mediante un análisis de colocaciones adjetivas y verbales; asimismo, después se recopiló un corpus de comentarios en redes sociales de 1,841,385 palabras para extraer palabras clave, las cuales se utilizaron como rasgos para mejorar la precisión de los algoritmos en tareas de clasificación automática.

El análisis de los datos mostró que algunos adjetivos y verbos se relacionan exclusivamente o con mayor intensidad con el lema HOMBRE o con MUJER; los hombres se relacionan con mayor intensidad con adjetivos relacionados con la agudeza mental, la sexualidad y la fertilidad, mientras que las mujeres se relacionan con adjetivos relacionados con la salud, el estado civil y las afiliaciones religiosas. Con respecto a las colocaciones verbales, las mujeres se asocian más fuertemente con verbos relacionados con expresiones de emociones y los hombres con verbos que denotan un comportamiento violento. Posteriormente, en las tareas de clasificación que se realizaron con el corpus extraído de las redes sociales, las palabras clave obtuvieron un 98% de precisión al clasificar los textos y un 92% al clasificar casi 7,500 comentarios. Cuando se eliminaron las palabras clave de la tarea de clasificación y se

utilizaron todas las palabras para llevar a cabo dicha tarea, la precisión descendió en promedio 17 %.

Palabras clave : Lingüística de corpus, clasificación automática de textos, palabras clave

Abstract

This thesis uses lexical analysis tools, quantitative processing methods, and natural language processing procedures to identify, analyze, and classify the presence of linguistic sexism in our daily lives, both subtle and obvious. The experiments explore how collocational analysis can be used to examine the linguistic representation of women and men in Spanish news articles online. Additionally, the research analyzes how keyword extraction, a technique from Corpus Linguistics, can produce features that enhance the accuracy of algorithms in Automatic Text Classification tasks within the Natural Language Processing field. Specifically, this thesis focuses on extracting keywords from a corpus to classify misogynistic language.

To conduct the experiments, fourth-generation corpora and computer-mediated communication sources were utilized for data collection and corpus building. The NOW corpus, which contains 7.2 billion words, was queried initially to uncover gender linguistic representations through an adjectival and verbal collocational analysis. Later, a corpus consisting of 1,841,385 words was compiled to extract keywords used as features to enhance the accuracy of algorithms in automatic classification tasks.

The analysis of the data showed that some adjectives and verbs patterned exclusively or more strongly with the lemmas MAN and WOMAN. Men patterned more strongly with adjectives related to mental sharpness, sexuality, and fertility issues, while women patterned with adjectives related to health issues, marital status, and religious affiliations. Concerning verbal collocations, women collocated more strongly with verbs related to emotional vocal expressions, and men with verbs denoting violent behavior.

In the classification tasks, the extracted keywords achieved a 98% accuracy when classifying the texts and a 92% accuracy when classifying nearly 7,500 comments. Removing the keywords from the classification task and using all words as features led to a 15% drop in accuracy.

Abbreviations and Acronyms

ADESSE	Alternancia de Diátesis y Esquemas Sintático-Semántico
AL	Applied Linguistics
ATC	Automatic Task Classification
BNC	The British National Corpus
CALL	Computer-Assisted Language Learning
CDA	Critical Discourse Analysis
CL	Corpus Linguistics
CMC	Computer-Mediated Communication
CMDA	Computer-Mediated Discourse Analysis
FCDA	Feminist Critical Discourse Analysis
LGBT	Lesbian, Gay, Bisexual, and Transgender
LL	Log-likelihood
MI	Mutual Information
ML	Machine Learning
NLP	Natural Language Processing
NOW	News on the Web
TF-IDF	Term Frequency times Inverse Document Frequency
VSM	Vector Space Model
WEKA	The Waikato Environment for Knowledge Analysis

Table of Contents

<i>Agradecimientos</i>	<i>ii</i>
<i>Resumen</i>	<i>iv</i>
<i>Abstract</i>	<i>vi</i>
1 Introduction	1
1.1 Relevance of study	3
1.2 Map of the thesis	4
2 Research questions and hypothesis	8
2.1 Statement of the problem	8
2.2 Research Questions	9
2.3 Hypothesis	10
3 Theoretical Framework	11
3.1 Corpus Linguistics	11
3.1.1 Collocations	13
3.1.1.1 Mutual Information	15
3.1.2 Keywords Analysis and Keyness	17
3.1.3 Corpus Linguistics and Machine Learning Resources	21
3.1.3.1 NOW Corpus	21
3.1.3.2 Concordancers Generations	22
3.1.3.3 Weka	27
3.1.3.4 Verb database classification and adjective taxonomy.	30
3.1.3.4.1 ADESSE	30
3.1.3.4.2 Adjective Supersenses Classification Framework	31
3.1.3.4.3 Violentómetro	33
3.2 Computers and human language	34
3.2.1 Natural Language Processing	35
3.2.2 Machine Learning	35
3.2.2.1 Automatic Text Classification	38
3.3 Language and the World Wide Web	43
3.3.1 The Web and Corpus Linguistics	43
3.3.2 Social Media and Language Research	46
3.3.2.1 YouTube	49
4 Literature Review	53
4.1 Language and Gender	53
4.1.1 Deficit model	53
4.1.2 Dominance model	54
4.1.3 The difference (cultural) model	57
4.1.4 The “dynamic” or “Social Constructionist” model	59
4.2 Gender and Corpus Linguistics (Corpus studies on Gender)	60

4.3	Language and Machine Learning	70
5	<i>Methodology</i>	85
5.1	Collocational Analysis and The NOW Corpus	86
5.2	YouTube Corpus and Keyword Analysis	92
5.3	Automatic Text Classification	98
6	<i>Results and Analysis</i>	105
6.1	Results of Experiments of the NOW Corpus	106
6.1.1	Adjectives	107
6.1.2	Verbs	118
6.2	Results of Machine Learning Experiments	124
6.2.1	Preliminary Experiment of “Violentómetro”	125
6.2.2	Automatic Text Classification (Video Files)	128
6.2.3	Text Automatic Classification (users’ comments)	131
7	<i>Discussion</i>	135
8	<i>Conclusion</i>	155
	<i>References</i>	158

List of Figures

Figure 1 Sample concordances.	23
Figure 2 Most popular tools used for analyzing corpora (Tribble 2012 in Anthony, 2015).	25
Figure 3 AntConc software tool.	26
Figure 4 WEKA Workbench	28
Figure 5 Weka Explorer interface.	29
Figure 6 IPN's Violentómetro	34
Figure 7 Visual representation of data fields.	36
Figure 8 Document classification process. Taken from (Dalal & Zaveri, 2011).	42
Figure 9 Frequency distribution of gendered titles according to Baker, 2013.	69
Figure 10 Locations where the tweets of the SELA corpus were compiled. Taken from (García-Díaz et al., 2021)	77
Figure 11. Linguistic features with the highest information gain. Taken from (García-Díaz et al., 2021).	82
Figure 12 Corpus NOW interface. Parameters to search the adjectival collocations.	86
Figure 13 Adjectival Collocations of the lemma MUJER (WOMAN)	87
Figure 14 The collocates and their collocations in context.	88
Figure 15 Classification of the adjectives based on the SuperSenses taxonomy.	89
Figure 16 Corpus NOW interface. Parameters to search the verbal collocations.	90
Figure 17 Verbal Collocations of the lemma MUJER (WOMAN).	91
Figure 18 The collocate and its collocations in context.	91
Figure 19 Classification of the verbs based on the ADESSE taxonomy.	92
Figure 20 YouTube video about femicide.	93
Figure 21 YouTube user's comments on femicide.	94
Figure 22 AntConc interface. Parameters set to generate the keyword list.	96
Figure 23 Keyword list obtained from comparing Viogendis and Viomujdis.	97
Figure 24. Classified comments according to their classes.	101
Figure 25 Weka interface showing instances (comments) and attributes (features).	102
Figure 26 Boolean scheme.	102
Figure 27 Constellation network of the collocations of the Behavior category.	108
Figure 28 Constellation network of the collocations of the Body category.	111
Figure 29 Constellation network of the collocations of the Mind category	114
Figure 30. Constellation network of the collocations of the Social category	117
Figure 31 Constellation network of the category of Communication (ADESSE).	119

Figure 32 Constellation network of the category of Life (ADESSE).	121
Figure 33 Constellation network of the category of Competition (ADESSE).	123
Figure 34 Samples of comments being classified.	132

List of Tables

Table 1 Contingency table for keyness calculation. Rayson, P. (2013).	20
Table 2 Top-level classes in ADESSE. Taken from García-Miguel & Albertuz, 2005.	31
Table 3 Change subclasses in ADESSE. Taken from García_Miguel & Albertuz, 2005.	31
Table 4 Supersense taxonomy for adjective classification.....	33
Table 5 Vector space model representation	40
Table 6 Four levels of CMDA (taken from Herring, 2013).....	48
Table 7 Results of the first subtask over the test corpora. Taken from (Canós, 2018).....	72
Table 8 Results of the second subtask over test corpora. Taken from (Canós, 2018).....	73
Table 9 Examples of text for each misogyny category. Taken from Anzovino et al. (2018).	74
Table 10 Accuracy performance for misogynistic language identification. (Anzovino et al., 2018)	75
Table 11 System Results per team in subtask A of the HatEval task in Spanish. Taken from (Plaza- del-Arco et al., 2019)	76
Table 12 MisoCorpus Classification. Taken from (García-Díaz et al., 2021).....	78
Table 13 Accuracy and standard deviation of the baseline model (BoW). (Taken from García-Díaz et al., 2021)	79
Table 14 Accuracy and standard deviation of AWE (Average Word Embedding) for VARW, SELA, DDSS, and MisoCorpus-200 evaluated with a 10 cross-fold validation. (Taken from García- Díaz et al., 2021).....	79
Table 15. Accuracy and standard deviation of LF features for VARW, SELA, DDSS, and MisoCorpus-200 evaluated with a 10 cross-fold validation. (Taken from García-Díaz et al., 2021)	80
Table 16 Accuracy and standard deviation of AWE and LF features for VARW, SELA, DDSS, and MisoCorpus-2020 when evaluated with ten cross-fold validation. Taken from (García-Díaz et al., 2021).....	80
Table 17 Comparison of all the subsets when executed with each feature and when these were combined. Taken from (García-Díaz et al., 2021)	81
Table 18 Comparison of the accuracy between European Spanish and Latin American Spanish when applying AWE, LF and AWE+LF.	82
Table 19 Experiments in this research study	85
Table 20 YouTube Corpora.....	95
Table 21 Comments made by YouTube users.	95
Table 22 Keyword analysis carried out for each corpus.	97
Table 23 Vector Space Model	98

Table 24 Features in each one of the three sub-experiments.	99
Table 25 Sample of VSM with 30 features.....	100
Table 26 Sample of VSM with 242 features.....	100
Table 27 Keywords removed to assess their weight in the text classification.....	103
Table 28 Results obtained for each one of the categories in the Supersenses classification.	106
Table 29 Vector space model representing the frequency of features in two classes. (Violentómetro).....	126
Table 30 Results obtained in the experiment. Features taken from the Violentometro.....	127
Table 31 Vector space model representing the frequency of features in three classes. (Violentómetro).....	128
Table 32 Results obtained in the experiment once Features were adjusted (Violentómetro).	128
Table 33 Vector space model representing the frequency of features in the three classes. (keywords).	129
Table 34 Results obtained in the experiment with keywords as features (30 features).	130
Table 35. Results obtained in the experiment with keywords as features (243 features).	130
Table 36 Results obtained when classifying comments. (1,756 features).....	132
Table 37 Results obtained when the keywords were excluded.	133
Table 38 Adjectival collocation of MAN and WOMAN.....	138
Table 39 Verbal collocations of the lemmas MAN and WOMAN.....	139
Table 40 Accuracy of the best algorithms	141
Table 41 Results of the automatic text classification with the string to word vector filter.....	142
Table 42 Ranking of the attributes (keywords) with higher information gain.	144

1. Introduction

This thesis is founded on two beliefs. Firstly, that academic research should not be isolated from real-world issues. Secondly, that Corpus Linguistics and Computational Linguistics, as the two branches of Linguistics most linked to technology, can provide valuable tools for tackling the problems we face, and increasing our understanding of ourselves as individuals and as a society.

It is evident that violence is one of the most pressing issues in Mexican society today, and it is a phenomenon that permeates every aspect of life. Violence is not limited to physical harm and the countless deaths that have caused immense pain over the last fifteen years; it also numbs and desensitizes society at large. Violence is evident in the psychological and social spheres and is reflected in the attitudes and ideologies of some members of society, who justify and normalize mistreatment and violence. The language that is used to justify violence is perpetuated by everyday people, the media, and societal institutions, and it is not surprising to utilize linguistic knowledge to expose this language, which contributes to the existence of discourses that lead to violence towards certain social groups, particularly women.

This thesis analyzes two linguistic corpora, namely the Spanish New on the Web (NOW) and a corpus of comments posted by social media users (YouTube), to expose the sexist attitudes that permeate language and society. The purpose is to carry out collocational analyses and automatic text classifications of gendered texts and comments using keywords as features. Unlike other linguistic works that focus on Discourse Analysis, this thesis utilizes lexical analysis tools, quantitative processing methods, and natural language processing procedures to identify and analyze the subtle and blatant presence of sexism in everyday language.

This thesis proposes that automatic or semi-automatic quantitative linguistic analyses are crucial for unmasking the everyday violence in language that often goes unnoticed. Micromachismo, which is often perceived as language without malice, becomes a problem of significant social consequences when it is revealed to be a pattern repeated thousands of times in language and multiplied at all levels and in all places.

To achieve the above objective, the thesis starts with a comparison of the most significant and relevant collocations accompanying the lemmas hombre ‘man’ and mujer ‘woman’ in the Latin American press. The patterns of co-occurrence of adjectives and verbs with these lemmas reveal an ideology where women are subjected to objectification, deindividuation, and marginalization. Behind the subtlety of stereotyped language, there is a disregard and disdain that are common currency and can lead to hatred. After conducting a formal collocational analysis of the written press, the thesis proceeds to the second significant analysis, which is a corpus of comments on gendered language posted by social media users on YouTube. This corpus is utilized to carry out automatic text classifications using linguistic gendered features.

The analysis of two distinct corpora reveals a convergence in the study of sexist language and stereotypes. Due to their different natures, these corpora are analyzed from two perspectives originating in separate fields. Edited texts (NOW corpus) are typically analyzed quantitatively in Corpus Linguistics, while unedited texts (YouTube corpus) full of errors and typos are analyzed in Computational Linguistics using a variety of tools and algorithms.

Research on Language and Gender has a lengthy history in linguistics, with studies ranging from attributing linguistic differences between men and women to deficits on the part of women to considering gender as a social construct (Litosseliti, 2014; Coates, 2015; Kendall & Tannen, 2015; Flowerdew & Richardson, 2017, and Weatherall, 2005). While extensive research in Corpus Linguistics has addressed how men and women are represented in different corpora in the past decade (Baker, 2010/2013; Caldas-Coulthard & Moon, 2010; Moon, 2014; McEnery & Baker, 2015), this field has primarily focused on English and has been relatively neglected in the Spanish language. This thesis aims to expand research in this area by utilizing tools and techniques from Corpus Linguistics (CL) to explore the topic further.

Moreover, this thesis uses tools and techniques from CL to enhance automatic text classification (ATC) tasks in the field of Natural Language Processing (NLP). ATC is a Machine Learning (ML) task, a branch of Artificial Intelligence in NLP. It assigns a document to a class in a set of categories based on its content and extracted features. ATC

utilizes features like token n-gram, character n-gram, bag-of-PoS, embedding, morphology, and pragmatic linguistic features to evaluate models. Machine Learning extracts knowledge from data, allowing systems to learn from it. ATC has numerous practical applications, such as content management, spam filtering, opinion and sentiment analysis, improving search results in search engines, ranking or grouping of results, online reviews of products, text mining, and information retrieval (Sebastiani, 2005; Dalal & Zaveri, 2011; and Vajjala et al., 2020). While hate speech investigations are a crucial aspect of ATC research, this thesis also employs data on topics like ethnicity, immigration, gender identities, and misogynistic language, among others, to conduct classification experiments.

The critical point of this thesis is how the keyword extraction technique, a typical Corpus Linguistics technique, significantly improves Automatic Text Classification techniques accuracy when applied in Computational Linguistics. This technique helps identify gendered comments.

1.1 Relevance of the study

- 1) There are several reasons why this research is relevant. First and foremost, this thesis addresses the representation of women and men in online Spanish news, a topic that has received extensive research attention in English, but not in Spanish. By conducting a collocational analysis, this research study has amplified linguistic patterns, providing a foundation for researchers to engage in more qualitative discussions. This is where the significance of the collocational analysis lies. A collocational analysis, particularly when conducted on a specific corpus, can expand the possibilities for research findings in fields such as linguistics, gender studies, and discourse analysis, among others.
- 2) Feature selection is one of the main challenges in automatic text classification (ATC). With thousands of words and other linguistic items to choose from, selecting efficient features can be difficult. Many of these features may be non-informative and yield conflicting or poor results. In the natural language processing field, there are various feature selection methods grouped into four major categories: filter model, wrapper model, embedded model, and hybrid model (Deng et al., 2019; Yang et al., 2012; Liu & Yu, 2005). Common methods for measuring the goodness of features in ATC tasks

include bag-of-word, TF-IDF, and information gain (IF). The use of keywords as a feature selection method in ATC tasks is what I propose in this thesis. Its relevance lies in the fact that keywords yielded favorable results across different ATC tasks. My hope is that experienced scholars in natural language processing will explore the process of obtaining keywords in corpus linguistics and extrapolate this idea to inform automatic text classification.

- 3) During the process of engineering the automatic text classification experiments, a corpus was built; this corpus contains instances of misogynistic language in Spanish and was used to run the ATC tasks. One of the most important disadvantages when using corpora is that these are too generic, may not be suitable for certain purposes, or users may be limited to run searches. Besides that, when one talks about corpora, scholars now tend to think about a corpus as containing hundreds of millions of words; however, the construction of such corpora is expensive and time-consuming. Given this, the relevance of this thesis concerning corpus building lies in the fact that a corpus was constructed to address a specific topic and need. What I attempt to convey is that researchers may endeavor in the construction of specific or specialized corpora which may allow them to expand their research areas and findings.
- 4) In conclusion, my hope is that this thesis contributes to promoting research that advocates for a more comprehensive approach to linguistics. By utilizing interdisciplinary areas of research, linguists can expand their research horizons and deepen their understanding of linguistics. I believe that this thesis can inspire researchers to think beyond the traditional boundaries of linguistics and embrace a multidisciplinary perspective to enrich their research.

1.2 Map of the thesis

This thesis is organized into eight distinct chapters: introduction, theoretical framework, literature review, methodology, results and analysis, discussion, and conclusion.

1. Introduction

In this chapter, I offer a comprehensive overview of this thesis, outlining how I utilized a combination of Corpus Linguistics and Computational Linguistics (Natural Language Processing) tools and techniques to conduct my research. Furthermore, I emphasize the significance of this thesis by underscoring the interdependent relationship between CL and NLP, and highlight the benefits of interdisciplinary-oriented research. To aid the readers in navigating this thesis, I present a detailed map that serves as a useful guide to locate specific information.

2. Research Questions and Hypothesis

In this chapter, I provide an overview of my research by outlining the main topics and ideas that I intend to investigate. I present the key questions that I will answer in the Results section, and conclude by describing the hypotheses that I will attempt to prove.

3. Theoretical Framework

In this chapter, I provide an introduction to major concepts, tools, statistical measures, and techniques utilized in both CL and NLP. Given that some readers may not be familiar with the disciplines presented in this thesis, I describe concepts such as concordances, keywords, collocations, keyness, and mutual information, among others. Additionally, I provide an overview of the different classification frameworks employed in the collocational analysis and highlight how the worldwide web has contributed to CL in corpus

building. Finally, I discuss significant concepts and techniques in Machine Learning, and provide insight into how ATC tasks are executed.

4. Literature Review

The fourth chapter consists of three sections. The first section provides an overview of relevant studies that examine the relationship between gender and language. Given that this thesis revolves around the symbiotic connection between how women and men use and are portrayed in language, I describe early studies that attempt to provide an explanation for gender-based linguistic differences.

In the second section, I delve into studies that utilize corpora as a primary instrument to explore how gender is represented in language. Unlike the studies reviewed in the first section that rely on a limited amount of data and sometimes on the researchers' opinions, the research studies presented in the second part employ a more substantial amount of data to identify linguistic patterns that illuminate how women and men are depicted.

Finally, the last section of this chapter provides an overview of automatic text classification research studies that focus on hate speech. These studies examine the effectiveness of different features to determine which algorithms generate the most accurate classifications.

5. Methodology

In this chapter, I provide a detailed description of the processes involved in conducting the collocational analysis in the NOW corpus, including the categorization of collocates according to Supersenses and ADESSE classifications. Additionally, I outline the process of

constructing the YouTube corpus, including the intricacies involved in its development. I discuss the methods used to obtain keywords and how they were subsequently used in ATC tasks. Finally, I explore how keywords were utilized as features in various tasks to evaluate their effectiveness in different classification contexts.

6. Results and Analysis

In this chapter, I present and analyze the results of the adjectival and verbal collocational analyses. Additionally, I provide an overview of the outcomes of various automatic text classification tasks that utilize keywords as features.

7. Discussion

In this section, I integrate the findings from the ATC tasks and collocational analysis. I also address the research questions that have guided this thesis and highlight the importance and benefits of building bridges not only between different fields within linguistics but also with other disciplines.

8. Conclusion

In the final chapter, I summarize the key findings of this thesis. I also provide a critical evaluation of the limitations of this research study and suggest avenues for future research.

1 Research questions and hypothesis

In this chapter, I will delve into the statement of the problem, exploring the challenges and opportunities associated with using tools and techniques from multiple disciplines to investigate a specific topic. Additionally, I will examine the complexities that arise when working with internet-sourced data. Finally, I will introduce the research questions and hypotheses that guide this study.

1.1 Statement of the problem

In the introduction of this thesis, I contend that quantitative studies of linguistics often suffer from a lack of dialogue between different subfields. This issue is not unique to linguistics, but rather pervasive throughout various disciplines and fields of knowledge. The analytical tradition in the West has resulted in compartmentalization and (hyper)specialization, leading to the approach of breaking problems into smaller parts that can be solved more effectively. However, this reductionist approach ignores potentially complementary approaches from different subdisciplines. In this thesis, I aim to address this issue by linking corpus linguistics and automatic text classification tasks, combining tools and techniques to address a specific problem.

Corpora, twenty years ago, were compiled, designed, processed, and analyzed manually or semi-automatically. Today, fourth-generation corpora are built of texts downloaded from the internet based on criteria as varied as their origin or seeds, with texts assessed for relevance and reliability. However, typed texts that are part of computer-mediated communication (CMC) sources pose unique challenges to linguistic analysis. These texts are transient and usually far from the standard variety of the language, replete with dialectal instances, slang, and references that quickly expire, making analysis and classification difficult. Additionally, these texts are typically full of errors of all kinds, including digitization, orthographic, and semantic.

A prominent example of these challenges is former US President Donald Trump's tweet in 2017, "Despite the constant negative press coffee," where the word "covfefe" remains a mystery. This example illustrates the precariousness of language used in social networks and the impossibility of editing it. Another example in Spanish is the common confusion between the bigram "a ver" and its homophone "haber." However, in a large corpus, all possible confusions and errors (e.g., aver, aber, a ber, ha ver, ha ber, haver) will undoubtedly appear. This problem is compounded by the shallow orthography of Spanish, where Grapheme-Phoneme correspondence is highly regular. Correcting texts typed in social networks is virtually impossible, given that it is impossible to know the speaker's intent, the precise error, or if there is a practical and reliable way to clean up the text. Without a minimum of errors, it is impossible to tag or analyze collocations, colligations, or n-grams.

Despite these challenges, social networks remain invaluable tools for determining public opinion, points of view, social engagement, and the world around us. While the quality of texts and corpora depends on the care with which they are written, the knowledge of the writer, and the revisions to which they are subjected, social networks pose unique challenges to linguistic analysis. Thus, while the use of corpora based on social networks may be limited, social networks can provide essential insights into what society is interested in and how people engage with their environment.

1.2 Research Questions

This thesis aims to bridge the fields of Corpus linguistics and Natural Language Processing, specifically Machine Learning. The research study aims to evaluate the practicality of utilizing keyword extraction, a Corpus Linguistics technique, as a feature selection method to enhance automatic text classification tasks. A crucial aspect of this study involves the creation of a corpus, which was utilized to assess whether keywords could serve as features in ATC tasks. Additionally, gender and language are critical topics in this research study, given the focus on classifying misogynistic language. In light of these goals, the research questions that guide this study are as follows:

1. How can corpora built from online social networks help reveal gender representations?
2. To what extent is keyword analysis an effective feature selection method in machine learning, and how can its efficiency be measured?
3. What are the differences between traditional corpora and those built from online resources?
4. What are the possibilities and complexities involved in constructing a corpus from the web?

1.3 Hypothesis

From the research questions above, the following hypothesis may be derived:

1. Data obtained from social networks can provide unique insights that are not typically available through traditional data collection methods, allowing marginalized communities to have a platform to express their opinions.
2. Keyword analysis, as applied in corpus linguistics, can serve as a feasible feature selection method for those without technical expertise in traditional machine learning feature selection methods.
3. As keywords reflect the main focus of a text, their use in algorithms enhances or at least maintains the accuracy of ATC tasks.
4. Differences between traditional and online CMC corpora can be attributed to differences in the corpus nature, sample selection, and content quality.
5. Collecting linguistic data from the web provides access to a wealth of specialized, current, and detailed material on various topics

2 Theoretical Framework

In this research study, I draw on a range of tools and techniques from both linguistics and related disciplines such as Natural Language Processing (NLP) and Machine Learning. This chapter is organized into three main sections, each of which comprises sub-sections that delve into important topics related to this research.

In the first section, I describe the various linguistic analyses and statistical metrics employed in the experiments carried out in this study. I also discuss the resources used, which are drawn from both the Corpus Linguistics and Machine Learning fields. Furthermore, I outline the verbal and adjectival taxonomies used in the data analysis process.

The second section provides a general overview of the Natural Language Processing field, with a focus on Machine Learning as an approach to solving NLP tasks. This section also outlines the procedure involved in text classification tasks, which is central to this research project.

The third and final section of this chapter elaborates on the use of the World Wide Web to generate linguistic data in the form of corpora, as well as the use of computer-mediated communication (CMC) outlets like YouTube and Twitter to conduct research on various linguistic areas. It is important to note that while I do not provide an in-depth description of all these fields, I provide enough information to help readers understand the multidisciplinary nature of this study.

Corpus Linguistics

The present research study draws heavily on Corpus Linguistics (CL), a field that offers a set of procedures and methods for studying language use. In the following pages, we will define and elaborate on several key constructs in this area that have informed our project. According to McEnery and Hardie (2011), while some procedures in CL are still developing, others, such as concordancing, are well-established. CL involves the empirical investigation of language variation and use, utilizing electronic corpora and computer tools to examine "real-

life" language use (Evert, 2008). Biber and Reppen (2015) suggest that this approach yields research findings with greater generalizability and validity than other methods would permit. While CL is primarily considered a methodological perspective for researching language phenomena, some scholars view it as a theoretical approach. Tognini-Bonelli (2001) argues that:

Corpus work can be seen as an empirical approach in that, like all types of scientific inquiry, the starting point is actual authentic data. The procedure to describe the data that makes use of a corpus is therefore inductive in that it is statements of a theoretical nature about the language or the culture which are arrived at from observations of the actual instances. The observation of language facts leads to the formulation of a hypothesis to account for these facts; this, in turn, leads to a generalization based on the evidence of the repeated patterns in the concordance; the last step is the unification of these observations in a theoretical statement. (p. 2)

Tognini-Bonelli (2001) argues that Corpus Linguistics (CL) provides a contextual theory of meaning, which has led to the development of new theories of language. Similarly, Teubert (2005) characterizes CL as an empirical field that focuses on the study of authentic language data. While the methods used to analyze corpus data vary, the corpus itself, not CL as a field, is considered by some to be theory. McEnery and Hardie (2011) dispute the idea of CL as a theory and maintain that CL is a methodological perspective. Nevertheless, many scholars agree that CL combines the activities of data gathering and theorizing, leading to a qualitative change in our understanding of language (Halliday, 1993).

The terms corpus-based and corpus-driven linguistics illustrate the dichotomy between using corpus data to validate or refine theories (corpus-based) versus using corpus data as a source of hypotheses (corpus-driven) about language (McEnery & Hardie, 2011). Within CL, the corpus is the primary source of data, and it is a collection of spoken or written texts that meet specific design criteria based on their intended purpose and scope (Weisser, 2016). Typically, a corpus is a finite body of text sampled to represent a particular language variety that can be

stored and manipulated using computers (McEnery & Wilson, 2001). Corpora help researchers identify elements and structural patterns in language use and map out language systems (Kennedy, 2014). Although electronic text collections are the norm, the nature of the corpus may vary depending on its intended use. Corpora provide data for various types of linguistic analysis, which can be either qualitative or quantitative. Collocational studies are a traditional type of linguistic analysis within CL, and since this research project includes a collocational analysis, the following paragraphs will delve deeper into this type of analysis.

2.1.1 Collocations

The search for collocations of the lemmas “hombre” and “mujer” in the NOW corpus is the foundation of this research project, and it has led to other experiments in this study. When examining collocations, it is important to distinguish between the “node” and the “collocates.” The former refers to the word being analyzed, while the latter refers to words that occur in the proximity of the node (Sinclair, 1991). Before delving into technical definitions of collocations, Hunston (2002) asserts that “Collocations are the tendency of words to co-occur in a biased manner; for instance, the word *toys* frequently co-occur with *children* rather than with “women” or “men” (p. 68). She uses this example to provide a logical explanation for this co-occurrence, based on the fact that toys generally belong to children rather than adults. Evert (2008) draws on the work of Firth (1957) to argue that the meaning of a word (the node) can be characterized to some extent by its most typical collocates. Collocations can be identified via their frequency or, more commonly, using statistical measures called association measures (Brezina, 2018).

Firth (1957) described collocation as words that frequently occur with other words, but this definition is general and limited because it only relies on simple frequency counts of co-occurrences. McEnery and Hardie (2011) argue that collocations represent the idea that significant aspects of the meaning of a word (or another linguistic unit) are not contained within the word itself, considered in isolation, but rather subsist in the characteristic associations that the word has with other words or structures that it frequently co-occurs with

(p. 123). Firth's work emphasizes the importance of frequency in determining the statistical significance of collocations, but this view has been heavily criticized. Two techniques have been employed to determine the significance of a collocation: a non-statistical (collocation-via concordance) and a statistical one (collocation-via significance). In the former technique, the researcher intuitively scans the concordance lines and identifies the items and patterns that re-occur in the vicinity of the node; in this technique, researchers identify the frequency count but do not execute any statistical testing of the important collocations they find. In the latter technique, the statistical one, "the frequency of each word within the window of the text defined by the span around the node word is compared against its frequency in the rest of the corpus, and if the difference between the frequencies is sufficiently great, the word being examined is identified as a collocate of the node word" (McEnery & Hardie, 2011).

As we have seen, there are competing definitions of collocations in different fields such as Phraseology, Computational Linguistics, Discourse Analysis, and Lexicography. Bartsch (2004) establishes the following criteria, based on both a quantitative and qualitative approach to collocations, for empirical studies that focus on computer-aided extraction of collocation candidates from the corpus:

- 1) Within a span of 3:3 (or 5:5) words to the left and right of a node word (the search word).
- 2) Two (or more) words that co-occur recurrently with the node word.
- 3) And whose frequency of co-occurrence can be said to be statistically significant according to at least one of the three statistical algorithms employed (the threshold values are $MI \geq 3$ and $t\text{-score} \geq 2,576$ for 95% certainty and a significantly high chi-square rating).

After meeting the initial requirements, the collocates undergo a secondary qualitative assessment based on the following criterion:

- 4) must be in direct syntactic relation with each other, and
- 5) display either lexically and/or pragmatically constrained lexical selection, or
- 6) have an element of semantic opacity such that the meaning of the collocation cannot be said to be reducible as a function of the meanings of the constituents.

(p. 76-77)

The collocations are considered to be constrained lexically and/or pragmatically based on specific criteria, namely the recurrent co-occurrence of at least two lexical items in direct syntactic relation to each other (Bartsch, 2004). It is worth noting that empirical studies established criteria are the primary focus of Corpus Linguistics (CL), as is the case with this research study.

In a more recent study, Brezina, McEnery, and Watman (2015) outline several criteria for identifying collocations. Firstly, they revisit three traditionally accepted criteria in academia, namely distance, frequency, and exclusivity. Secondly, citing Gries (2013), Brezina et al. (2015) add three more criteria, namely directionality, dispersion, and type-token distribution. Finally, they argue for considering connectivity between individual collocates as the seventh criterion, stating that the collocates of words do not occur in isolation but are part of a complex network of semantic relationships that reveal the text or corpus's meaning and semantic structure.

Collocations are commonly associated with a lexical relationship between individual words, but they can also occur between words and grammatical categories or markers. These types of associations are referred to as "colligations" (see Xiao, 2015) and analyzed through collostructional analysis (Stefanowitsch & Gries, 2003). However, this research study did not conduct colligation analysis.

In the preceding paragraphs, I described how collocations are defined, identified, and operationalized, emphasizing that statistical significance is one way to identify collocations, as was done in this research study. The following section will describe the Mutual Information (MI) metric used to identify the lemmas' collocations.

2.1.1.1 Mutual Information

The identification of collocations involves various approaches, each with their own strengths and limitations. Statistical significance is a crucial factor in identifying collocations, and researchers often use association measures such as mutual information (MI), t-scores, and log-likelihood tests. These association measures are categorized into two groups: measures

of effect size (MI and MIk) and measures of significance (z-score, t-score). Effect size measures determine the strength of the relationship between words by comparing observed occurrences with expected frequencies. On the other hand, significance measures aim to assess the evidence for a positive association between words, regardless of the corpus size. MI is an information-theoretic measure that determines the extent to which the occurrences of one word determine the occurrence of another word, and vice versa. It is based on the mutual information concept from Information Theory and is widely used in natural language processing research (Manning & Schütze, 1999; Evert, 2008).

In the following paragraphs, I will explain MI in detail as it was the statistical measure that was most relevant to this project. Essentially, MI quantifies the amount of information one random variable provides about another. As Hunston (2002) aptly states, MI is a statistical measure that gauges the strength of association between two words in a corpus based on their independent relative frequency (p. 72). The MI equation is given as follows,

$$MI = \frac{\text{Log}(F_{n,cN}/F_n F_c S)}{\text{Log}2}$$

where N represents the total number of words, F_(n) is the frequency count of the node, F_(c) denotes the frequency of the collocate, F_(n,c) is the frequency of the node and collocate co-occurring within a specified span, S is the size of the corpus, and Log2 is a constant approximately equal to 0.301. An MI score greater than 3 is considered statistically significant.

The MI measurement determines the amount of information that one word provides about the likely occurrence of another (Clear, 1993). To understand this concept better, let's consider an example. Suppose the word "man" appears ten times in a corpus of 10 million words, resulting in a probability of 0.000001. If we observe the word "clever" five times in the same corpus, and in each instance, "man" follows "clever," the probability of seeing "man" increases to 0.5. Hence, the appearance of "clever" significantly enhances the chances of finding "man" nearby. Conversely, if we observe the word "man," we can expect to find "clever" nearby. Clear (1993) notes that the MI value averages the association in both directions.

- a) MI is a measure of the strength of collocations whereas t-score is a measure of certainty of collocations.
- b) The value of an MI-score is not particularly dependent on the size of the corpus whereas the t-score is; this is due to the amount of evidence is taken into account; the larger the corpus is, the more significant a large number of co-occurrences is.
- c) Based on the above, MI scores can be compared across corpora, even if the corpora are different sizes, but absolute t-scores cannot be compared across corpora (though it is reasonable to compare t-scores rankings) because the size of the corpus will affect the t-scores.
- d) Top collocates from a point of view of MI-score tend to give information about its lexical behavior, but particularly about the more idiomatic (fixed) co-occurrences. In contrast, from a t-score point of view, top collocates tend to give information about the grammatical information of a word.
- e) The collocates with the highest t-score tend to be frequent words (whether or not they are grammatical words) that collocate with a variety of items. The collocates with the highest MI scores tend to be less frequent words with restricted collocations.

As mentioned earlier, several association measures are available, and the choice of the best measure depends on the aspects we intend to highlight. For this research, we opted to use the MI measure because it underscores the rarity and exclusivity of collocation relationships. This feature favors collocates that almost exclusively appear in the company of the node, even if it is only once or twice in the entire corpus (Brezina, 2018, p. 71). In this section, we have demonstrated how the MI measure was applied to analyze the collocates of the lemmas man ‘hombre’ and woma ‘mujer’ in the NOW corpus. In the next section, we will discuss a keyword analysis that was conducted as part of the second stage of this research study.

2.1.2 Keywords Analysis and Keyness

In the second stage of this research study, the focus shifted to identifying keywords in three distinct corpora. In this section, I will discuss the concepts of keyword and keyness and their

significance in corpus research. Keyword analysis is widely used across Linguistics and Applied Linguistics (AL), ranging from genre analysis to critical studies, for a variety of purposes, including general genre characterization and identification of text-specific ideological issues (Pojanapunya & Todd, 2018). The goal of identifying keywords in the three corpora in this research study was to use them to conduct automatic text classification experiments, thereby improving the accuracy of algorithms.

Before delving into the concept of keyness, I will provide a brief overview of the keyword concept, as these two constructs are interrelated. According to Stubbs (2010), there are three different definitions (senses) of keywords derived from distinct academic traditions, which are only marginally compatible. The first sense defines keywords as a "focal point around which entire cultural domains are organized," and it is explicitly cultural. However, this view has been heavily criticized as there is no explanation of how these words relate to any theory of how the language's vocabulary is organized. The second sense is statistical, and this research study adheres to this view. Statistical keywords are words that are significantly more frequent in a sample text than would be expected, given their frequency in a large general reference corpus (Stubbs, 2010). This approach provides an empirical discovery method based on frequency and distribution. The third sense focuses on discovering speech acts and cultural schemas through culturally significant units of meaning, which may be overlooked by the introspective data used in speech act theory.

It is essential to note that in this research project, keywords should not be interpreted as words that are "key" because they may have a particular cultural, social, or political significance. Instead, keywords were obtained using a statistical process. In the next section, I will elaborate on the concept of keyness and how it can be used to identify significant keywords in the corpora analyzed in this study.

Culpeper (2009) defines style markers as words whose frequency significantly differs from their frequency in a norm, which precisely corresponds to the concept of keywords. Both style markers and keywords are based on the notion of repetition, but not all repetition is considered, only that which statistically deviates from the pattern formed by that item in another context (p. 4). According to Baker (2004), Scott (1998) identified three types of

keywords: proper nouns, keywords (words that indicate the "aboutness" of a particular text and are recognizable by humans as key), and high-frequency words, such as "because," "shall," or "already," which may indicate style rather than aboutness. Keyword analysis is a popular technique in corpus linguistics and can be used for various purposes, such as providing a descriptive account of different genres or identifying traces of discourse within language (Baker, 2004).

To identify significant differences in keywords between two corpora, a keyness statistical measure is used. Keyness refers to the quality that words may have in a given text or set of texts, indicating their importance in reflecting the overall topic of the text and avoiding trivial or insignificant details. Scott and Tribble (2006) argue that a keyness metric is useful in identifying differences and similarities in keyword analysis.

Previous studies on keyword analysis have employed various statistical measures, such as log-likelihood (LL) or chi-square statistics. However, as Pojanapunya and Todd (2018) note, there is no clear best statistical practice for identifying keywords, and different statistics may be more or less appropriate for different purposes. It is important to select an appropriate metric that goes beyond statistical significance tests, as this is crucial for the accuracy and validity of the analysis.

Previous research has examined the use of effect size and significance test statistics in identifying keywords and found that the former is more suitable for critical research, while the latter is better for genre-oriented research (Pojanapunya & Todd, 2018, p. 160). Effect size statistics indicate the magnitude of an observed finding, whereas statistical significance tests show the level of confidence we can have in the difference observed. Gabrielatos (2018) recommends using effect size statistics to establish keyness, but also suggests using a statistically significant metric as a supplement when analyzing differences in keywords across corpora.

The process of calculating keywords involves three stages as summarized by Rayson (2013). In the first stage, a word frequency list is generated for each of the two texts being compared. This list includes the different forms of each word (type) and the number of times they occur (tokens), as well as the total number of words in each text.

The second stage involves comparing the two frequency lists using a selected keyness statistic measure, which assesses the relative frequency of each word in the two texts. The larger the difference in relative frequencies, the larger the keyness value.

In the third and final stage, the words are sorted in order of their keyness value. Words with higher keyness are considered more relevant as they indicate what occurs in the first text.

To represent this process mathematically, the table below shows how each frequency is obtained:

	Corpus 1	Corpus 2	Total
Frequency of a word	a	b	$a + b$
Frequency of other words	$c - a$	$d - b$	$c + d - a - b$
Total	c	d	$c + d$

Table 1. Contingency table for keyness calculation. Rayson, P. (2013).

Table 1 represents the frequencies of words in two corpora. The variables "a" and "b" refer to the frequency of words in the two corpora being compared, while the variables "c" and "d" refer to the size of each corpus. The expected values can be calculated using the formula below:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

The expected values are the averages for each word adjusted for the corpus size. In the formula above, "N" refers to the total number of words, and "O" corresponds to the observed value. The expected values are represented in Table 1 as c and d . So, we calculate $= c \times (a + b) / (c + d)$ and $= d \times (a + b) / (c + d)$.

The final log-likelihood value is then calculated using the following formula:

$$LL = 2 \sum_i O_i \ln \left(\frac{O_i}{E_i} \right)$$

The formula above represents the distance of the word frequency in each corpus from the previously calculated expected or average values. In terms of Table 1, $LL = 2 \times ((a \times \ln(a)) + (b \times \ln(b/E2)))$.

To enhance the efficacy of text classification experiments, it was necessary to identify words with the highest keyness when comparing three corpora. These keywords were then used as part of a feature engineering process to evaluate the accuracy of algorithms, both with and without these keywords. In the forthcoming methodology chapter, I will provide detailed information regarding the nature of the corpora and how the comparison was conducted to identify the most significant keywords.

2.1.3 Corpus Linguistics and Machine Learning Resources

The next sub-section will describe the resources used for data collection, pre-processing, and analysis. While many of these resources were created for use in corpus linguistics studies, others were borrowed from other fields to aid in the analysis of the data for this research project.

2.1.3.1 *NOW Corpus*

The NOW corpus, which was utilized in the collocation analysis of the first experiment, is a component of the Corpus del Español (News on the Web/NOW), created by Mark Davies and hosted by Brigham Young University (BYU). This corpus is a compilation of 7.2 billion words sourced from web-based newspapers and magazines from twenty different Spanish-speaking countries, and is constantly expanding through automated scripts. The NOW corpus is updated with 180-200 million new words per month (from roughly 300,000 new articles), which amounts to approximately two billion new words each year. The automated scripts

used in this process include the acquisition of 10,000-15,000 URLs from Google News, downloading the web pages with HTTrack, cleaning the content with JusText to eliminate boilerplate material, tagging and lemmatizing the texts with CLAWS 7, removing duplicates based on n-grams, and integrating the text into an existing relational database architecture. The NOW corpus offers various benefits, such as the ability to search based on criteria such as country, source, and period, as well as the capability to research what is happening in the language at present time. Additionally, the corpus architecture enables users to search for specific words, phrases, or families of new words, and provides tools for KWIC and collocation analysis. BYU also hosts numerous other corpora, including The Intelligent Web-Based corpus (14 billion words), the Corpus of Contemporary American English (COCA) (1.0 billion words), the Wikipedia Corpus (1.9 billion words), and the Coronavirus Corpus (635 million+ words), among others. The NOW corpus was instrumental in the first stage of this research study, as it facilitated the collocation analysis to observe how men and women are depicted in digital newspapers. This approach is gaining popularity in fields such as Linguistics and Gender Critical studies, as described in the literature review chapter. In the second stage of the research project, the concordancer software was used to prepare the data for the text classification experiments.

2.1.3.2 Concordancers Generations

When dealing with large corpora consisting of millions of words, it becomes imperative to utilize computer software to efficiently search and extract relevant information. One such tool used to explore corpora is a concordancer. Concordancers are automated systems that compile and display concordances of specific types of occurrences or tokens in a corpus. Please refer to Figure 1 for an illustration.

k Los niños lo adopte una verdadera familia hombre y mujer no dos pinches vatos k se
 UN LECHO FAMILIAR DE MAMA MUJER Y PAPA HOMBRE NO COMO ELLOS DICEN O QUIERAN, SON UNOS
 se habla @Cr P exavto familias hecs de hombre y mujer que al tener sexo da como
 la verdadera palabra de Dios mas q el hombre la quiera modificar q mal ARREPIENTANSE, JE
 que disque su biblia, jesus andubo con puro hombre y compartio su amor con puro hombre ¿ Que
 puro hombre y compartio su amor con puro hombre ¿ Que acaso eso lo ase ser tremendo marico
 hagan lo que quieran Dios le da al hombre libre albedrio, ni modos el que se ensucie
 pero el infierno no fue hecho para el hombre, fue hecho para el ene" "@Gonzalo Tancara n

Figure 1. Sample concordances.

Concordancers have become invaluable tools for linguists to carry out keyword, n-gram, collocation, and concordance analyses, as well as generate frequency lists. Some concordancers even allow users to search for suffixes, multiple words, regular expressions, part-of-speech tags, and other annotations embedded within the corpus. CL scholars have identified several generations of concordancers, each with varying capabilities. In the following paragraphs, I will describe these different types of concordancers.

First-generation concordancers

The first-generation concordancers played a crucial role in identifying the key areas that needed to be improved to enhance the development of future concordancers. At the outset, these concordancers had limited capabilities and were only able to perform a concordance analysis or a word frequency list. Additionally, they struggled to process characters outside of the non-accented Roman alphabet, and any non-standard characters were replaced by designated character sequences. Consequently, there was no consensus on standard conventions for marking up language. These limitations prompted the development of the next generation of concordancers.

Second-generation concordancers

The emergence of personal computers brought about the availability of concordancers such as the Kaye concordance (Kaye, 1990) and the Longman Mini-Concordancer (Chandler, 1989). However, despite this development, many of the issues encountered in first-generation concordancers were not resolved. The variety of analyses that could be conducted remained limited, and there was still no consensus regarding character representation across various formats.

Third-generation concordancers

The third-generation concordancers represent a significant improvement over their predecessors. These tools offer a range of analyses beyond the KWIC analysis, including statistical measures that allow for a more robust analysis. One of the most critical developments of this generation was the implementation of the Unicode standard, which allowed for a corpus analysis across different writing systems. Notable examples of third-generation concordancers include WordSmith (Scott, 1996), MonoConc (Barlow, 1999), and AntConc (Anthony, 2005), which offer concordances, frequency lists, n-grams, collocations, and keyword analysis. However, some scholars, such as McEnery and Hardie (2011), have pointed out that these concordancers may not have reached a stable level of maturity. They suggest that incorporating tools such as collocation networks (Phillips, 1989), multidimensional analysis (Biber, 1988), and collocation analysis (Stefanowitsch & Gries, 2003) could further expand the range of linguistic analyses and research questions. Additionally, third-generation concordancers may struggle to handle corpora over 100 million words.

Fourth-generation concordancers

The fourth-generation concordancers shifted their focus to web-based interfaces. Instead of merely expanding corpus analysis tools, developers of fourth-generation concordancers sought to address issues such as limited computing power, incompatibilities between operating systems, and ethical concerns regarding the distribution of corpora (McEnery & Hardie, 2011). To circumvent these ethical challenges, corpus developers made their materials available via web-based interfaces, allowing users to enter search queries and receive results without having complete access to the corpus (Rayson, 2015). Some notable web interfaces include corpus.byu.edu (Davis, 2020), CQPweb (Hardie, 2020), and SketchEngine (Kilgarriff, 2020). Figure 2 displays some of the most popular computer software tools used for information retrieval.

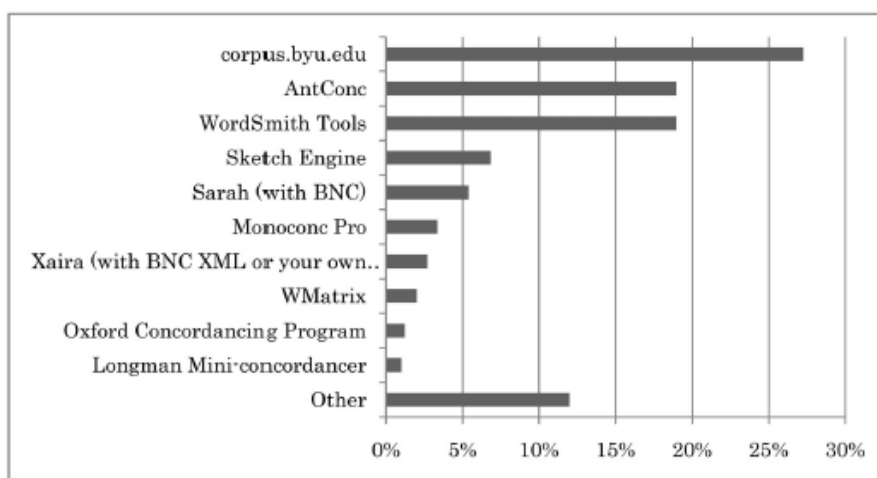


Figure 2. Most popular tools used for analyzing corpora (Tribble 2012 in Anthony, 2015).

With the increasing power of computers and technological advancements, tools for information retrieval have undergone significant developments. Concordancers, in particular, have expanded their applications to include areas such as psychology and other fields that rely on language, texts, or documents for analysis. However, there are still limitations that concordancers place on corpus analysis. As Anthony (2013) notes, the functionality offered by software tools determines the research methods available to researchers, and therefore the design of these tools is becoming increasingly important with the growth of corpora and complexity of statistical analysis. To facilitate research, Rayson (2015) suggests that the fifth generation of concordancers should consider both the disciplinary needs of end-users and interoperability between different software tools. Additionally, corpus linguists may benefit from learning to program and working closely with computer scientists and software engineers to develop the next generation of corpus tools.

For this research study, the AntConc concordance tool, a third-generation concordancer, was utilized to identify keywords for use in the ATC tasks. AntConc enables users to compare a corpus against a reference corpus and identify keywords with higher statistical significance, which were then used to classify comments in several machine learning experiments. The AntConc corpus analysis toolkit is available for Windows, Mac, and Linux-based systems and offers a concordance, word, and keyword analysis generator, tools for cluster and lexical bundles, and a word distribution plot. AntConc also allows for the analysis of tagged corpora.

Figure 3 displays the AntConc interface (Version 3.4.4), which provides access to various linguistic analysis tools.

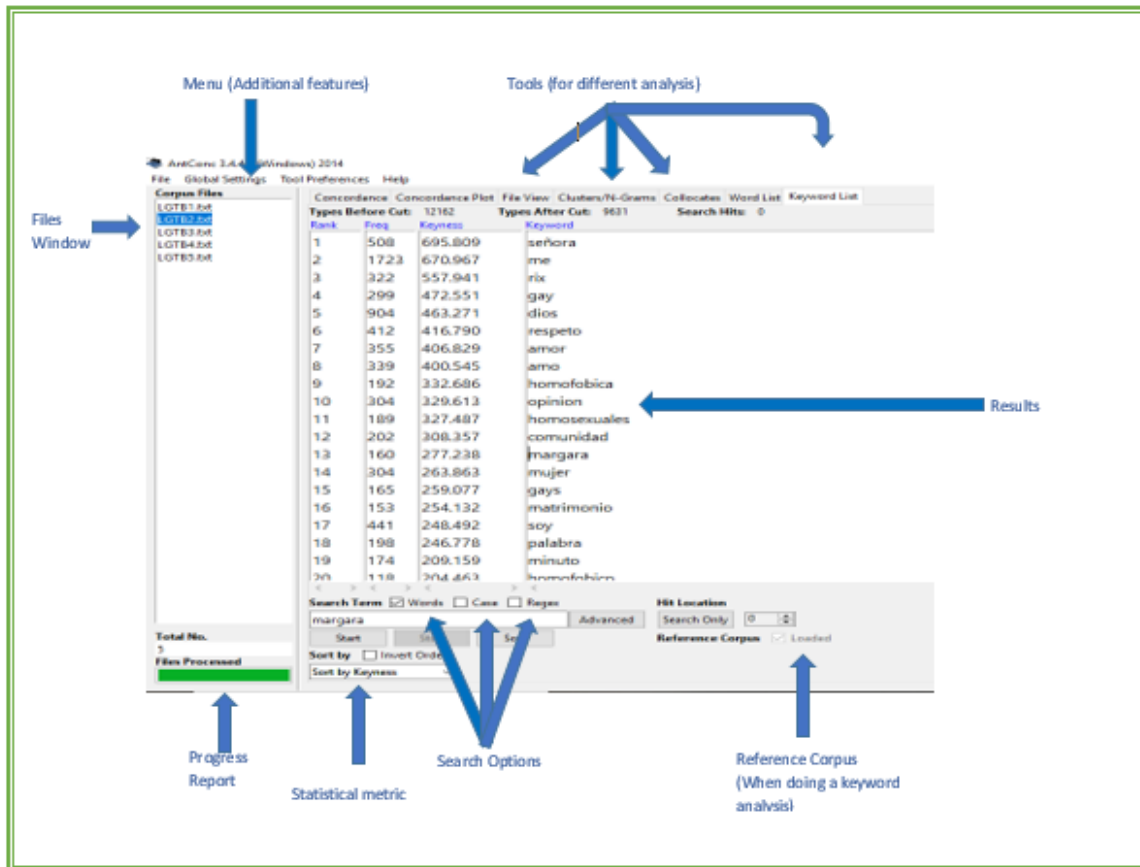


Figure 3. AntConc software tool.

In the next few paragraphs, I will provide a brief overview of the main features of AntConc, including the concordance tool, keyword tool, and collocate tool. Anthony (2013) outlined the following functionalities of the concordance tool:

1. Users can search for substrings, words, or phrases, and can specify whether the search is case-sensitive or insensitive.
2. Regular expressions (REGEX) can be used for complex searches.
3. Clicking on a search term in the KWIC (Keyword in Context) result display will take the user to the original data file.

4. A concordance search term plot tool is available to help users identify how a search term is used in the target corpus.

For this research project, the keyword list tool in AntConc was used to generate a list of words that were later used to classify comments in machine learning experiments. This tool allows users to identify words that appear with unusual frequency in a corpus compared to a reference corpus. Users can choose to calculate the keyness of words using either a log-likelihood or a chi-squared metric. The keyword tool displays each word with its frequency and keyness value, and clicking on a keyword takes the user to the concordance to see how the word is used in context.

AntConc also offers a collocate tool that allows users to search for words (collocates) that appear around a search term (node). The search term can be a word, phrase, or regular expression, and users can specify the window span for the collocates. The tool also allows users to sort the collocates by frequency or statistical measures such as mutual information or T-score. Although AntConc may not be suitable for large corpora, Anthony Laurence notes an increasing interest in working with small and specialized corpora in corpus linguistics. AntConc is regularly updated, with several releases since its launch in 2002.

In summary, I have discussed the NOW corpus and AntConc software, as well as the concepts of collocations and keyword analyses, and the importance of mutual information and keyness in corpus linguistics research. In the next section, I will describe other tools and resources that supported this research project from various areas.

2.1.3.3 *Weka*

In the second stage of the text classification experiments, both AntConc and Weka software were utilized. The Waikato Environment for Knowledge Analysis (WEKA) software, which was developed at the University of Waikato in New Zealand, provides access to a range of machine learning and data mining tools. It offers a variety of learning algorithms for data preprocessing, manipulation, evaluation, and visualization, including algorithms for classification, regression, clustering, and attribute selection. By applying learning algorithms

to a dataset and analyzing the output, users can gain insight into the data. They can also train data using different algorithms for classification and prediction purposes and compare their performance to select the best one for a given task. Notable algorithms available in WEKA include Naïve Bayes, Support Vector Machine, Decision Trees (such as J48), and Random Forest, among others. In addition, WEKA has a package management system that enables the installation of third-party libraries and the use of packages of interest.

The WEKA workbench features different graphical interfaces, each with its own purpose. These interfaces include Explorer, Experimenter, KnowledgeFlow, Workbench, and Simple CLI (see Figure 4).

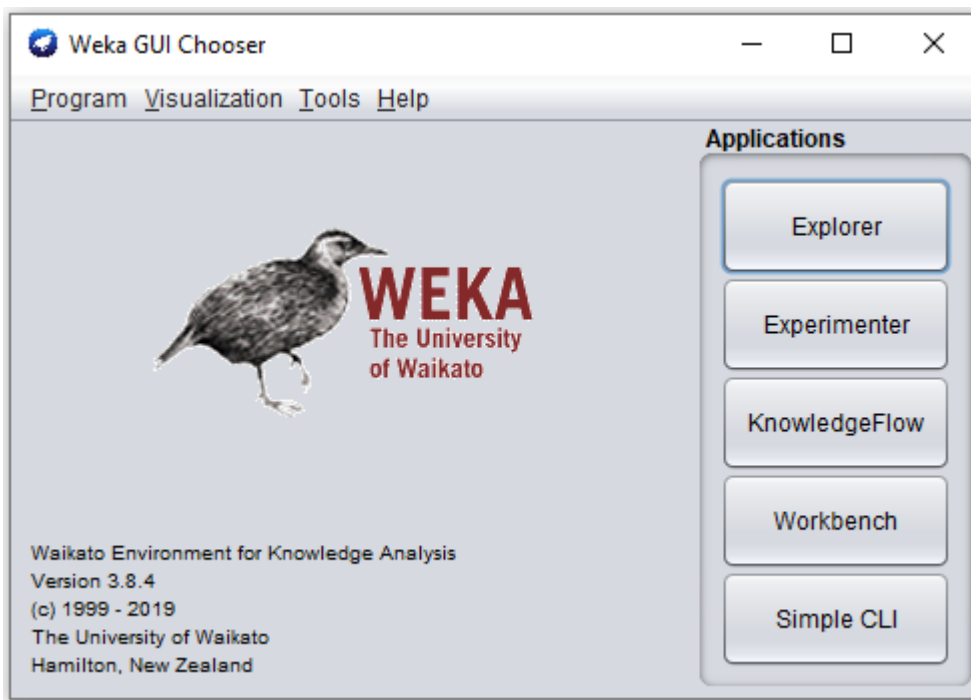


Figure 4. WEKA Workbench

The primary interface utilized in this research project was the Explorer, and thus, the subsequent paragraphs will detail the principal features of the various panels found within the interface. Figure 5 provides a visualization of the Explorer interface.

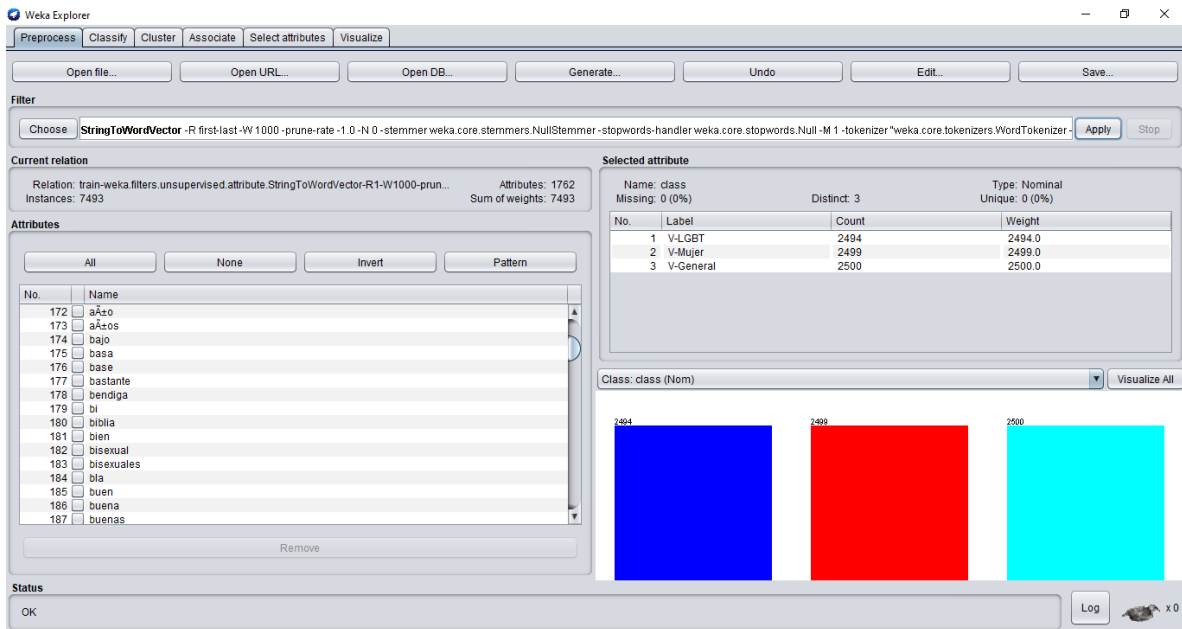


Figure 5. Weka Explorer interface.

The Explorer interface of WEKA provides different panels for data mining tasks, each with specific features that enable users to carry out the analysis of their data. At the top of the interface, users can observe the different panels that serve the different data mining tasks that WEKA supports. The “Preprocess” panel allows users to upload their data files, which must be in WEKA’s ARFF or CSV format, and provides options to edit, delete, or filter instances or attributes as needed.

Moving on to the “Classify” panel, users can access several classification and regression algorithms, and execute a cross-validation test or percentage split test on their data. They can also upload their test data set for analysis. The “Cluster” panel is available for running unsupervised clustering algorithms, although it is important to note that WEKA does not incorporate some of the most popular clustering algorithms.

The “Select Attributes” panel enables users to identify the most important and predictive attributes in their data using a range of algorithms designed for this purpose. Finally, the “Visualize” panel provides users with interactive, two-dimensional plots of their data, allowing them to interact with data points and selected portions of the data.

WEKA software serves as a useful tool for researchers who do not have a background in Computer Science, providing them with access to learning algorithms for their research studies. For linguists who may be hesitant to learn programming to address their research interests, WEKA offers a middle ground and a first approximation to Computational Linguistics and Natural Language Processing. A more detailed description of the different interfaces available in WEKA can be found in Frank et al. (2016).

2.1.3.4 Verb database classification and adjective taxonomy.

To carry out collocation and keyword analysis in this research, corpora were utilized. A more in-depth collocational analysis was conducted to identify the verbs and adjectives that collocated with the lemmas hombre ‘man’ and mujer ‘woman’. This involved classifying the adjectives and verbs according to two different taxonomies: one for verbs and another for adjectives. In the subsequent paragraphs, I will provide a description of the ADESSE database for Spanish verbs and the Supersenses adjective taxonomy.

2.1.3.4.1 ADESSE

ADESSE (Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español) is a syntactic-semantic annotated database for Spanish verbs, developed by the University of Vigo. This ongoing project provides a corpus-based description of verb valency through semantic features, such as verb senses, verb class, and the semantic role of arguments. ADESSE is a manually annotated corpus based on a syntactically analyzed corpus, and it contains 1.5 million words, 159,000 clauses, and 3,450 different verb lemmas. García-Miguel and Albertuz (2005) identify 12 verb classes that reflect large semantic domains, which are further subdivided into 51 subclasses. These 12 verb classes are grouped into larger macro classes.

MACROCLASS	CLASS	VERBS
1 Mental	11 Feeling	186
	12 Perception	72
	13 Cognition	122
2 Relation	21 Attribution	132
	22 Possession	117
3 Material Processes	31 Space	513
	32 Change	394
	33 Other facts	205
	35 Behavior	152
4 Communication		258
5 Existence		115
6 Causative and dispositive		57
TOTAL VERBS		

Table 2. Top-level classes in ADESSE. Taken from García-Miguel & Albertuz, 2005.

As previously mentioned, certain verbs can belong to multiple semantic subclasses; for example, the verb “cambiar,” which falls under five distinct subclasses. Please refer to table 3.

SUBCLASS	VERBS
3200 General	14
3210 Creation	30
3220 Destruction-Consumption	35
3230 Modification	298
3231 Personal Care	17

Table 3. Change subclasses in ADESSE. Taken from García-Miguel & Albertuz, 2005.

It is important to note that the ADESSE classification is still a work in progress and its semantic organization is constantly being refined. Unlike WordNet, ADESSE does not place strict limits on how verbs should be classified, recognizing that verbal meanings are multidimensional and highly flexible. As García-Miguel and Albertuz (2005) state, verb classification is only possible by identifying the basic dimensions of meaning they convey and keeping them separate from contextual influence.

While a detailed description of the ADDESSE project is beyond the scope of this discussion, its flexibility allowed for a more in-depth analysis of the verbs that collocate with both lemmas MAN and WOMAN, which will be presented in the results and analysis chapter. Meanwhile, the adjectives were classified using the Supersense taxonomy, which is described in the following paragraphs.

Adjective Supersenses Classification Framework

To classify the adjectives obtained from the collocations of both lemmas in the NOW corpus, a suitable taxonomy was necessary. While several taxonomies were considered, including those proposed by Peters & Peters (2000) and Dixon & Aikhenvald (2004), manually classifying the adjectives was difficult because the classification categories were not purely semantic in structure. Additionally, no existing classification taxonomies in Spanish were deemed suitable due to the wide variation in adjectival classification based on syntactic and semantic features (Fernandez Alonso, 2015; Romero, 2010).

Ultimately, the “Supersense taxonomy” developed by Tsvetkov et al. (2014) was chosen. This taxonomy was originally used to classify adjectives in GermanNet and was adapted for use in a supervised classification experiment with English adjectives in WordNet. The Supersense taxonomy consists of thirteen coarse semantic classes, each of which is followed by finer-grained subcategories. This allows for the annotation of adjectives at varying levels of granularity, even when classification into coarser classes is difficult. Table 4 provides an overview of the Supersense taxonomy used for the classification of adjectives in this study.

Words	Supersenses	Sub-classes
purple, shiny, taut, glittering, smellier, salty, noisy	Perception	color, lightness, taste, smell, sound
compact, gigantic, circular, hollow, adjacent, far	Spatial	dimension, direction, localization, origin, shape
old, continual, delayed, annual, junior, adult, rapid	Temporal	time, age, velocity, periodicity
gliding, flowing, immobile	Motion	Motion

creamy, frozen, dense, moist, ripe, closed, metallic, dry	Substance	consistency, material temperature, physical properties
rainy, balmy, foggy, hazy, humid	Weather	weather, climate
alive, athletic, muscular, ill, deaf, hungry, female	Body	constitution, affliction, physical sensation, appearance
angry, embarrassed, willing, pleasant, cheerful	Feeling	feeling, stimulus
clever, inventive, silly, educated, conscious	Mind	intelligence, awareness, knowledge, experience
bossy, deceitful, talkative, tame, organized, adept, popular	Behavior	character, inclination, discipline, skill
affluent, upscale, military, devout, Asian, arctic, rural	Social	stratum, politics, religion, ethnicity, nationality, region
billionth, enough, inexpensive, profitable	Quantity	number, amount, cost, profit
important, chaotic, affiliated, equal, similar, vague	Miscellaneous	order, completeness, validity

Table 4. Supersenses taxonomy for adjective classification.

Upon evaluating whether the Supersense taxonomy could aid in the classification of adjectives in this study, it became clear that its subcategories greatly improved the classification process. Thus, the decision was made to adapt the taxonomy to Spanish for the purposes of this experiment, resulting in a more effective tool for consistent classification. The results of this adjectival classification will be presented and discussed in detail in Chapter 6.

2.1.3.4.2 *Violentómetro*

This research study encompasses several major analyses. The first analysis focuses on collocational analysis, which involves identifying verbal and adjectival collocations of the lemmas using the NOW corpus. The second major analysis involves conducting text classification experiments using keywords obtained from YouTube corpora, which were constructed from comments made in videos discussing topics related to women and men.

As part of the second major analysis, the first text classification experiment utilized the verbs listed in the *Violentómetro* as features, rather than the features obtained from the YouTube corpora, to evaluate how they would perform in the experiment. The *Violentómetro* is a

classification system that uses verbs to assess the level of danger in events that women may encounter. It was created by the National Polytechnic Institute's Institutional Management Program with a Gender Perspective and provides users with the ability to recognize instances of gender violence in everyday situations, ranging from behaviors involving verbal or psychological abuse to those describing life-threatening events. Since the YouTube corpora examined gender relationship dynamics, the Violentómetro classification was deemed a valuable source of features for the first classification experiment. The results of this experiment will be discussed in further detail in subsequent chapters.



Figure 6. IPN's Violentómetro

2.2 Computers and human language

The research project is distinguished by its interdisciplinary nature. It utilizes tools from Corpus Linguistics to inform text classification tasks in Natural Language Processing (NLP). Additionally, identifying collocations related to gender adds to the feature engineering

process, a traditional technique in NLP. Although there has been a debate on the distinction between computational linguistics and NLP, the two fields share significant common ground. The focus of the project is to improve the accuracy of various algorithms, with the second major section of this chapter providing a general overview of the NLP field.

2.2.1 Natural Language Processing

Natural language processing (NLP) is an interdisciplinary field that intersects computer science, machine learning, and linguistics. Its primary focus is to build systems capable of processing and understanding human language. NLP has become increasingly popular in various industries, including marketing, healthcare, finance, law, and retail, among others. This field has given rise to a wide range of applications and techniques, including those found in email platforms such as Gmail and Outlook, voice-based assistants like Apple Siri and Amazon Alexa, search engines like Google and Bing, and machine translation services such as Google Translate and Amazon Translate, among others (Vajjala et al., 2020).

According to Vajjala, Majumder, Gupta, and Surana (2020), there are three approaches to solving NLP problems: heuristics, machine learning, and deep learning. A heuristic rule-based NLP system, which is primarily based on word-level formation and regular expressions, uses resources like dictionaries and thesauruses to carry out lexicon-based sentiment analysis. On the other hand, a machine learning-based NLP system utilizes various forms of data, including textual, images, speech, and structured data. Document classification is a successful example of this approach, which has had a significant impact due to the relative simplicity of the learning models needed for algorithm training (Pustejovsky and Stubbs, 2012).

The experiments conducted in this research project utilized the machine learning approach to classify both YouTube videos and comments. The next section of this paper will delve deeper into the machine learning approach to solving NLP problems. The last approach to NLP is a deep learning-based system that uses neural networks to handle complex unstructured data in text classification. Although technically a subset of machine learning,

deep learning employs unstructured data, making it more efficient in analyzing complex data sets. Recurrent Neural Networks (RNNs), one of the various types of neural networks, takes into account that language is sequential, enabling it to interpret sequential information for improved prediction. While RNNs are effective in processing sequential information, one disadvantage is that they cannot remember long contexts, making it difficult to perform well in longer texts.

Machine Learning

This research study underwent a transformation from an initial focus on a pure CL approach to incorporating tools from Machine Learning (ML). In this section, I will provide a general overview of the ML field and explain how it intersects with and informs CL. Figure 7 illustrates how ML fits into the broader computer science landscape as a branch of artificial intelligence with a wide range of applications.

By incorporating ML techniques, this research study was able to enhance its analytical capabilities and draw on a broader set of tools to address research questions. ML has become a vital field in computer science, with applications in areas beyond artificial intelligence. When combined with CL, the result is a powerful approach that can provide insights into complex phenomena.

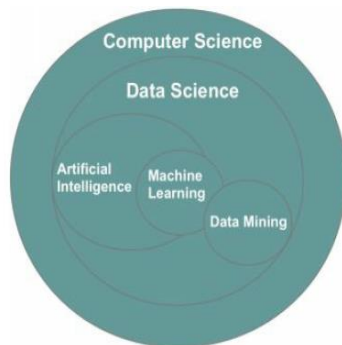


Figure 7. Visual representation of data fields.

Machine Learning (ML) is a field focused on extracting knowledge from data, enabling systems to learn and improve over time. Essentially, ML is a form of artificial intelligence that allows a system to learn from data. As Marsland (2014) suggests, imagine playing Scrabble against a computer. Initially, you may win every game, but after playing many

times, the computer begins to beat you, and eventually, you never win. Once the computer has learned to beat you, it can apply the same strategies against other players without starting from scratch each time. This example provides an approximation of how algorithms learn from data, as they are fed training data, they develop more accurate models over time. Later in this section, we will delve into the details of this process.

Machine learning systems can be categorized based on whether they have been trained with human input or not. The following are the main categories:

Supervised learning: A training set of examples with the correct responses (targets) is provided and, based on this training set, the algorithm generalizes to respond correctly to all possible inputs. This is also called learning from exemplars.

- a) Unsupervised learning: Correct responses are not provided, but instead the algorithm tries to identify similarities between the inputs so that inputs that have something in common are categorized together. The statistical approach to unsupervised learning is known as density estimation.

(Marsland, 2014, p. 6)

- b) Reinforcement learning: It is a behavioral learning model. The algorithm receives feedback from the analysis of the data so the user is guided to the best outcome. Reinforcement learning differs from other types of supervised learning because the system isn't trained with the sample data set. Rather, the system learns through trial and error.
- c) Neural networks and deep learning: Deep learning is a specific method of machine learning that incorporates neural networks in successive layers to learn from data iteratively. Deep learning is especially useful when you're trying to learn patterns from unstructured data. Deep learning, complex neural networks, are designed to emulate how the human brain works so computers can be trained to deal with abstractions and problems that are poorly defined.

(Hurwitz & Kirsch, 2018, p. 17)

This research study employed a supervised learning approach because the experiments aimed to train algorithms to learn patterns by establishing relationships between variables with labeled datasets. In supervised learning, machines are fed with labeled sample data with various features (represented as x) and correct value output (represented as y). The algorithm

then deciphers the patterns in the data and creates a model that reproduces the underlying rules with new data. Some of the most important supervised algorithms are Support Vector Machine, Naïve Bayes, K-nearest Neighbor, Decision Trees, and Random Forest.

While corpus linguistics (CL) uses massive corpora, computers assist researchers in linguistic analysis. Machine learning (ML) focuses on automating knowledge discovery and linguistic modeling by analyzing annotated corpus material, where the annotation is the target of the learning. However, fully understanding human language still poses challenges for computers. Statistical techniques have helped narrow the gap, but ML techniques perform better when provided with pointers, such as keywords, to what is relevant in a dataset. Pustejovsky and Stubbs (2012) note that metadata must be accurate and relevant to the task the machine is performing for the pointers to be useful. Through such additional information, algorithms can learn more efficiently and effectively.

As previously mentioned, one of the most popular tasks in machine learning is document (text) classification. This research study conducted several text classification experiments based on keywords. In the following paragraphs, we describe in detail the procedure for executing such tasks.

2.2.1.1 Automatic Text Classification

The research study focused on Automatic Text Classification (ATC) as a central task. The comments gathered for each YouTube video were compiled into three different corpora: Viomujdus, Viogendis, and LGBTdis, which were later renamed as V-Mujer, V-General, and V-LGBT for running other experiments. The comments were classified based on the most relevant features or keywords for each corpus. In the following paragraphs, I will provide a detailed explanation of the ATC procedure.

ATC is a discipline that intersects ML and Information Retrieval (IR), and shares several characteristics with other tasks such as text mining. As such, ATC can be considered as an instance of text mining. According to Sebastiani (2005), the applications of ATC technology include:

Newsire filtering (grouping news stories according to thematic classes)

- Patent classification (to organize taxonomies to detect existing patents related to a new patent)
- Web page classification (grouping web pages (sites) according to the taxonomies classification schemes typical of web portals).

These applications share a common characteristic in that they use a thematic approach to their classification procedures. However, it is worth noting that ATC is not limited to thematic domains. Sebastiani (2005) outlines additional applications of ATC in other domains, including:

- Spam filtering (grouping email messages as either legitimate or spam)
- Authorship attribution (the automatic identification of the author of a text among a predefined set of)
- Author gender detection (related to authorship attribution but the task here is to identify if the author of the text is male or female).
- Affective rating (deciding of a product review is a thumbs up or a thumbs down.)

One area of ATC application is opinion classification, which can involve determining if the information in a text is objective or subjective, or if the opinions expressed in the text are positive or negative, as well as the degree to which opinions are expressed. In general, ATC involves using machine learning to automatically assign a text document to predefined classes, based on textual features extracted from the document. According to Dalal and Zaveri (2011), a common approach to ATC involves the following steps:

Document pre-processing

- I) Feature extraction/selection
- II) Model selection
- III) Training and testing the classifier

In the first phase of Automatic Text Classification (ATC), stop-words, which are non-specific and do not aid in discrimination among classes, are eliminated. This includes functional words such as articles, prepositions, and auxiliary verbs. Additionally, stemming is used to

reduce words to their base form, consolidating singular, plural, and different tenses into a single word. This process significantly reduces the size of the documents, and if the data comes from web sources, it undergoes further pre-processing.

The second phase focuses on identifying important words in the documents, which can be done using statistical or semantic approaches such as the TF-IDF (term frequency – inverse document frequency) model or the Latent Semantic Indexing (LSI), respectively. Other methods include Mutual Information (MI), which is commonly used in statistical language modeling of word association, and Information Gain (Info Gain), which is frequently employed as a term goodness criterion in the field of machine learning. The important words identified in this phase are referred to as features, attributes, or variables.

During this phase, each document is represented as a document vector to reduce the complexity of the documents and make them easier to handle. This indexing preprocessing results in a vector space model (VSM) representation, where each column stores the features, and each row stores the instances (documents). The representation adopts a Multinomial Model in which each vector retains the information regarding the frequency of each occurrence (feature) in every instance (document). Table 5 illustrates this representation, which is useful for retaining information about the frequency of occurrences of the feature terms in each document.

	Vector	Matrices	
	Feature 1 (keyword)	Feature 2	Feature 3
Text 1(Viomujdis 1)	0		
Text 2(Viomujdis 2)	2		
Text 3	1		
Text 4	16		

Table 5. Vector space model representation

It is worth noting that the VSM, also known as Bag of Words (BoW), has certain limitations. For instance, it results in a high-dimensional representation and fails to take into account the correlation with adjacent words and semantic relationships between terms (features) in a document. To address these issues, term weighting methods have been developed to assign appropriate weights to the terms (Korde & Mahender, 2012). The three most commonly used weighting schemes are Boolean, Word Frequency, and TF-IDF.

The Boolean weighting scheme assigns a value of 1 to a term if it appears in the document and a value of 0 if it does not.

$$a_{ik} = \begin{cases} 1 & \text{si } f_{ik} > 0 \\ 0 & \text{en otro caso} \end{cases}$$

Where f_{ik} is the frequency of the word i in the document k .

In the word frequency scheme, the frequency of each word in every document is taken into account.

$$a_{ik} = f_{ik}$$

Where f_{ik} is the frequency of the word i in the document k .

The TF-IDF scheme takes into account the frequency of a word in each document and across all documents in a class. This method, which stands for Term Frequency multiplied by Inverse Document Frequency, measures the relevance of a word in a collection of documents. To calculate TF-IDF, the frequency of a word in a document is multiplied by the inverse document frequency of the word across a set of documents.

$$a_{ik} = f_{ik} \times \log \left(\frac{N}{n_i} \right)$$

In this formula, $a_{(ik)}$ is a product of f_{ik} and $\log(N/n_i)$, where f_{ik} is the frequency of the word i in document k , N is the total number of documents in the class, and n_i is the number of documents that contain the word i . To illustrate, suppose a document has 100 words and the word “powerful” appears three times. The term frequency for “powerful” is $(3/100) = 0.03$. If there are 10 million documents and “powerful” appears in 1000 of them, the inverse

document frequency is $\log(10,000,000/1000) = 4$. The resulting TF-IDF weight is obtained by multiplying these values: $0.03 * 4 = 0.12$.

In the third phase, a suitable machine learning algorithm is used to train the text classifier. Depending on the number of classes and features, several algorithms can be employed, including Naïve Bayes, Neural Networks, Support Vector Machines (SMV), and Decision Trees. In the fourth phase, the trained classifier is tested using either a test set of text documents or by partitioning the training set and using cross-validation. If the trained classifier's classification accuracy is found to be acceptable for the test set, the model is then utilized to classify new instances of text documents. Figure 8 depicts a generic view of the document classification process.

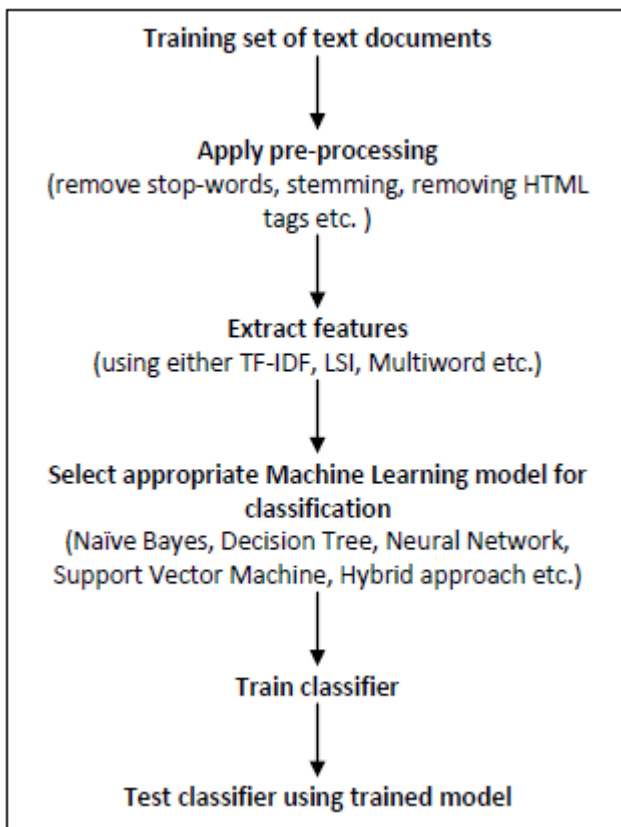


Figure 8. Document classification process. Taken from (Dalal & Zaveri, 2011).

The previous paragraphs provided an overview of how text classification tasks are typically conducted. In the upcoming methodology chapter, I will outline the specific approach that was taken to apply the keywords obtained from the YouTube corpora to the text classification process.

2.3 Language and the World Wide Web

The use of digital communication data has experienced an exponential expansion, leading to research being conducted across various disciplines such as Sociology, Communication, Psychology, and Linguistics. Additionally, other fields, including Medicine, Business, Law, and Data Science, have also come to rely on digital information to conduct research. Since language serves as a means to research subjects in all these disciplines, linguistics plays a central role in this new genre.

In this research study, the corpora used were derived from the World Wide Web and social media platforms. Therefore, it is essential to understand the significance of the web as a source of language data. Furthermore, it is important to recognize how social media has contributed to Corpus Linguistics research.

2.3.1 The Web and Corpus Linguistics

To determine whether the web can be considered a corpus, Kilgarriff and Grefenstette (2003) question previous definitions and requirements for a corpus. While some scholars stress that a corpus should be of finite size, machine-readable format, a standard reference, and representative, many corpora used in research do not meet all of these criteria. Kilgarriff and Grefenstette (2003) emphasize the need to distinguish between the questions "What is a corpus?" and "What is a good corpus for certain kinds of linguistic studies?" in order to avoid conflating different issues. If we limit the definition of corpus to "a collection of text," without adhering to prescriptive requirements, then the web can be considered a corpus.

Once this semantic issue has been settled, it is important to describe how the web has been used in linguistic research. Hundt, Nesselhauf, and Biewer (2007) outline two approaches commonly employed in corpus linguistics when using the web as a source of data.

- a. With the help of internet-based engines, the web can be used as a corpus itself (“Web as corpus”)
- b. The web can alternatively be used as a source for the compilation of large offline monitor corpora (“Web for corpus building”).

(p. 2)

Hundt et al. (2007) point out limitations of the first approach and acknowledge certain disadvantages, including the quality of information, the transience of the web, and limited access to information. Conversely, they identify three advantages of the “Web for corpus building” approach, namely, control, accessibility, and level of analysis. With respect to control, researchers have the ability to decide which texts to include in their database, allowing them to be more familiar with the content. Regarding accessibility, standard software can be used once the corpus has been transferred offline. Finally, offline corpora can be annotated, allowing for a wider range of analyses to be conducted. The corpus is useful for studying linguistic behavior and forms used in socializing through digital discourse, with lexical, syntactic, semantic, and discourse issues being among the areas that can be examined. Discourse behavior, such as identity, politeness, rhetorical strategies, gender, power, and ideology, can be studied in particular. It should be noted that the term “Web as Corpus” is generally used to describe all aspects of empirical language research based on textual material collected from the web, including those that should be labeled as “Web for corpus building” according to the definition given above (Bergh & Zanchetta, 2008).

Based on the preceding discussion, the “Web for corpus building” approach appears to be better suited for corpus linguistics as linguistic analysis demands different types of data. For instance, if the research project is concerned with gendered discourse, a standard corpus such as the British National Corpus (BNC) may not be sufficient, and a corpus specific to a particular genre may be required. Nevertheless, regardless of the discipline, the web as a data

source has become popular due to its massive size, extensive linguistic, geographic, and social range, currency, multimodality, and broad availability at minimal cost (Fletcher, 2004).

However, the “Web for corpus building” approach has not been without criticism, with the quality of web texts being one of the issues raised. Web texts may contain more errors compared to traditional electronic text corpora. However, despite this drawback, web texts are useful and attractive for online language references due to their vast size and ease of searching (Lew, 2012, p. 298).

Since the advent of the web as a data source, corpora’s size has grown exponentially. The first computer-based corpus, the Brown Corpus, contained one million words in 1960, while the British National Corpus contained 100 million words in 1980. However, the size of these corpora pales in comparison to those generated through the web, which now range in the billions of words.

The use of the web as a tool to build corpora began in 1999 in computational linguistics. A research project presented a method to disambiguate all the nouns, verbs, adverbs, and adjectives in a text using the senses provided by WordNet, ranking the senses using statistics gathered from the internet for word-word occurrences (Mihalcea & Moldovan, 1999). Jones and Ghani (2000) demonstrated how to automatically construct queries to access documents on the World Wide Web a year later. Since then, the web has been used not only in computational linguistics but also in various other disciplines, not only for information retrieval purposes or building corpora but also for all types of linguistic analysis.

Web texts serve as a source of not only valuable and otherwise inaccessible linguistic data but also new and rare words that are absent in existing corpora. Moreover, web texts are freely available, vast in number and volume, constantly updated, and reflect the latest language usage (Renouf, 2007, p. 42). The web’s impact on corpus linguistics and other disciplines is undeniable, with access to diverse linguistic data resulting in a broader range of linguistic analyses. In addition, the emergence of new genres and types of text through web data has expanded the scope of research in linguistics. For example, researchers examine

how limited formatting and cumbersome input technology encourage short sentences, abbreviations, and emoticons in mobile phone text messages (Lindquist, 2009, p. 224).

In light of the above, the following section will discuss how social media and computer-mediated communication (CMC) sources have extended the scope of linguistic analysis, given that the research project involved compiling corpora from YouTube.

2.3.2 Social Media and Language Research

In the previous section, I discussed how the web has contributed to corpus linguistics by enabling the construction of corpora, and the subsequent opportunities for linguistic analysis. However, the internet has also provided a unique opportunity to study language use in social media contexts, where communication takes place through various formats such as messages, images, and videos. These social networks, falling under the umbrella of Computer-Mediated Communication (CMC), are designed to facilitate communication and build social relationships.

This section will explore how linguists can utilize CMC and social networks for research purposes. While there are various types of CMC environments, such as Twitter, Facebook, YouTube, electronic mail, instant messaging, chats, discussion forums, blogs, and video conferencing, this section will focus on social media and social networks interchangeably, both of which prioritize social interaction.

To provide a general overview of the opportunities presented by CMC and social networks, Page, Unger, Zappavigna, and Barton (2014) have detailed some of the topics that can be researched:

- Linguistic practices: what people do with language, the regular behaviors that develop within particular communities, and how language is used to perform particular identities (for instance, linguists might analyze how a forum community uses narratives/stories to enhance

group cohesion, or how Facebook friends code-switch between different languages to signal their linguistic identities).

- Texts/utterances: collections of words, clauses, and sentences arranged deliberately in a structure with a clear communicative function. When a certain type of text becomes easily recognizable, this is often referred to as a genre, e.g., a comment thread on a newspaper site. This level of language is also sometimes referred to as discourse.
- Clauses and sentences: strings of words arranged in a structure, often described as syntax or grammar.
- Lexemes or words: units of meaning consisting of one or more morphemes, like ‘eggs’
- Morphemes: the smallest units of meaning, e.g., ‘egg’, which calls up a certain concept in our minds, or ‘-s’ to indicate plurality.
- Phonemes: individual sounds/signs that make up spoken or signed words; and graphemes, e.g., letters or characters in writing.

(p. 31)

When it comes to linguistic analysis, the potential for exploration expands even further when dealing with both written and spoken language. Additionally, computer-mediated communication (CMC) sources have become an essential tool not just in Linguistics but in other fields like Machine Learning and Data Mining from Computer Sciences.

Linguistics has already started to integrate CMC into various disciplines, such as discourse analysis, sociolinguistics, and language learning, among others. The emergence of Computer-Mediated Discourse Analysis (CMDA) has already yielded numerous discourse-focused studies that use CMC environments like Twitter, Facebook, and YouTube for data collection and analysis. These studies can vary widely in their focus, but Herring (2013), who is a proponent of CMDA, identifies four levels that encompass most studies in this field. Table 6 provides more information on these levels.

Levels	Issues	Phenomena	Methods
Structure	Orality, formality, complexity, efficiency, expressivity, genre characteristics, etc.	Typography, orthography, morphology, syntax, discourse schemata, formatting conventions, etc.	Structural/descriptive linguistics, text analysis, stylistics
Meaning	What is intended, what is communicated, what is accomplished	Meaning of words, utterances (speech acts), exchanges, etc.	Semantics, pragmatics
Interaction management	Interactivity, timing, coherence, repair, interaction as co-constructed, etc.	Turns, sequences, exchanges, threads, etc.	Conversation analysis, ethnomethodology
Social phenomena	Social dynamics, power, influence, identity, community, cultural differences, etc.	Linguistic expressions of status, conflict, negotiation, face-management, play, discourse styles/lects, etc.	Interactional sociolinguistics, critical discourse analysis, ethnography of communication

Table 6. Four levels of CMDA (taken from Herring, 2013)

According to Androutsopoulos (2006), CMC provides a new and valuable empirical arena for various research traditions within sociolinguistics. Online ethnography, for example, allows researchers to investigate internet cultures, chart the dynamics of online activities related to offline events, or study identity constructions. While CMC presents challenges when identifying gender, social class, or race, it still offers opportunities for exploring language variation. Sociolinguistic research on CMC can also focus on interactional linguistics, communities of practice, or gendered discourses. With the former, linguists can explore how linguistic structures and language use are displayed in CMC and whether interactional coherence is affected by the nature of online communication. The latter two approaches can reveal how the widespread use of online interaction is challenging and rewriting the idea of community, as well as exploring gender differences in language use online and how this relates to prior research.

Language learning is one area where the impact of CMC sources within linguistics is especially evident. Within the field of Computer-Assisted Language Learning (CALL), CMC sources have been used in language learning settings for some time. Early studies suggested that the use of synchronous CMC not only increased the amount of language learners produced but also broadened the variety of language forms and functions they employed. Furthermore, it reduced anxiety levels among participants and promoted more balanced participation compared to face-to-face discussions. However, as new media applications

emerge every day, they present both opportunities and challenges that require further research and exploration. With the proliferation of online applications, video tutorials, and language learning websites, researchers can now explore the content of the material, interactions developed in online applications, and pedagogical practices demonstrated in video tutorials or online classes.

In summary, the internet and CMC sources have significantly contributed to language research. In the following section, I will discuss how YouTube, a social media outlet, has been used in several research studies, including the construction of corpora.

2.3.2.1 YouTube

In the second stage of this research study, the corpus was expanded through the creation of a new corpus using YouTube social media videos focused on gender issues. The decision to use YouTube was due to its popularity as a platform for online video sharing and its potential for capturing people's opinions. However, this new corpus also presented some limitations, mainly due to the nature of the language used in the videos. Nonetheless, the corpus was still useful in guiding feature selection for text classification experiments.

The YouTube website provides a user-friendly interface that allows almost anyone to publish videos online, making it a valuable source for data collection. Although the technical aspects of the platform are not relevant to this research, it is worth noting that YouTube's popularity and accessibility make it an excellent source for collecting language and gender-related data. It is essential to recognize that when studying the collocations of "man" and "woman" and the classification of comments based on gender-related topics, we are dealing with gendered discourses. We are examining how men, women, or the LGBTQ+ community are portrayed through language. Therefore, the selection of a social network to collect data was a critical decision. YouTube was chosen because of its potential to provide valuable insights into people's opinions on gender issues.

In summary, the decision to use YouTube as a source for data collection was based on its popularity and accessibility, making it a useful platform for examining language and gender-related issues.

It is important to note that YouTube was originally designed to broadcast media entertainment content; in other words, it was foremost a commercial enterprise (Burgess & Green, 2013, p. 76). However, as people started using this platform to share videos, it was clear that an unintended use was being developed, and that is that YouTube became a place for cultural participation (Burgess & Green, 2013). Nowadays, users not only consume content but also interact socially with others. Unlike Twitter and Facebook where social networking is based on personal profiling, in YouTube, the video content is the main vehicle for communication. Therefore, taking into account that YouTube enables cultural participation by ordinary citizens. It is through the videos that people set the topics that are to be shared and discussed for members within a community. Burgess and Green (2013) doubt that YouTube developers ever intended to create a space for cultural participation; furthermore, they ask to consider the idea that “YouTube may be generating public and civil value as an unintended and often unsupported consequence of the practices of its users.” (p. 76). Taking into account that many of the videos shared via YouTube originate in the everyday lives of its users, YouTube represents a place where people’s culture takes place. In this sense, YouTube sets the conditions to establish not only local but also broader communities of practice in which its users can express their identities, share their values, engage with others, negotiate meaning, and encounter cultural differences.

Very often antagonism and controversy arise in the YouTube communities and this is one of the reasons I opted to rely on YouTube for data collection. The antagonism may uncover discourse practices deeply rooted in controversial topics such as gender inequalities or same-sex marriage. Moreover, it is through interactions that antagonism or controversies contribute to developing new literacies, new cultural forms, and new social practices that are constructed, challenged, rejected, or adopted. Besides the antagonism that derives from online interactions, there is another issue that has attracted attention when researching online interaction and that is the anonymity that users benefit from. Pihlaja (2014), citing Hardaker (2010) mentions that CMC sources offer a high degree of anonymity which may foster the

effect of deindividuation that may lead users to develop a sense of impunity, loss of self-awareness, and a likelihood of acting upon normally inhibited impulses. However, anonymity also presents users the opportunity to engage in conversations with people that will not otherwise occur due to the nature of topics. In other words, users can set the agenda for those topics that they considered ought to be discussed not only among online communities but also among different communities across societies. YouTube presents a space for disenfranchised communities, a space that is usually not offered by those in power or by mainstream media. In his work, *Outline: Trans Self-Representation and Community Building on YouTube*, Raun (2016) researched how some members of the Trans community have relied on YouTube to establish social interactions which are cut off from their offline lives. He describes that Trans people are increasingly stepping out of the shadow of pathologization and secretiveness to tell their life stories and share information to connect with like-minded others while using YouTube as a platform. In a struggle to exercise agency, YouTube allows Trans to challenge how they are represented in mainstream media and enforce how they want to be perceived. In this sense, “Representation carries a special political weight for minority groups and plays a significant role in the formation and visibility of social movements and identities” (Raun, 2016, p. 22).

In this section, I have described how the World Wide Web has influenced corpus linguistics mainly in the area of corpus building. I have also detailed how CMC sources have been utilized in different areas within linguistics such as discourse analysis, sociolinguistics, and language learning. Finally, I described how the YouTube social media outlet has become a place for cultural participation where disenfranchised communities can have their voices heard. The world wide web has become a rich source of linguistic data, and the opportunities that this presents not only to linguists but also to other disciplines are vast.

It is worth noting that YouTube was originally created as a commercial platform for broadcasting media entertainment. However, as people began to share their own videos, it became clear that YouTube had become a place for cultural participation. Unlike social networking sites such as Twitter and Facebook, where personal profiling is the basis of social interaction, YouTube primarily relies on video content as a vehicle for communication. This platform allows for the cultural participation of ordinary citizens, providing a space for

people to set topics for discussion and engage with others in local and global communities of practice. Burgess and Green (2013) doubt that YouTube developers ever intended for it to become a space for cultural participation, suggesting that its cultural value is an unintended consequence of its users' practices. However, YouTube has enabled individuals to express their identities, share their values, negotiate meaning, and encounter cultural differences.

Antagonism and controversy often arise in YouTube communities, making it a suitable platform for data collection. Such interactions may reveal discourse practices rooted in controversial topics, such as gender inequalities or same-sex marriage. Furthermore, the resulting antagonism and controversies contribute to the development of new literacies, cultural forms, and social practices, which are challenged, adopted, or rejected. Although anonymity on CMC sources may foster deindividuation and a loss of self-awareness, it also allows users to engage in conversations with people with whom they may not interact otherwise. This feature enables users to discuss topics across different communities and societies that they consider essential for discussion. In this way, YouTube provides a space for disenfranchised communities that is not typically offered by those in power or mainstream media. Raun's (2016) study on the Trans community highlights the ways in which members of this group have used YouTube to establish social interactions that are separate from their offline lives. YouTube has allowed Trans individuals to challenge mainstream media representations and enforce how they want to be perceived.

In summary, the World Wide Web has provided rich linguistic data, which has influenced corpus linguistics, mainly in corpus building. Additionally, CMC sources have been used in various linguistic areas, such as discourse analysis, sociolinguistics, and language learning. YouTube, in particular, has become a space for cultural participation, where people can express their identities and engage with others across communities. This platform allows disenfranchised communities to have their voices heard and to challenge mainstream media representation.

3 Literature Review

In this chapter, I will explore research studies conducted in the field of Language and Gender. First, I will discuss early and significant studies that utilized a qualitative approach. Next, I will examine research studies that employed Corpus Linguistics as the primary research method and adopted a quantitative approach to data analysis. Finally, I will explore studies that integrated tools and resources from both Corpus Linguistics and Machine Learning. This section will not only demonstrate the integration of these disciplines in linguistic analysis but also elaborate on how this integration can inform the Language and Gender field.

3.1 Language and Gender

The study of Gender and Language has evolved through four distinct theoretical approaches: the “deficit” approach, the “dominance” approach, the “difference” approach, and the “dynamic” or “social constructionist” approach (Litosseliti, 2014; Coates, 2015). However, as research in the field of Gender and Language becomes more diverse, it has moved beyond simply exploring connections between gender and language. Instead, it now encompasses a broader scope that addresses socio-cultural conventions.

Deficit model

One of the earliest works on gender and language was Otto Jespersen’s “Language: Its nature, development, and origin” in 1922, which falls under the ‘deficit’ model. This approach attributed gender differences in language use to innate characteristics and abilities of men and women. Jespersen claimed that women used more adverbs of intensity due to a tendency towards hyperbole, did not finish their sentences because of a lack of forethought, and had a less extensive vocabulary than men. Additionally, women were thought to avoid vulgarity and swearing, while men were considered the primary innovators of language. However, these claims lacked methodological rigor and were based on personal beliefs rather than ethnographic or anthropological research.

Despite its lack of scientific validity, Jespersen's work was influential in stimulating further research and challenging preconceptions about women's language use. However, it is now widely recognized that gender differences in language use are not innate, but are rather shaped by social practices and cultural conventions. Therefore, contemporary research in Gender and Language has shifted towards exploring how gender and language intersect with broader socio-cultural factors, such as power dynamics, socialization, and identity construction.

the highest linguistic genius and the lowest degree of linguistic imbecility are very rarely found among women. The greatest orators, the most famous literary artists, have been men; but it may serve as a sort of consolation to the other sex that there are a much greater number of men than women who cannot put two words together intelligibly, who stutter and stammer and hesitate, and are unable to find suitable expressions for the simplest thought. Between these two extremes, the woman moves with a sure and supple tongue that is ever ready to find words and to pronounce them in a clear and intelligible manner. (Jespersen 1922, p. 253, from Thomas, 2013)

Jespersen's theories on language differences between men and women centered around denigrating women's linguistic abilities while elevating men's. While Jespersen's work is now considered to be flawed, it did pave the way for further research into gender and language.

One of the earliest studies on this topic was conducted by Swift and Miller (1981/2001), who explored issues such as the use of "man" as a false generic, gendered double standards, and assigning gender to gender-neutral terms. They also suggested non-sexist alternatives to the pronoun "he" and examined clichés like "man-in-the-street," masculine gender titles for jobs that can be performed by both men and women, and prefixed compounds like "man-made." This study represented an early attempt to analyze gender and language using a lexical and syntactical approach. Even today, researchers continue to investigate these discrepancies.

Jespersen and Swift and Miller both argued that sexist language is not limited to English and is instead reflective of the prejudices present in the societies where these languages have evolved. Schulz (1975) similarly showed that words for women tend to develop in a derogatory direction compared to words for men, as seen in the example of “bachelor” versus “spinster” and “master” versus “mistress.” Schulz also demonstrated that this trend can be observed in other languages, such as Spanish, where the word "mujerzuela" has a secondary meaning of “prostitute” whereas “Hombrezuelo” does not necessarily is pejorative.

In addition to examining language that portrays women as sex objects, researchers employing the “deficit” model also investigated depictions of women in domestic roles that trivialize them, such as the portrayal of women as weathergirls. While early research in this area had its limitations, it did bring to light the pervasive influence of an androcentric ideology that is deeply ingrained in patriarchal societies and permeates language.

Dominance model

The second approach to gender and language research focused on the relationship between gender and power status in determining speech styles. This model explains the linguistic differences between women and men in terms of men’s dominance and women’s subordination, considering women as an oppressed group.

Early research that supported the dominance model focused on examining linguistic features such as questions, hedges, interruptions, qualifiers, back-channeling, topic initiation, and topic control, which were argued to reflect and perpetuate male dominance (Litosseliti, 2014). One of the most influential works on gender and language, which encompasses both the “deficit” and “dominance” models, is Robin Lakoff’s *Language and Woman’s Place*. Lakoff aimed to provide evidence of inequality in society through the analysis of language use.

Lakoff (1975) observed that women’s language varies in the use of lexical items, such as colors, which she argued was due to the expectation that women were relegated to making non-crucial decisions such as finding fine discrimination among colors. She also noted that women tended to use weaker expletives, such as “oh dear,” instead of “damn,” which

reflected the strength of their emotions. Furthermore, Lakoff argued that women's heavily qualified statements and use of tag questions were signs of their uncertainty and efforts not to impose their subordinate view onto their interlocutors.

Like Jespersen, Lakoff's data-gathering techniques deviated from accepted research conventions, relying on introspection and analyzing her own speech and that of her acquaintances. She also used "case studies" from different institutions, such as academia, the arts, politics, and the media, to illustrate the relationship between gender and power. Lakoff attempted to find a rationale for men and women's linguistic differences in what she called a "socialization process," arguing that society keeps female children in line by criticizing or scolding them if they "talk roughly" like a boy.

Despite the introspective nature of her data collection, Lakoff's work was evaluated within the context of the time it was published. Her claims served as a means for academia to evaluate the androcentric view as a phenomenon that pervades across disciplines and people's everyday lives.

While Lakoff did not question the origin of linguistic differences between women and men, Spender (1980) argued that a male-dominant society promoted the belief that men were the superior sex and that social institutions and practices were organized accordingly. As a result, the meaning in language was defined by men. Spender challenged research that accepted male language as the norm, arguing that the deficiency lay not with women's language but with the social order. She claimed that men controlled the meaning of public discourse, shaping both the meaning and form of language. Spender's work laid the foundation for feminist linguistics, which seeks to identify, demystify, and resist the ways language is used to create and sustain gender divisions and inequalities (Talbot, 1988, in Litosseliti, 2014).

Fishman's (1978) research on everyday conversation used a dominance model to observe the representation of power and how it helped establish relationships between women and men. She defined power as the ability to impose one's definition of what was possible, right, rational, and real. Fishman's data included fifty-two hours of recorded conversation between three couples, and she found that women tended to ask more questions than men, which she interpreted as a way of providing conversational support. Fishman also observed that women

used minimal responses like “yeah,” “umm,” and “huh” to show they were attending to what was being said, whereas men used them to display a lack of interest. Contrary to Jespersen’s claim that women spoke more, Fishman found that men produced over twice as many statements as women and that men’s responses were often longer. Despite two of the women identifying themselves as feminists and all three men sympathizing with the women’s movement, men controlled the conversation topics, dropping those introduced by women and pursuing those introduced by men. Zimmerman and West’s (1975) study on power and gender in conversational patterns similarly found that men exercised dominance in their interactions with women, including interrupting, controlling turn-taking, and showing inattentiveness.

These studies demonstrate the need for a multidisciplinary approach to studying gender and language. While some, like Lakoff, attribute linguistic differences to women’s subordinate role in society and others, like Spender, attribute them to male dominance, the difference model, which will be described in the following paragraphs, attributes the differences to the belief that women and men belong to different cultures.

3.1.1 The difference (cultural) model

The difference theorists analyze linguistic variations between women and men based on the social and cultural contexts in which they interact. Specifically, these theorists argue that socialization into different subcultures and gender roles can influence speech styles. Maltz and Borker (1989/2012) proposed a new framework for examining differences in the speech patterns of women and men. Their approach emphasized cultural differences between genders, rather than psychological differences or power differentials. In the difference model, also known as the “Two cultures approach,” it is believed that men and women have different genderlects, which can lead to miscommunication. By examining social interactions, it is possible to understand the source of male-female differences in language use.

Research on children’s play cited by Maltz and Borker (1989/2012) suggests that girls tend to use language to create and maintain relationships of closeness and equality, to criticize

others in acceptable (indirect) ways, and to interpret accurately and sensitively the speech of other girls. In contrast, boys tend to use language to assert dominance, attract and maintain an audience, and assert oneself when others have the floor. The difference model argues that these linguistic differences are culturally determined, unlike the deficit model which places blame on women for differences between genders, or the dominance model which blames powerful groups. Crawford (1995) similarly argued that communication between genders is communication across cultures, and that gender roles are enacted in specific ways depending on the context.

Consider the difficulties of talk, say, a person of Italian background and one from Japan. Even if the two share a common language, they may have trouble communicating because they are likely to have different ways of expressing politeness, conversational involvement, and so forth. The “two-culture” approach proposes that talk between women and men is fraught with potential misunderstanding for much the same reasons that communication across ethnic groups is. (p. 86)

Tannen (1990) argued for the adoption of a sociolinguistic approach to examining gender differences in language use. She noted that such an approach could reveal that boys and girls grow up in different cultures with varying conversational expectations, which may cause friction between them. Tannen (1994) also suggested that linguistic characteristics, such as pacing, pausing, and attitudes toward simultaneous speech, could be attributed to the speaker’s style and their relationship with others. In addition, linguistic strategies can serve as both control and connection maneuvers in family and human interactions.

Similarly, Coates (2015) emphasized the importance of analyzing linguistic variation in relation to social class, speech style, and other non-linguistic variables, including ethnicity, age, and gender. It is believed that cultural differences, such as pressure on girls to be polite and on boys to be competitive, can lead to the adoption of different interaction styles and linguistic choices.

However, difference theorists tend to overlook the power dimension in gender and language research and reduce the relationship to a simplistic cultural distinction. This approach fails to consider social hierarchy and the patriarchal context in which interactions occur. Romaine (2000) criticized traditional linguistic studies for their narrow operationalization of social variables and the failure to explain how power relations are maintained and recreated at the interactional level. The difference model expands the scope of gender and language research but marginalizes the influence of patriarchal and androcentric ideologies on linguistic differences between women and men.

3.1.2 The Dynamic or Social Constructionist model

The previous approaches to gender and language were criticized for their simplistic approach to gender as a social category, which led to the emergence of the social constructionist approach. Under this approach, gender is seen as a social construct, and language is viewed as discourse that produces, rather than reflects, gender as an important social category (Wheaterall, 2005). As a result, the interdisciplinary investigation has become the central approach of the field, which centers around issues such as power relations, gender identity, masculinities, institutional discourses, queer theory, and theoretical approaches like critical discourse analysis (CDA), feminist critical discourse analysis (FCDA), socio constructionism, and poststructuralism, among others (Flowerdew & Richardson, 2017).

CDA is an interdisciplinary approach that aims to understand the relations between discourses and issues such as power relations, ideologies, institutions, inequalities, identities, and social changes (Van Dijk, 2015). FCDA, on the other hand, seeks to examine how gendered assumptions and power asymmetries are produced, sustained, negotiated, and contested in specific communities and discourse contexts (Lazar, 2014). Poststructuralism is another critical approach to gender and discourse that seeks to deconstruct the constructions and structures within discourses (Baxter, 2003).

It is important to note that Butler (1990) argued that gender was performative, which means that people use language and other aspects of behavior to perform a male or female identity,

rather than speaking a certain way because they are male or female. Therefore, research on language and gender has increasingly shifted from gender differences to gendered discourse, which analyzes language within specific situated activities that reflect the importance of culturally defined meanings of linguistic strategies and gender (Kendall & Tannen, 2015).

3.2 Gender and Corpus Linguistics (Corpus studies on Gender)

In the previous section, I described the early stages of the Language and Gender field, which focused on analyzing small amounts of data using a descriptive and qualitative approach. It is worth noting that during this period, most studies focused on how men and women used language, with little emphasis on how they were talked about. Over the last twenty years, CL has played a crucial role in the development of Language and Gender as a field of inquiry. As a result, the research approach has shifted from using small sets of data to analyzing vast amounts of data comprising millions of words using techniques from CL and corpus query tools.

Corpus Linguistics has proven to be a valuable method of linguistic inquiry, not only in Language and Gender but also in fields such as Translation Studies, Discourse Analysis, Applied Linguistics, and Sociolinguistics. Its potential applications are also recognized in areas like forensic linguistics, language teaching, teacher education, media studies, gender studies, and academic and workplace discourses (Hyland et al., 2012).

As Language and Gender is a branch of Sociolinguistics, Baker (2010) emphasizes that CL can assist Sociolinguistics in providing data, computational tools, and procedures to identify language patterns and frequencies. However, some critics argue that CL is too focused on quantifying, which may oversimplify, stereotype, or reinforce prejudice. Nonetheless, Baker (2010) suggests that CL can complement existing paradigms rather than replace them. McEnery and Baker (2015) support this notion and argue that while early corpus analyses tend to be quantitative, as research progresses, the analysis becomes more qualitative and context-driven, relying less on computer software. It is worth noting that even as early as 1992, Leech and Fallon had recognized that corpora could be a source of comparative

information on various social, political, and cultural aspects. Furthermore, they envisioned that corpora could go beyond isolated word frequencies to analyze collocations in context.

In previous research on language and gender, there was a heavy focus on identifying and evaluating how men and women used language, while overlooking how they were represented in language. However, Baker (2012) acknowledged that corpus research could benefit gender studies not only by analyzing how men and women use language, but also by examining their linguistic representation. This approach involves analyzing language from a discourse point of view, which introduces new issues such as power, identity, and diverse theoretical stances. Through discourse analysis, researchers can examine to what extent language is gendered and how gender is constructed, performed, represented, and indexed through discourse (Sunderland, 2004). One key study in this area is Sunderland's work on gendered discourses concerning classrooms, parenting magazines and news reports, and children's magazines, although these studies involved small amounts of data. To identify gendered discourses, Baker (2012) suggests building a corpus and using computer software to identify repetitive linguistic patterns related to gender.

Several studies have relied on corpus linguistics tools to analyze gender and language, including Rayson et al.'s (1997) study, which analyzed 4.5 million words of transcribed speech from the spoken demographic section of the British National Corpus. This analysis considered social differentiation in the use of English vocabulary based on gender, age, social group, and geographical regions. The findings showed that male speakers used more frequently words such as *fucking, fuck, shit, hell, crap, mate*, and *guy*, discourse markers such as *er, aye, right, and okay*, and the articles *a* and *the*, while female speakers rely more frequently on words such as the pronouns, *he, she, I, me, him*, discourse markers such as *mm, really, and oh*, and adjectives such as *lovely* and *nice*. Another interesting finding in this research is the lexical variation by age group; the analysis showed that on one hand, words such as *fucking, my, shit, fuck, okay, me, really, and cos* (because) among others are more common among people whose age is under 35. On the other hand, it was observed that words such as *er, mm, said, says, well, yes, the, and he* among others are more frequently employed by people whose age is over 35 years of age. Rayson's study differentiated and compared linguistic styles of men and women, but its contribution lies in considering age, social group,

and geographical regions as possible explanations for language differences. Other studies in this area focus on analyzing the representation of gender through language, which marks a different approach to researching language and gender.

One of the earliest studies to adopt a gender representation approach and emphasize the importance of word frequency is the “Longman Grammar of Spoken and Written English” (LGSWE), a 40-million-word corpus analyzed by Biber et al. (1999) using authentic data to describe what occurs and how often in different registers. The authors analyzed four registers, including conversation, action, newspaper language, and academic prose. Their findings revealed that in the LGSWE, there were 620 nouns ending with “-man” compared to only 40 that ended with “-woman” per million words. For example, the noun “police” occurred 35 times with “-man” and 5 times with “-woman,” while the noun “business” occurred 40 times with “-man” and had no occurrences with “-woman.” The authors also identified feminine terms without an equivalent masculine term with negative connotations, such as *spinster-bachelor*, *mistress-master*, *tigress-tiger*, and *witch-wizard*. According to Biber et al. (1999), the uneven distribution of masculine and feminine terms reflects the societal bias in the English language, where men hold more positions of power and authority than women. Another reason for this discrepancy is that masculine terms are often used for both men and women, but not vice versa.

Similarly, Romaine (2000) conducted a study on the British National Corpus (BNC) using collocational and frequency analysis. She found only 25 occurrences of “lady of the house,” 3 of “woman of the house,” none of “gentleman of the house,” and 8 of “man of the house.” Romaine (2000) noted that neutral terms are used to perpetuate the inequalities expressed by sex-marked terms, where women are more likely to be referred to as “chairperson” than “chairwoman.” Based on a frequency analysis of these terms in a 3-million-word sample, Romaine found that “chairman” occurred 1,142 times, “chairperson” appeared only 130 times, but “chairwoman” was used only 68 times. Her argument that “neutral terms are used to perpetuate the inequalities expressed in sex-marked terms” is one of the first examples of a constructionist approach to language and gender that considers the role of language in perpetuating gender inequalities.

Overall, the studies discussed here highlight the quantitative and qualitative differences between masculine and feminine terms in the English language and how they reflect and perpetuate gender inequalities in society.

Sigley and Holmes (2002) investigated the relationship between gender and corpus linguistics tools by examining several corpora, including the Brown Corpus of America, the Lancaster-Oslo-Bergen (LOB) corpus, the Wellington Corpus of New Zealand English (WWC), the Freiburg-Brown Corpus of American English (Frown), and the Freiburg-LOB corpus of British English. Their findings indicate that the frequency of sexist suffixes such as -ess and the use of pseudo-polite terms such as lady/ladies have declined since 1960. Additionally, the use of man as a generic term has decreased in written material, and the use of gender premodified terms such as female lawyer has also declined. Interestingly, the frequency of women and woman doubled, while man and men significantly decreased. However, the authors did not provide any hypotheses to account for these findings, but it could be theorized that the increase in the use of woman/women and the decrease in man/men could be due to the fact that the analyzed corpora were composed of published texts written by scholars who are more aware of gendered discourse. To corroborate these findings, it would be suitable to analyze other types of corpora. Another technique used in Corpus Linguistics to examine language and gender is the analysis of collocations.

Hunston (2002) explains that strong collocations become fixed phrases that represent a package of information, and as a result, the assertion behind the phrase is less open to questioning. For example, the collocation between illegal and immigrant has a high mutual information (MI) score and could lead people to accept the idea that moving from one country to another under some circumstances is reprehensible and illegitimate. Similarly, Baker (2010), who revisited Stubbs (1996), examined the collocation working mother and found that it contained the implicature that what mothers do at home is not viewed by society as real work. Both Hunston (2002) and Baker (2010) argue that if collocations and fixed phrases are repeatedly used as unanalyzed units in media discussion and elsewhere, people may come to think about things in those terms. These arguments align with the social constructionist

approach, which stresses the need to research how social power and inequality are enacted, reproduced, legitimated, and resisted through discursive practices. Such analyses are advocated by CDA and FCDA and are necessary for the gender and language field to adopt a gender and discourse approach.

Gesuato (2003) conducted a collocational analysis on the terms woman, man, boy, and girl. Her study revealed that words that frequently co-occur in a language form “constellation” of repeated meanings that lead to the creation of conventional expressions and opinions. Gesuato analyzed the Usbooks, Ukbooks, Time, and Today components of the Cobuild online corpus, with a focus on identifying gendered equivalents for adult and child. She found that both man and woman were commonly associated with discourse domains related to physical attractiveness, age, physical appearance, and family/personal relationships. However, some adjectives were found to be more strongly associated with man than woman, such as size for physical appearance. Gesuato concluded that woman and girl were frequently associated with passivity, physicality, and negativity, while man and boy were associated with activity and cognitivity. Although her study attempted a socio-cultural analysis approach, it was limited to identifying the discourse domains that were most closely associated with the lemmas she employed.

In Romaine’s (2000) collocational analysis of the term doctor with titles such as lady, woman, and female, she found that lady doctor appeared 125 times, woman doctor appeared 20 times, and female doctor appeared 10 times. In contrast, there were no occurrences of gentleman doctor, one occurrence of man doctor, and 14 occurrences of male doctor. Romaine also found that the expressions career woman occurred 48 times, career girl occurred 10 times, and career lady only once, whereas career man appeared 6 times, and career boy and career gentleman did not appear. She attributed these discrepancies to the idea that only men had careers, and women who did so should be marked. Romaine added that it would be odd to call a woman a family woman since women are often assumed to be family-oriented. Such explanations for these findings suggest that there is a need to go beyond the language per se to uncover what lies behind such linguistic representations of both women and men. In Mautner’s (2007) collocational analysis, she employed a keyword and collocation approach to explore stereotypical constructions of age and aging. Using the 500-

million-word Bank of English corpora, Mautner found that the keyword *elderly* collocated more frequently with terms such as *woman*, *women*, and *lady* than with *men* or *gentleman*. Mautner also found that words such as *widow*, *lady*, and *woman* had a high mutual information score in association with *elderly*. Given that *elderly* is primarily associated with discourses of care, disability, and sickness, these findings suggest that feminine terms are commonly associated with discourses of aging and vulnerability.

In a similar study, Pearce (2008) analyzed the representation of the lemmas *MAN* and *WOMAN* in the British National Corpus (BNC), with a focus on how these lemmas behaved as subject and object, and the adjectives associated with them. Pearce claimed that the lemmas' collocates reflect persistent gender differences in the representation of men and women across various domains, such as power and deviance, social categorization, personality and mental capacity, and appearance and sexuality (p. 7).

Pearce's study revealed that verbs denoting physical strength and exercise of power, such as *dig*, *climb*, *jump*, *conquer*, *dominate*, and *lead*, collocated with *man* when used as subjects. No similar collocations were identified with *woman*. Regarding the same lemmas used as objects, Pearce found that verbs denoting actions of legal systems, such as *apprehend*, *arrest*, *convict*, and *sentence*, collocated with *man* but not with *woman*. Verbs denoting victimization by violence, such as *kill*, *wound*, *knife*, and *shoot*, collocated with *man*, while *assault*, *gag*, *rape*, and *violate* collocated with *woman*. Pearce also found that adjectives associated with physical size, power, and wealth, such as *big*, *fit*, *tall*, *great*, *powerful*, and *rich*, collocated with *man* but not with *woman*. However, adjectives associated with marital/reproductive status, nationality, religion, and ethnicity, such as *married*, *childless*, *Catholic*, and *American*, respectively, collocated more with *woman* than with *man*. While both Mautner (2007) and Pearce (2008) obtained interesting results, their studies only identified collocations and fell short in explaining the underlying reasons for such differences in representation. Categorizing the collocates into different domains could allow for a more in-depth analysis. However, such an analysis would require an interdisciplinary team that includes not only linguists but also scholars from sociology, anthropology, and gender studies.

Caldas-Coulthard and Moon's (2010) study is widely cited in the field of gender and language. The researchers analyzed the representation of man, woman, girl, and boy in tabloids (The Sun) and broadsheets in reports and features, using a corpus of around 45 million words for The Sun and 112 million words from the broadsheets. The study revealed several interesting findings.

Man was found to be similarly represented in both corpora. However, The Sun labeled man more frequently as a member of a team (right-hand man, extra man), and he was more likely to be identified as young in The Sun than in the broadsheets. In terms of physical portrayal, The Sun tended to use adjectives related to size, strength, and capacity, while the broadsheets showed more range in their use of adjectives, but these were infrequent in the corpus. The broadsheets also categorized man more frequently as someone with mental capacity (thoughtful, intelligent, cleverest).

In terms of the collocates of woman, The Sun categorized woman as career *woman*, *driver*, *reader*, *cop*, *judge*, *passenger*, *patient*, *teacher*, and *motorist*, among other adjectives, while the broadsheets categorized woman as career woman, *cleaning woman*, *working woman*, *president*, *prime minister*, *lawyer*, *police officer*, and *writer*, among other adjectives. These findings reveal sociolectal and demographic distinctions between the sets of occupations. Woman was found to collocate with adjectives indicating age and marital status in both sub-corpora. The broadsheets showed more collocates related to physical attributes and appearance, including adjectives like *fat*, *tall*, *small*, *naked*, *topless*, *pretty*, and *attractive*.

This study's major contribution is that it investigated how man and woman are represented in written media, allowing researchers to compare findings obtained in different corpora. This approach is much needed to make significant progress in the field, rather than simply confirming earlier research results.

In a more recent study, Moon (2014) examined English adjectives used to describe men and women, with a focus on age. Using the 450-million-word Bank of English (BoE), she investigated the collocates of man and woman when they were also categorized as young, middle-aged, or old. The aim of the research was to identify stereotypical characteristics associated with different ages. Moon noted that "collocates of young and old suggest that

they are not just counterparts in terms of age reference but also in evaluative orientation” (Moon, 2014, p. 7).

To illustrate how collocational pattern identification can reveal evaluative orientations, Moon conducted a preliminary search in the BoE. She found that adjectives such as *inexperienced*, *beautiful*, *fresh*, *attractive*, *trendy*, *single*, *healthy*, *vulnerable*, *pretty*, *talented*, *energetic*, *dynamic*, and *fit* collocated with “young and...,” while adjectives such as *sick*, *tired*, *infirm*, *frail*, *gray*, *fat*, *disabled*, *slow*, *poor*, *weak*, *wise*, *beautiful*, and *ugly* collocated with “old and...”. This analysis showed that most of the adjectives that collocated with “young” were positive and related to physical characteristics or potential, while the adjectives that collocated with “old” were mostly negative.

Regarding the collocations of “young men,” the most common collocates were *handsome*, *nice*, *bright*, *tall*, and *angry*. Other collocates related to the relationship domain (*gay*, *single*, *married*, *lonely*, *bisexual*, etc.) and physical attributes (*tall*, *thin*, *muscular*, *slender*, etc.) occurred frequently. Concerning the collocations of middle-aged women, there were fewer adjectival collocates. Some of these were positive in evaluative orientation, such as *elegant*, *attractive*, *beautiful*, and *healthy*, but there were also negative collocates, such as *single*, *lonely*, *plump*, *fat*, *stout*, and *bored*, that suggested negative traits for women in midlife.

In her research, Moon (2014) aimed to identify stereotypical characteristics associated with gender and age. However, her study had a few limitations. Firstly, she did not distinguish between the singular and plural forms of the words analyzed. Additionally, she did not consider the various text types and genres represented in the Bank of English (BoE). It’s important to note that while the studies I’ve discussed in this section shed light on the different representations of men and women, they may still be subject to criticism. To date, research in this area has only scratched the surface and has yet to comprehensively address whether gendered language use has evolved over time or in different genres. Although creating corpora is a resource-intensive task, even small-scale projects focusing on specific genres could help to rejuvenate the gender and language field, which has been stagnant for some time.

Baker's (2013) study aimed to revitalize the gender and language field by conducting a diachronic analysis of gender-marked language in four different corpora: The Lancaster (BLOB), the Lancaster-Oslo/Bergen corpus (LOB), the Freiberg-Lob (FLOB), and the British English 2006 (BE06). It is worth noting that previous research in this area utilized a corpus-driven approach, which narrowed its focus as patterns emerged. In contrast, Baker's study adopted a corpus-based approach, where researchers explore a predetermined hypothesis. However, Baker (2013) acknowledges that such an approach may miss certain gender-marked words and emphasizes the importance of intuition in selecting the words for investigation. Additionally, Baker recognizes that his research has limitations, including a small corpus size and the reflection of written and published rather than spoken or unpublished English. Therefore, he cautions about making conclusive remarks regarding language change and gender.

Baker (2013) analyzed various terms such as male and female pronouns, man, men, woman, women, boy, girl, as well as gender-related professions and terms of address such as Mr. and Ms. One of the significant findings showed a decrease in the use of male pronouns, whereas female pronouns showed a slight increase. However, despite this fluctuation, there remains a substantial gap between the use of male and female pronouns. In terms of inclusive language, such as him/her, s/he, he/she, he or she, and him or her, the analysis found an increase in their use between 1961 and 1991, but the total usage in 2006 was less than half that of 1991. This suggests that inclusive language strategies may not be becoming popular and may even die out.

Baker's analysis also showed that the use of the term "spokesman" has been consistent over the last 15 years, whereas "spokeswoman" did not occur in the BLOB and LOB corpus, but appeared 8 and 5 times respectively in the FLOB and BE06. "Spokesperson" did not occur in the BLOB and LOB but appeared 2 and 4 times in the FLOB and BE06. Finally, Baker's study examined gendered titles such as Mr., Mrs., Miss, and Ms. He summarizes his findings in the figure below.

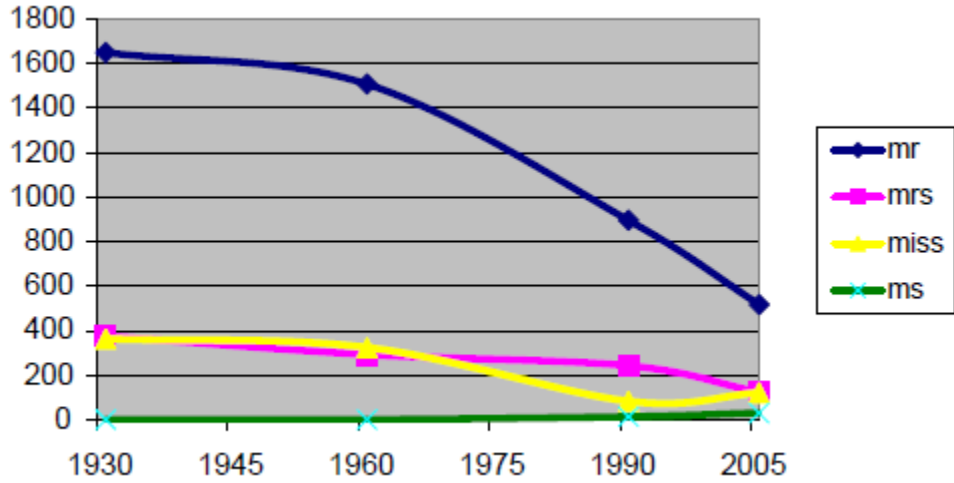


Figure 9. Frequency distribution of gendered titles according to Baker, 2013.

In his study, Baker (2013) analyzed the use of gender-marked language in four different corpora and found that while there was a decrease in the use of male pronouns, the gap between the use of male and female pronouns remained substantial. He also noted that the use of inclusive terms had decreased over time and that the frequency of gendered titles such as Mr., Mrs., Miss, and Ms. had declined as well, possibly due to increased awareness of gender inequality.

This progress in the field of gender and language has come a long way from the early days of introspective and subjective data analysis. The “two-culture” approach, which attributes linguistic differences between men and women to the social roles they enact in different contexts, still dominates much of the current research. However, there has been some movement towards adopting the “socio-constructionist” approach, which seeks to question gendered assumptions that are taken for granted.

In recent years, there has been an incursion of the gender and language field into the natural language processing (NLP) field. Studies addressing misogynistic language and hate speech directed towards women have contributed to expanding and informing both the gender and language field and NLP studies. While these studies may not be traditionally considered part of the gender and language field within NLP, they can still be valuable in addressing gendered issues in language.

3.3 Language and Machine Learning

In this final section, I will describe studies that have utilized Machine Learning to conduct automatic text classification. Specifically, machine learning algorithms were employed to classify files based on keywords and to classify human-annotated comments in a topic classification experiment. It should be restated that in this research, the task at hand is the detection of sexism in general, which could be seen as a form of opinion mining or sentiment analysis. Therefore, I believe that the general label of topic classification is a better fit for the experiments conducted here.

Topic classification has two primary uses in the social sciences: retrieving individual comments and tracing patterns and trends in issue-related activity (Hillard et al., 2008). For this research, the focus is on classifying files and comments based on keywords obtained from corpora related to gender violence, non-gender violence, and the LGBT community. Although not the main focus of this study, it also enables us to identify trends in language use and opinions regarding these matters. Topic classification involves assigning individual documents (comments) to a limited set of categories. It is valuable for its ability to limit research results to documents that closely match the user's interests, compared to less selective-based approaches.

Hillard et al. (2008) suggest that an ideal topic classification system for social sciences should possess four primary characteristics. Firstly, the categories should be able to distinguish and identify the documents' subject matter. Secondly, the categories must be accurate and reflect the document's content. Thirdly, the system should be reliable and capable of classifying documents even if the terminology used is changing. Finally, the system should identify documents that address the topic, even if they are not primarily about that subject.

Most of the studies presented in the following pages focus on identifying and classifying hate speech. Hate speech is defined as any communication that devalues a person or a group based on characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or any other trait (Schmidt & Wiegand, 2017). Although terms such as abusive language, misogynistic language, cyberbullying, and offensive and vulgar language are used

in many studies, for the purpose of clarity, I will use the term "hate speech" as an umbrella term for all these terms.

There is a growing interest in tasks such as classification and automatic detection of hateful online language in the field of natural language processing. To advance this research area, there are academic events held throughout the world that focus on these tasks. SemEval (Semantic Evaluation) is one such event, which is a series of international natural language processing research workshops that evaluate semantic analysis systems and create highly annotated datasets. In these workshops, tasks are assigned, and several teams develop computational semantic analysis systems that are compared to determine which system (model) is more accurate in such tasks (see Basile et al., 2019). Another similar academic event is IberEval, which promotes the development of human language technologies for Iberian languages. In these events, shared tasks are assigned, and each team develops natural language processing systems that are evaluated (see Fersini et al., 2018). These events involve groups of researchers competing to present the best NLP model to solve tasks, and one traditional task in these events is the classification and detection of hate speech, as well as sentiment analysis and opinion mining.

At IberEval 2018, a task was set for researchers to participate in called Automatic Misogyny Identification (AMI). The task involved using datasets of tweets in both English and Spanish, and 32 teams participated in the English language dataset while 24 teams participated in the Spanish language dataset. The training data for this task consisted of 3,251 and 3,307 tweets for English and Spanish, respectively, while the test data comprised 726 and 831 tweets for English and Spanish, respectively. The task had two main subtasks: the first was to discriminate and classify misogynistic tweets from non-misogynistic ones, and the second was related to misogynistic behavior and target classification.

In terms of misogynistic behavior, each tweet had to be classified according to a given taxonomy, and in terms of target classification, each tweet had to be classified as active (directed to somebody in particular) or passive (directed to potential receivers). In this event, Canós (2018) achieved the best results with the Spanish dataset. In his model, the preprocessing stage involved the following steps:

- All letters were converted to lower case.
- Multiple concatenated exclamation marks were replaced by the keyword MULT_EXCLAMATION.
- Multiple concatenated question marks were replaced by the keyword MULT_QUESTION.
- Multiple concatenated exclamation and question marks were replaced by the keyword MIXED_MARKS.
- URLs were replaced by the keyword URL
- Users' mentions were replaced by the keyword USER.
- Misogynistic hashtags were replaced by the keyword MISO_HASHTAG. A hashtag was considered misogynistic if it appeared only in several misogynistic tweets in the training corpora. Misogynistic hashtags in English are those that contain any of the words: *bitch, whore, hoe, cunt, womenare, womensuck*. Misogynistic hashtags in Spanish are those that contain any of the words: *feminanzi, perra*.
- The rest of the hashtags were replaced by the keyword HASHTAG.

(p. 90-91)

Canós (2018) employed the TF-IDF feature extraction in his model, which converted each tweet in the dataset into vectors, allowing identification of the importance of each word for each tweet. For the first subtask, all tweets were used to extract the vocabulary, while for the second subtask, only the misogynistic tweets were utilized for vocabulary extraction. Canós' system classified each tweet as either misogynistic or non-misogynistic, and only the former was categorized based on behavior and target, utilizing the Support Vector Machine (SVM) classification algorithm.

Among the 24 teams that participated in the first subtask using the Spanish language dataset, Canós (2018) presented the most accurate system with a precision score of 0.8147. In the second subtask, he presented the third most accurate system, as shown in Tables 7 and 8.

Language	Accuracy	System Ranking	Team Ranking
Spanish	0.8147	1	1
English	0.7493	23	10

Table 7. Results of the first subtask over the test corpora. Taken from (Canós, 2018)

Language	F _{avg}	F _{category}	F _{target}	System Ranking	Team Ranking
Spanish	0.4328	0.3234	0.5422	6	3
English	0.3263	0.1477	0.5049	20	6

Table 8. Results of the second subtask over test corpora. Taken from (Canós, 2018)

The results obtained with the English and Spanish language datasets varied, and the author suggests that this could be due to the different choice of misogynistic hashtags in the two languages, as well as the possibility that misogynistic language is more prevalent in English than in Spanish. Anzovino et al. (2018) also conducted a study on automatic detection and categorization of misogynistic language in Twitter, using a modified version of the system presented in IberEval 2018. They began by designing a taxonomy to distinguish misogynistic messages and characterized the different types of manifestations among them. This taxonomy was adapted from Poland (2016) and is as follows:

1. Discredit: slurring over women with no other larger intention.
2. Stereotype and Objectification: To make women subordinated of description of women’s physical appeal and /or comparisons to narrow standards.
3. Sexual Harassment and Threats of Violence: to physically assert power over women, or to intimidate and silence women through threats.
4. Dominance: to preserve male control, protect male interests, and exclude women from conversation.
5. Derailing: to justify abuse, reject male responsibility, and attempt to disrupt the conversation to refocus it.

Table 9 displays examples of messages that were categorized according to the taxonomy developed by Anzovino et al. (2018).

Misogyny category	Text
Discredit	I've yet to come across a nice girl. They all end up being bitches in the end #WomenSuck
Stereotype	I don't know why women wear watches, there's a perfectly good clock on the stove. #WomenSuck
Objectification	You're ugly. Caking on makeup can't fix ugly. It just makes it worse!
Sexual Harassment	Women are equal and deserve respect. Just kidding, they should suck my dick
Threats of Violence	Domestic abuse is never okay.... Unless your wife is a bitch #WomenSuck
Dominance	We better not ever have a woman president @WomenSuckk
Derailing	@yesallwomen wearing a tiny skirt is "asking for it". Your teasing a (hard working, taxes paying) dog with a bone. That's cruel. #YesAllMen

Table 9. Examples of text for each misogyny category. Taken from Anzovino et al. (2018)

To label the data, Anzovino et al. (2018) relied on a previous study that had identified influential features for identifying hateful speech. These features included words such as *xist*, *sexi*, *ka*, *sex*, *kat*, *exis*, *exis*, *xis*, *exi*, *xi*, *bitc*, *ist*, *bit*, *itch*, *itc*, *fem*, *ex*, *bi*, *irl*, *wom*, and *girl*. They also added other words that reflected different categories of misogyny, representing potential actions against women. A gold standard data set was labeled by two annotators, and discrepancies were resolved by a third annotator. This data set was used as a quality control to compare with the second data set, which was labeled through the CrowdFlower platform. The data set comprised 4,454 tweets, balanced between misogynous and non-misogynous tweets.

In the automatic categorization of misogynous tweets, different features were used to identify the best results. These included N-grams, Bag-of-POS, stylistic linguistics features such as the length of comments, number of adjectives, and number of users mentions, and embeddings. These features were run individually and then used in combination with linguistic features to achieve the highest accuracy. The authors employed several classification algorithms, including Support Vector Machine, Random Forest, Naïve Bayes, and Multi-Layer Perceptron Neural Network, which are considered effective text categorization algorithms.

The findings indicate that the Token n-grams in combination with the Support Vector Machine outperformed the other classifiers.

Features combination	RF	NB	MPNN	SVM
Char n-grams	0.7930	0.7508	0.7616	0.7586
Token n-grams	0.7856	0.7432	0.7582	0.7995
Embedding	0.6893	0.6834	0.7041	0.7456
Bag-of-POS	0.6064	0.6031	0.6017	0.5997
Linguistic	0.5831	0.6098	0.5963	0.5348
Char n-grams, Linguistic	0.7890	0.7526	0.7443	0.7627
Token n-grams, Linguistic	0.7739	0.7164	0.7593	0.7966
Embedding, Linguistic	0.6830	0.5878	0.7014	0.6556
Bag-of-POS, Linguistic	0.6069	0.6286	0.5997	0.5799
All Features	0.7427	0.7730	0.7613	0.7739

Table 10. Accuracy performance for misogynistic language identification. (Anzovino et al., 2018)

Plaza-del-Arco et al. (2019) conducted a SemEval workshop study aimed at detecting and classifying hate speech against immigrants and women. To this end, they used a dataset (HatEval dataset) consisting of tweets in Spanish and English, collected from July to September 2018. The data was obtained by monitoring potential victims of hate accounts, downloading the history of identified haters, and filtering Twitter streams with keywords, hashtags, and stems. The researchers used derogatory words and highly polarized hashtags as keywords to collect the corpora, which they believed would help distinguish between hate speech, offensiveness, and stance. The most common words in the corpus included *migrant*, *refugee*, *#buildthatwall*, *bitch*, and *hoe* in English, and *inmigra-*, *arabe*, *sudaca*, *puta*, *callate*, and *perra* in Spanish. The HatEval dataset comprises 19,600 tweets, with 13,000 in English and 6,600 in Spanish. Of these, 9,091 targeted immigrants, while 10,509 targeted women.

During the annotation process, each tweet was labeled with a 0 if it was non-hateful and a 1 if it contained hate speech (HS). Another field was used to indicate the recipient of the tweet (TR); tweets directed at an individual were marked with a 1, while those aimed at a group received a 0. A third field, aggressiveness (AG), was used to indicate whether a tweet was aggressive (1) or not (0). To evaluate their model, Plaza-del-Arco et al. (2019) considered only the tweet text and the HS field. The sentences were preprocessed and converted into

feature vectors using the term frequency (TF) statistical feature. The researchers employed a voting ensemble classifier that combined predictions from three classifiers: Logistic Regression (LR), Decision Tree (DT), and Support Vector Machine (SVM).

In this study, Plaza-del-Arco et al. obtained better results with tweets in Spanish, ranking 14th out of 41 participants. The evaluation metrics included accuracy (Acc), precision (P), recall, and F1-score (F1). Table 11 shows the results obtained by Plaza-del-Arco with the Spanish dataset. However, the researchers did not achieve good results with the English dataset. They attribute this discrepancy to the fact that Spanish and English behave differently when it comes to using xenophobic words. For example, the word “puta” in Spanish can be used to offend someone, but it can also express surprise (¡Put a madre!). Furthermore, the researchers argue that the language used to insult women and immigrants is different and requires different NLP systems.

User name (r)	Test			
	P	R	F1	Acc
franco1q2 (1)	0.734	0.741	0.73	0.731
luiso.vega (2)	0.729	0.736	0.73	0.734
fimplaza (14)	0.707	0.713	0.707	0.711
SVC baseline (21)	0.701	0.707	0.701	0.705
DA-LD-Hildesheim (40)	0.493	0.494	0.493	0.511

Table 11. System Results per team in subtask A of the HatEval task in Spanish. Taken from (Plaza-del-Arco et al., 2019)

In the three previous studies, researchers used Spanish tweets to automatically classify and identify misogynistic content. However, García-Díaz et al. (2021) conducted a more extensive investigation that not only aimed to detect misogynistic tweets on Twitter, but also to classify messages that harass women in Spanish from both Spain and Latin America. In addition, they focused on identifying tweets related to violence against women and general traits associated with misogyny. This new approach allowed for a cultural analysis of misogyny in the Spanish-speaking world.

The corpus compiled in García-Díaz’s research was labeled as containing misogyny and was divided into three subsets:

- I. Violence Against Relevant Women (VARM). This data subset contained tweets directed to Greta Thunberg and some Spanish politicians
- II. European Spanish vs Latin American Spanish (SELA). Due to differences observed in previous Spanish language datasets which are suspected of compromising the accuracy of classification, this research opted to compile tweets from Latin American and European people. The tweets from Latin America were obtained from latitude: -0.1596997 , longitude: -78.452125313 , radius: 1,500 km whereas the Spanish tweets were collected from latitude: 40.416705 , longitude: -3.703583 , radius: 520 km. See Figure 10 to identify the location.
- III. Discredit, Dominance, Sexual Harassment, and Stereotype (DDSS). To compile this subset, keywords were used to identify tweets that relate to discredit, dominance, sexual harassment, or stereotypes.

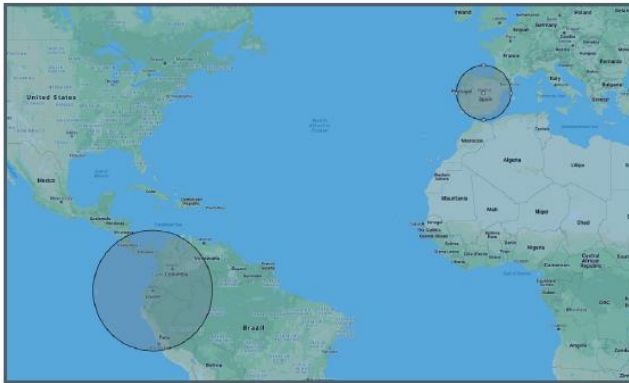


Figure 10. Locations where the tweets of the SELA corpus were compiled. Taken from (García-Díaz et al., 2021)

The Miso-Corpus-2020 retrieved a total of 32,969 tweets, which were then preprocessed to obtain 7,682 tweets related to misogyny. It should be noted that some tweets may appear in multiple subsets, which explains the discrepancy between the total number of tweets and the sum of the tweets in each subset. The distribution of tweets among the three subsets is shown in Table 12.

Name	Misogyny	Not misogyny	Mean of annotations
VARW	2094	2094	2.1529
SELA	2081	2081	2.3076
DDSS	1665	1665	2.1595
MisoCorpus-2020	3841	3841	2.2240

Table 12. *MisoCorpus Classification. Taken from (García-Díaz et al., 2021)*

In the preprocessing stage, the following procedures were taken:

- All tweets were converted to lowercase
- Blank lines and HTML tags were removed.
- Mentions were removed.
- Misspellings were fixed.
- Repeated symbols were removed
- The hashtag #feminista was converted to the word feminist.

In this experiment, both sentence embedding features and linguistic features were utilized. After the tweets were preprocessed, sentence embeddings were generated by computing the Average of Word Embeddings (AWE). While word embeddings aim to capture the meaning of individual words in a sentence, sentence embeddings aim to encode the meaning of the entire sentence, allowing for a better understanding of context and the intended meaning. Additionally, sentence embeddings enable sentences to be clustered based on similarity, as demonstrated by Reimers and Gurevych (2019).

The linguistic features used in this classification were extracted using the Linguistic Inquiry and Word Count (LIWC) tool, which can identify emotional, cognitive, and structural components present in verbal and written speech samples. A total of 253 different linguistic features were grouped into ten categories, including Figures of speech (FSE), Pragmatics (PRA), Morphological features (MOR), Grammar and spelling mistakes (ERR), Part of Speech (PoS), Punctuation and symbols (SYM), Twitter features (TWI), Sociolinguistics (SLI), Topics (TOP), Sentiment lexicon (SEN), and Stylometry (STY).

To assess the predictability of these features, the authors conducted experiments using the WEKA platform, employing the Random Forest (RF), Sequential Minimal Optimization (SMO), and Linear Support Vector Machines (LSVM) classification algorithms. Each model

was trained using a 10-fold cross-validation and evaluated based on its accuracy and standard deviation. We established a baseline by applying a model based on the Bag of Words (BoW) technique. The results of this initial experiment are presented in Table 13.

Feature set	RF		SMO		LSVM	
	ACC	SD	ACC	SD	ACC	SD
VARW	78.930	2.214	78.524	2.241	80.053	1.613
SELA	76.967	2.728	76.918	2.141	78.476	1.625
DDSS	74.734	1.795	74.003	2.568	77.698	1.742
MisoCorpus-2020	76.215	0.972	73.798	1.038	77.060	1.443

Table 13. Accuracy and standard deviation of the baseline model (BoW). (Taken from García-Díaz et al., 2021)

The LSVM classification algorithm outperformed the others in all subsets and the entire MisoCorpus-2020. After establishing the baseline, the sentence embedding model was compared with the LSVM model. Table 14 demonstrates how the model improved with the sentence embedding feature. The LSVM classifier achieved the best result on the MisoCorpus-2020, as shown in Table 13 with an accuracy of 77.06. However, when using word embeddings, the accuracy increased to 80.825, as displayed in Table 14.

Feature set	RF		SMO		LSVM	
	ACC	SD	ACC	SD	ACC	SD
VARW	82.092	1.224	84.886	1.276	84.480	1.951
SELA	81.307	2.111	82.100	1.733	81.859	1.973
DDSS	79.063	2.009	81.360	1.586	81.148	1.745
MisoCorpus-2020	77.232	1.497	81.020	1.505	80.825	0.844

Table 14. Accuracy and standard deviation of AWE (Average Word Embedding) for VARW, SELA, DDSS, and MisoCorpus-200 evaluated with a 10 cross-fold validation. (Taken from García-Díaz et al., 2021)

All algorithms demonstrated improvement; however, SMO yielded the highest accuracy for both the VARW subset and the MisoCorpus-2020. In an effort to further improve results, the corpus was evaluated using linguistic features, as shown in Table 15.

Feature set	RF		SMO		LSVM	
	ACC	SD	ACC	SD	ACC	SD
VARW	81.112	1.040	82.403	2.042	82.283	1.535
SELA	81.740	1.949	80.057	2.067	81.115	1.865
DDSS	77.613	2.190	77.976	2.056	79.245	1.078
MisoCorpus-2020	79.237	1.757	78.938	1.631	79.263	1.658

Table 15. Accuracy and standard deviation of LF feature for VARW, SELA, DDSS, and MisoCorpus-200 evaluated with a 10 cross-fold validation. (Taken from García-Díaz et al., 2021)

The authors found that incorporating linguistic features led to a decrease in accuracy not only for the entire MisoCorpus-2020 but also resulted in increased standard deviation in some cases. This sacrifice in accuracy was deemed acceptable in exchange for greater interpretability of the results. Finally, the corpus was evaluated using both sentence embeddings and linguistic features, and the results are presented in Table 16.

Name	RF		SMO		LSVM	
	ACC	SD	ACC	SD	ACC	SD
VARW	82.092	1.224	84.886	1.276	84.480	1.951
SELA	81.307	2.111	85.175	1.450	83.734	1.915
DDSS	78.912	2.690	81.208	1.766	80.755	1.113
MisoCorpus-2020	79.302	1.497	85.175	1.450	82.882	1.291

Table 16. Accuracy and standard deviation of AWE and LF feature for VARW, SELA, DDSS, and MisoCorpus-2020 when evaluated with ten cross-fold validation. Taken from (García-Díaz et al., 2021)

Table 17 provides a comprehensive view of the accuracy of all subsets when utilizing each feature individually as well as in combination, with SMO achieving the highest accuracy in both individual and combined features not only in the subsets but also in the entire MisoCorpus-2020.

Classifier	Model	VARW	SELA	DDSS	SMC-2020
RF	BoW	78.930	76.967	74.734	76.215
	AWE	82.092	81.307	79.063	77.232
	LF	81.112	81.740	77.613	79.237
	AWE+LF	82.092	81.307	78.912	79.302
SMO	BoW	78.524	76.918	74.003	73.798
	AWE	84.886	82.100	81.360	81.020
	LF	82.403	80.057	77.976	78.938
	AWE+LF	84.886	85.175	81.208	85.175
LSVM	BoW	80.053	78.476	77.698	77.060
	AWE	84.480	81.859	81.148	80.825
	LF	82.283	81.115	79.245	79.263
	AWE+LF	84.480	83.734	80.755	82.882

Table 17. Comparison of all the subsets when executed with each feature and when these were combined. Taken from (García-Díaz et al., 2021)

Table 17 displays the combined accuracy of SMO using both the SELA subset and the MisoCorpus-2020, which yielded an accuracy of 85.175. The authors posit that these findings suggest that the Average Word Embedding and linguistic features are complementary and do not contradict each other. Additionally, the results obtained by combining both features outperformed those obtained by the baseline model. Using the information gain measure, the authors identified the linguistic features that contributed most significantly to class prediction. The Sentiment Lexicon (SEN) feature, which encompasses offensive language, yielded the highest information gain at 0.126654. The PoS-words-feminine percentage, which refers to the number of grammatically feminine words, was the second most relevant feature in classification, with an information gain of 0.083707. Figure 11 displays the linguistic features with the highest information gain.

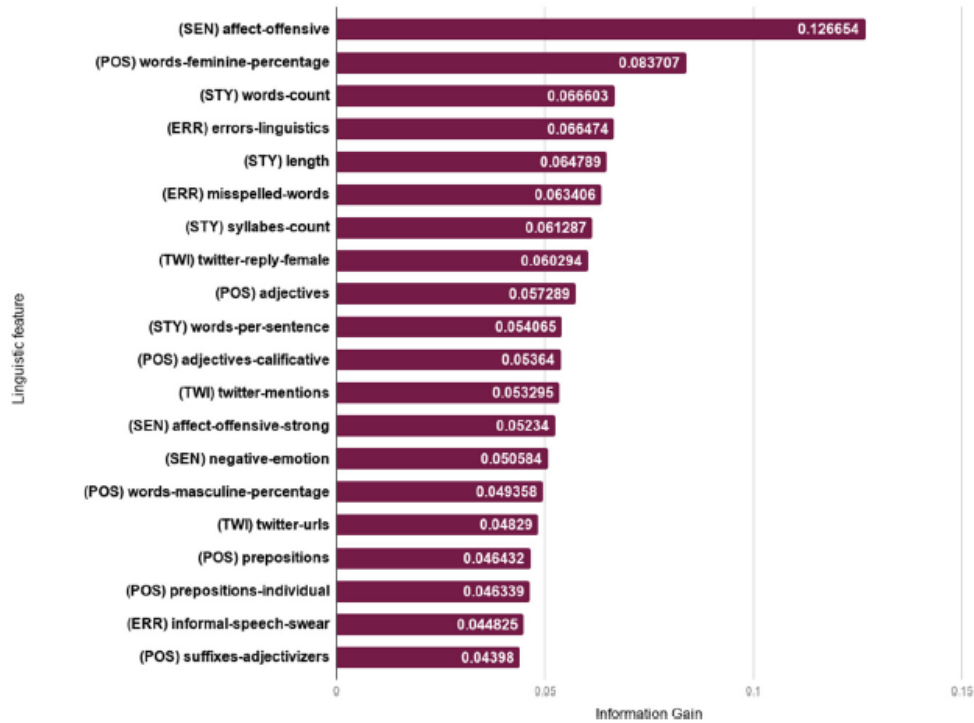


Figure 11. Linguistic features with the highest information gain. Taken from (García-Díaz et al., 2021)

Concerning the SELA corpus, this was divided into two subsets. The training set consisted of 2,976 tweets from Spain, out of which 1,488 were labeled as misogynistic and 1,488 as non-misogynistic. The test set comprised 984 tweets from Latin America, with 492 tweets labeled as misogynistic and 492 as non-misogynistic. It is worth noting that some tweets were excluded from the experiment due to their challenging nature, which accounts for any disparity between the number of tweets used in this experiment and the total number of tweets in the SELA corpus. Table 18 presents the results of this experiment.

Feature set	RF	SMO	LSVM
AWE	57.3171	61.0772	61.3821
LF	69.1057	69.7154	72.4593
AWE+LF	69.4106	75.1016	73.1707

Table 18. Comparison of the accuracy between European Spanish and Latin American Spanish when applying AWE, LF and AWE+LF

SMO achieved the highest accuracy when AWE and LF were combined. AWE yielded the worst results with the three algorithms. When executed individually, LF showed a significant

improvement when compared to AWE alone. According to the authors, these results suggest that misogyny shares common features from different backgrounds, including linguistic features and other features related to the usage of Twitter.

In the study that I have just described, the authors classified tweets from three different corpora and the results showed what features provide the highest accuracy. Yet, several issues were identified. First, it seems that some tweets were included in more than one dataset; since this could influence the results and mislead the comparisons that were made, it would be necessary not to include the tweets in some datasets if these are already included in another dataset. Another issue is the lack of consistency in the annotation; to avoid such inconsistency, more annotators could be included to have a better consensus regarding how the tweets are labeled. One major contribution of this study is the comparisons of how hate speech is represented in Spanish from Europe and Spanish from Latin America; however, instead of using a part of the SELA (Spanish) tweets as a training set and some part (Latin America) as the test set, both sets should be executed with both features and see what the results are. Furthermore, it would be interesting to see the linguistic features with more information gained in both the Spanish and Latin American subsets.

In this section, I have presented studies that have sought to identify hate speech in different corpora using data in Spanish. The anonymity afforded in online social networks allows the proliferation of hate speech directed to women and immigrants among other communities. Given this, methods to automatically detect hate speech are needed; therefore, the natural language processing and machine learning communities have engaged in developing models to serve these purposes. For a comprehensive review of the state of the art of how hate speech has been researched and what techniques and languages have been employed see Poletto et al. (2020). From a purely linguistic perspective, identifying hate speech allows us to have a better insight into how these communities are represented in language.

In this chapter, I have described how the language and gender study fields have developed and how CL has made inroads within the intersection of both. I have also presented studies that show how misogynistic language has been researched in the NLP field. In the following

chapter, I will describe the procedures taken to carry out the data collection process and how such data was analyzed.

4 Methodology

This chapter is divided into three sections, each of which describes a procedure I used to collect and analyze data.

The first section explains how I identified adjectives and verbs that collocated with the lemmas “hombre” and “mujer” using the News on the Web (NOW) corpus. I established parameters to identify the most relevant collocates and used the Supersenses framework and ADESSE classification to categorize the adjectives and verbs.

The second section describes how I built the YOUTUBE corpus and the issues I faced with this type of corpus. I also explain how I used the AntConc query tool to identify keywords for each of the corpora, which were later used in automatic classification experiments.

In the final section, I describe the first experiments that took into account keywords derived from the corpora built for this research, as well as the violentómetro classification. I provide details on how these keywords were used in WEKA software to perform automatic classification of videos. Additionally, I explain how techniques from Machine Learning were used in the final experiments, including classification of texts (comments) using the string-to-word vector filter. Table 19 summarizes the experiments discussed in this section.

Experiments	Corpus
Collocational analysis of the lemmas MAN and WOMAN	News on the Web
Automatic Text Classification (Violentómetro and keywords as features)	Viomujdis, Viogendis, LGBTdis (videos used as texts)
Automatic Text Classification based on keywords as features (<i>Word to string vector</i>)	Viomujdis, Viogendis, LGBTdis (Youtube comments)

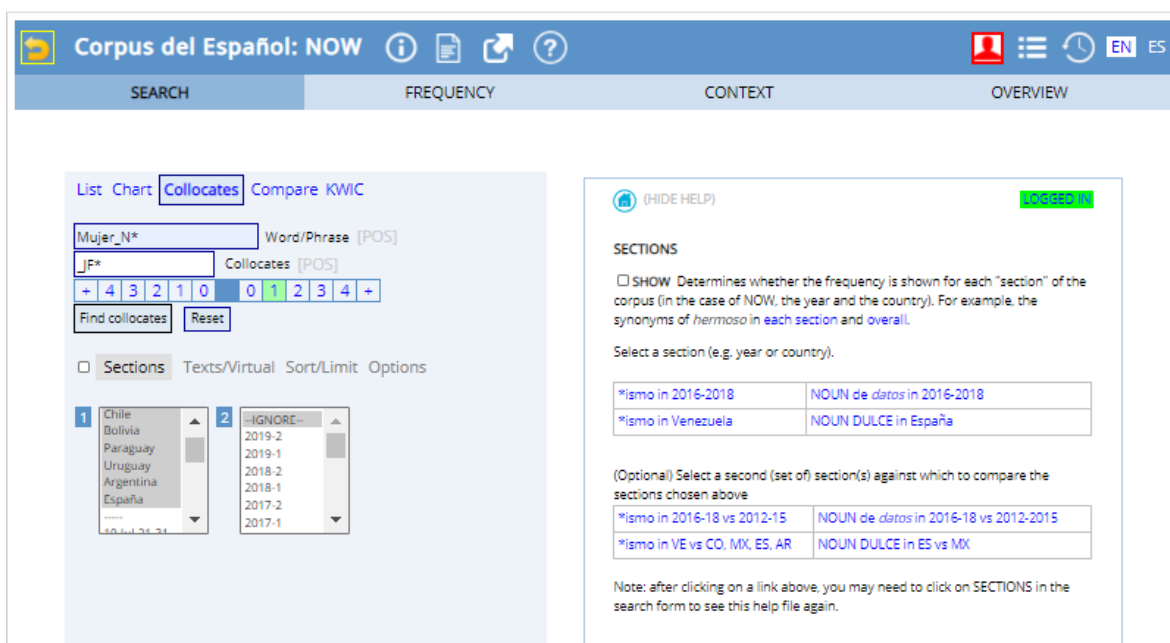
Table 19. Experiments in this research study

4.1 Collocational Analysis and The NOW Corpus

This section provides a detailed account of the procedure used in the first experiment to collect data, which focused on identifying adjectival and verbal collocations of the lemmas hombre ‘man’ and mujer ‘woman’. The experiment involved analyzing the NOW corpus to identify adjectives and verbs that collocated with these lemmas.

Initially, adjectival collocations were searched by using a window span of one space to the right and left of the lemma. This approach enabled the identification of collocates located immediately before and after the node word. The search was conducted on web-based news articles and newspapers from 2012 to 2018 in Latin American countries and Spain for both “woman” and “women” as well as the lemmas “man” and “men”. Additionally, adjectives with a mutual information score above 3 were retrieved.

Figure 12 displays the interface of the NOW corpus and the parameters entered to find collocates to the right of the lemmas.



The screenshot shows the 'Corpus del Español: NOW' interface. The top navigation bar includes 'SEARCH', 'FREQUENCY', 'CONTEXT', and 'OVERVIEW'. The search parameters are as follows:

- Word/Phrase [POS]: Mujer_N*
- Collocates [POS]: JF*
- Window span: + 4 3 2 1 0 0 1 2 3 4 +
- Buttons: Find collocates, Reset
- Sections: Texts/Virtual Sort/Limit Options
- Section 1: Chile, Bolivia, Paraguay, Uruguay, Argentina, España
- Section 2: IGNORE--

The 'SECTIONS' panel on the right includes a 'SHOW' checkbox and a table of section-specific collocates:

SECTIONS	
<input type="checkbox"/> SHOW	Determines whether the frequency is shown for each "section" of the corpus (in the case of NOW, the year and the country). For example, the synonyms of <i>hermoso</i> in each section and overall.
Select a section (e.g. year or country).	
*ismo in 2016-2018	NOUN de <i>datos</i> in 2016-2018
*ismo in Venezuela	NOUN DULCE in España
(Optional) Select a second (set of) section(s) against which to compare the sections chosen above	
*ismo in 2016-18 vs 2012-15	NOUN de <i>datos</i> in 2016-18 vs 2012-2015
*ismo in VE vs CO, MX, ES, AR	NOUN DULCE in ES vs MX

Note: after clicking on a link above, you may need to click on SECTIONS in the search form to see this help file again.

Figure 12. Corpus NOW interface. Parameters to search the adjectival collocations.

During this stage, the same parameters were used to search for every lemma. The results for the adjectival collocations of the lemma *mujer* ‘woman’ are displayed in Figure 13, with the collocations having the highest mutual information (MI) score appearing at the top of the list.

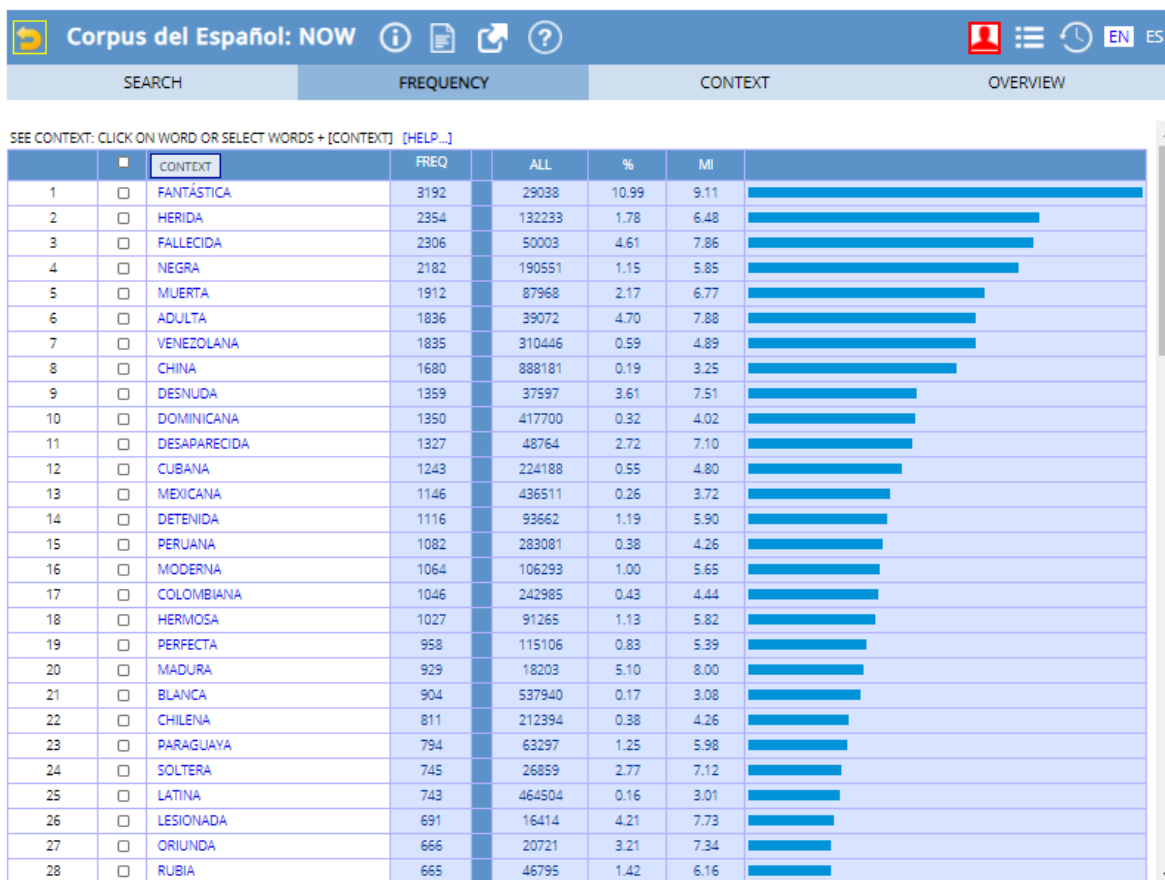


Figure 13. Adjectival Collocations of the lemma *MUJER* (WOMAN)

After obtaining the results in the NOW corpus, users can access the sentences from web-based newspaper articles where these collocations were used. This allows for an analysis of how the nodes and their collocations were used in context, as shown in Figure 14.

Corpus del Español: NOW

SEARCH FREQUENCY CONTEXT OVERVIEW

12-2-2013-1, 2013-2, 2014-1, 2014-2, 2015-1, 2015-2, 2016-1, México, Guatemala, Honduras, Nicaragua, El Salvador, Costa Rica, Panamá, Cuba, Puerto Rico, República, Ecuador, Perú, Chile, Bolivia, Paraguay, Uruguay, Argentina, España, 2017-1, 2017-2, 2016-2, 2018-1

200 500 1000

CONTEXT [?] SAVE LIST CHOOSE LIST CREATE NEW LIST [?] SHOW DUPLICATES

CONTEXT	A	B	C
Independiente de Hidal	A	B	C
La Prensa de Honduras	A	B	C
Semana.com	A	B	C
El Universal	A	B	C
Informador.mx	A	B	C
En Mayúscul	A	B	C
Teletrece	A	B	C
eju.tv	A	B	C
eju.tv	A	B	C
Diario EL PAIS Uruguay	A	B	C
Diario EL PAIS Uruguay	A	B	C
Diario Río Negro	A	B	C
El Sol de México	A	B	C
El Horizonte	A	B	C
El Siglo de Torreón	A	B	C
Reporte Indigo	A	B	C
La Voz de Michoacán	A	B	C
ELIMPARCIAL.COM	A	B	C
Fandango.lat	A	B	C
La Hora	A	B	C
Prensa Libre	A	B	C
La Tribuna.h	A	B	C
La Tribuna.h	A	B	C
La Prensa Gráfica	A	B	C
Fandango.lat	A	B	C
La Nación Costa Rica	A	B	C
En Mayúscul	A	B	C

Figure 14. The collocates and its collocations in context.

A total of 1,586 adjectival collocations were obtained during the first stage. To conduct a linguistic analysis and uncover how men and women are linguistically represented in the Spanish-speaking press, these adjectives needed to be classified. Various Spanish classification taxonomies were considered, but due to the significant variation in Spanish adjective classification based on syntactic and semantic features, an English classification taxonomy was adapted. The SuperSenses Taxonomy (shown in Table 4) was used to classify all the adjectives, enabling a more comprehensive analysis. Figure 15 displays some of the results of this classification in column J.

	A	B	C	D	E	F	G	H	I	J	AK
1	numero	Corp	Busqueda no	hombre -m	Colloc	Venta	Fecha	NU	CONTEXT	Taxonomia	MI
2	1	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	1	ADÚLTERA	Behavior	10.13
3	2	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	2	IMPORTUNA	Behavior	9.61
4	3	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	3	LUCHONA	Behavior	9.44
5	4	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	4	FANTÁSTICA	Behavior	8.72
6	5	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	5	CINCUENTONA	Temporal	8.61
7	6	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	6	MENOPÁUSICA	Body	8.6
8	7	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	7	HACENDOSA	Behavior	8.53
9	8	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	8	AFROAMERICANA	Ethnicity	8.41
10	9	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	9	FRÍGIDA	Body	8.11
11	10	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	10	SUMISA	Behavior	8.06
12	11	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	11	OBESA	Body	7.94
13	12	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	12	CORPULENTA	Body	7.86
14	13	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	13	DISCAPACITADA	Body	7.85
15	14	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	14	VERRACA	Behavior	7.85
16	15	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	15	COQUIMBANA	ethnicity	7.8
17	16	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	16	SEMIDESNUDA	Body	7.79
18	17	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	17	EBRIA	Body	7.77
19	18	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	18	CORAJUDA	Behavior	7.76
20	19	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	19	SORDOMUDA	Body	7.76
21	20	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	20	AGUERRIDA	Behavior	7.63
22	21	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	21	MADURA	Body	7.59
23	22	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	22	INDEFENSA	Victim	7.46
24	23	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	23	AMARGADA	Feeling	7.42
25	24	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	24	ADULTA	Temporal	7.37
26	25	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	25	FALLECIDA	Body	7.37
27	26	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	26	BARBUDA	Body	7.36
28	27	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	27	LESIONADA	Body	7.36
29	28	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	28	NONAGENARIA	Temporal	7.21
30	29	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	29	NIGERIANA	Ethnicity	7.17
31	30	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	30	TREINTAÑERA	Temporal	7.08
32	31	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	31	MALHERIDA	Victim	7.06
33	32	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	32	CAUCÁSICA	Ethnicity	7.04
34	33	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	33	DESNUDA	Body	7.04
35	34	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	34	NEIVANA	Ethnicity	7.02
36	35	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	35	VOLUPTUOSA	Body	7.02
37	36	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	36	SEXAGENARIA	Temporal	7
38	37	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	37	DESVALIDA	Body	7
39	38	NOW	Mujer_N	Mujer	_JF*	1D	19/10/2018	38	BOLEACÉTICA	MI	6.98

Figure 15 Classification of the adjectives based on the SuperSenses taxonomy.

The procedure for classifying the verbs was similar to the adjectives, with the only difference being the window span of one space to the right. Additionally, each node was searched for verbs in present and past tenses. Only the verbs with a MI above 3 were selected. Figure 16 depicts the interface where the lemma “woman” was entered, and the past tense verbs were filtered.

Figure 16 shows the search interface for 'Mujer_N*' with the following parameters:

- Word/Phrase: `Mujer_N*`
- Collocates: `vPRET (hablé)`
- Filter: `_VIS*`
- Sections: Uruguay, Argentina, España, 19-Jul-21-31, 19-Jul-11-20, 19-Jul-01-10
- Ignore: 2019-2, 2019-1, 2018-2, 2018-1, 2017-2, 2017-1

The 'SECTIONS' panel includes the following information:

SECTIONS

SHOW Determines whether the frequency is shown for each "section" of the corpus (in the case of NOW, the year and the country). For example, the synonyms of *hermoso* in each section and overall.

Select a section (e.g. year or country).

*ismo in 2016-2018	NOUN de datos in 2016-2018
*ismo in Venezuela	NOUN DULCE in España

(Optional) Select a second (set of) section(s) against which to compare the sections chosen above

*ismo in 2016-18 vs 2012-15	NOUN de datos in 2016-18 vs 2012-2015
*ismo in VE vs CO, MX, ES, AR	NOUN DULCE in ES vs MX

Note: after clicking on a link above, you may need to click on SECTIONS in the search form to see this help file again.

Figure 16. Corpus NOW interface. Parameters to search the verbal collocations

After entering the specified parameters, the collocates with MI above 3 were retrieved. In Figure 17, the verbs that collocated with the lemma woman are presented, along with the frequency of each verb's occurrence with the lemma. It is important to restate that the MI score highlights verbs that occur exclusively with the lemmas.

Figure 17 displays the results of the search for 'Mujer_N*' in the FREQUENCY tab. The table lists 28 collocates with their respective frequencies, percentages, and MI scores.

	CONTEXT	FREQ	ALL	%	MI
1	BOCINÓ	1	5	20.00	10.22
2	DESHABITÓ	1	8	12.50	9.54
3	AMAESTRÓ	1	13	7.69	8.84
4	NARCOTIZÓ	2	26	7.69	8.84
5	EMBADURNÓ	4	63	6.35	8.57
6	CONTRABANDEÓ	3	65	4.62	8.11
7	GRUÑÓ	6	140	4.29	8.00
8	EMPALIDECIÓ	1	26	3.85	7.84
9	SENTISTEIS	1	26	3.85	7.84
10	POLOLEÓ	2	53	3.77	7.82
11	YACIERON	1	29	3.45	7.69
12	TACONEÓ	1	32	3.13	7.54
13	MONETIZÓ	1	32	3.13	7.54
14	TIMÓ	4	149	2.68	7.32
15	ESCULCÓ	1	40	2.50	7.22
16	ERUCTÓ	1	42	2.38	7.15
17	VOLANTEÓ	4	176	2.27	7.08
18	CASTRÓ	6	269	2.23	7.06
19	TAJEÓ	1	46	2.17	7.02
20	FINGIÓ	81	4072	1.99	6.89
21	CURVÓ	1	51	1.96	6.87
22	FORCEJEÓ	62	3192	1.94	6.86
23	AGONIZÓ	18	930	1.94	6.85
24	RAPTÓ	20	1149	1.74	6.70
25	ENSILLÓ	1	60	1.67	6.64
26	ACOMPASÓ	1	61	1.64	6.61
27	LEVITÓ	1	61	1.64	6.61
28	AMAMANTÓ	7	441	1.59	6.57

Figure 17. Verbal Collocations of the lemma MUJER (WOMAN)

The interface once again provides users with access to the text where the collocations appeared, enabling them to examine the context of these occurrences and further the linguistic analysis. Refer to Figure 18 for a visual representation.

The screenshot shows the 'Corpus del Español: NOW' interface. At the top, there are navigation tabs: SEARCH, FREQUENCY, CONTEXT, and OVERVIEW. Below the tabs, there is a search bar and a list of results. The results are displayed in a table with columns for date, source, and text snippets. The text snippets contain the lemma 'MUJER' and its collocations, such as 'esta mañana. # El bebé fue robado de la clínica Centra, donde la mujer fingió ser enfermera y se los pidió a sus padres para presuntamente', 'por desgracia algunos de los proyectiles penetraron en un hogar aledaño, donde una mujer pereció víctima de las balas perdidas, y una niña', 'presunto delincuente. # Además, unos proyectiles penetraron en un hogar aledaño donde una mujer pereció por "balas perdidas" y una niña', 'para escapar. # Desde San Francisco de Macorís, en el 2016, una mujer fingió su secuestro y exigió a su hija y a una hermana que residen en', 'comerciantes hicieron esfuerzos para evitar que los agentes cumplan con su labor. Incluso una mujer fingió desmayar se en el preciso instante', 'contra ldlib, el último bastión opositor a Damasco en el país, y una mujer pereció en Hama por disparos de grupos de la oposición. # El Obse', 'de otras chicas en su celular. # Ocurrió a mediados de mayo, la mujer fingió estar "desesperada por sexo" y sugirió a el novio que hicieran', 'recuperó el dinero. # Hace menos de una semana, en Piura, una mujer fingió el robo de 4 600 soles que había sido recaudado en un puesto', 'hacia su casa. Los días previos a el encuentro con la adolescente, la mujer fingió estar embarazada en sus redes sociales y con sus vecinos. In', 'la Tinka cayó en contradicciones y su nerviosismo la delató. Divincri determinó que la mujer fingió robo para apoderar se de S/4, 600. # Se h', 'Hora Local Mujer se autosecuestro porque tenía mucha presión en su trabajo # Mujer fingió violento secuestro y aparentaba haber sido tor', 'de el acusado y los guardó en un placard. # Más tarde, la mujer fingió que iba a buscar leña a el monte, dejando a la menor sola', 'de el acusado y los guardó en un placard. Más tarde, la mujer fingió que iba a buscar leña a un sector montañoso, en inmediaciones de la', 'de el acusado y los guardó en un placard. # Más tarde, la mujer fingió que iba a buscar leña a el monte, dejando a la menor sola', 'canal 02). PE. 73047889 Surco: mujer envenenó a sus dos hijas e intentó quitar se la vida # Vecinos señalan que', 'abandonó el hogar Mujer envenenó a sus dos hijas. América Noticias # Una mujer envenenó a sus dos hijas y luego intentó quitar se la vida', 'que tres días antes discutió con su pareja y que él abandonó el hogar Mujer envenenó a sus dos hijas. América Noticias # Una mujer envene', '# Por Soy502 # 27 de marzo de 2019, 12: 03 # La mujer fingió su propio secuestro y logró que sus familiares pagaran para que la liberaran.', 'de " inentendible ". AR. 32802844 Una mujer fingió su muerte para que su novio dejara de ahorcar la Fue en Lanús', 'Ambos cuerpos quedaron tirados en el asfalto, pero debido a el fuerte impacto la mujer pereció de manera instantánea. Montoya Carrasco i', '# Texto encontrado sobre William Melton | The Guardian # Melton agregó que la mujer fingió su muerte, a través de un maniquí, para lueg', 'que permitan dilucidar cómo llegó el auto a ese punto. Asimismo, si la mujer pereció allí o fue trasladada sin vida desde otro lugar, explicaro', 'CEO de JetSmart. AR. 32687714 Una mujer fingió el secuestro de su bebé para ocultar que lo asesino y lo descuartizó #', 'suroccidental de San Cristóbal dos hombres murieron por disparo de arma de fuego y una mujer pereció a consecuencia de un disparo de b', 'mayor de Maquera Maquera. PE. 72672898 Mujer fingió tener autismo para agredir sexualmente a su cuidador # Racheld Childs, de EE', 'recoger sus documentos tuvo que confesar el hecho. | Fuente: Difusión # Una mujer envenenó a su pareja mezclando un potente veneno pi', 'confirmada después de culminar las investigaciones. # El hecho violento Testigos oculares narraron que la mujer fingió ser una pasajera, sin

Figure 18. The collocates and its collocations in context

A total of 3,200 verbal collocations were extracted from the NOW corpus. The next step was to classify these verbs using a taxonomy, which enabled a more extensive linguistic analysis of how men and women are represented. For this purpose, the ADESSE taxonomy was used. It should be noted that some verbs were left unclassified as they were not included in the ADESSE taxonomy. Column K in Figure 19 displays a sample of the results of this classification.

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	BUSQUED	NODO	Colocacio	Colocatio	VENTANA	FECHA	Context	Lemma	Taxonomia ADESE	FREQ	ALL	%	Mi		
32	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	ENVENENÓ	ENVENENAR	Modificación	22		1451	1.52	6.36	
33	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	DESCONFIÓ	DESCONFIAR	Creencia	8		566	1.41	6.26	
34	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	ESTORNUDÓ	ESTORNUDAR	Fisiología	1		71	1.41	6.26	
35	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	APUÑALÓ	APUÑALAR	Modificación	130		9735	1.34	6.18	
36	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	SUBSISTIERON	SUBSISTIR	Tiempo	2		157	1.27	6.11	
37	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	RADICÓ	RADICAR	Localización/Relación	232		18599	1.25	6.08	
38	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	ACUCHILLÓ	ACUCHILLAR	Modificación	32		2626	1.22	6.05	
39	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	CONSENTÍO	CONSENTIR	Permiso	25		2061	1.21	6.04	
40	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	PERECIÓ	PERECER	Vida	77		6469	1.19	6.01	
41	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	DEFECÓ	DEFECAR		3		256	1.17	5.99	
42	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	DEGOLLÓ	DEGOLLAR	Vida	18		1551	1.16	5.98	
43	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	MEMORIZÓ	MEMORIZAR	Conocimiento	3		261	1.15	5.96	
44	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	EMPUÑO	EMPUÑAR	Control	8		710	1.13	5.93	
45	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	SUPLANTÓ	SUPLANTAR	Sustitución	10		906	1.1	5.9	
46	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	INGIRIÓ	INGERIR	Ingestión	32		2985	1.07	5.86	
47	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	MURIÓ	MORIR	Contacto	4009		387971	1.03	5.81	
48	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	APRISIONÓ	APRISIONAR	Control	2		203	0.99	5.74	
49	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	CERCENÓ	CERCENAR	Modificación	10		1019	0.98	5.73	
50	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	DENUNCIÓ	DENUNCIAR	Valoración	3145		329314	0.96	5.7	
51	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	TARAREÓ	TARAREAR	Emisión de Sonido	2		218	0.92	5.64	
52	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	MEDICÓ	MEDICAR		2		221	0.9	5.62	
53	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	FALLECIÓ	FALLECER	Vida	2619		292149	0.9	5.6	
54	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	ALUMBRÓ	ALUMBRAR	Modificación	19		2144	0.89	5.59	
55	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	INCREPÓ	INCREPAR	Valoración	65		7371	0.88	5.58	
56	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	RELATÓ	RELATAR	Comunicación	1354		155452	0.87	5.56	
57	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	RELLENÓ	RELLENAR	Localización:Modificaci	7		816	0.86	5.54	
58	Mujer_N*	Mujer	_VIS*	vPRET(hablé) 1l		03/07/2019	RESULTÓ	RESULTAR	Atribución	2364		276511	0.85	5.54	

Figure 19. Classification of the verbs based on the ADESE taxonomy.

A collocational analysis was conducted using the data obtained in this stage to examine how men and women are portrayed in web-based news articles and newspapers. It is worth noting that the configuration of the interface of the NOW corpus has undergone changes since these searches were performed, which means that those attempting to replicate them may obtain slightly different results. In the next chapter, I will present the data obtained from the NOW corpus and delve into how analyzing the adjectival and verbal collocations of the lemmas hombre ‘man’ and mujer ‘woman’ can aid in examining gender representation in a corpus.

4.2 YouTube Corpus and Keyword Analysis

After obtaining the collocations in the NOW corpus, I aimed to conduct a more extensive analysis. Specifically, I sought to identify keywords related to men, women, and the LGBT community through an automatic classification experiment. To accomplish this, I needed to build a corpus that could help me identify such keywords. Unlike the NOW corpus, which contains web-based news articles that are generally edited for communication purposes and follow journalistic conventions, the new corpus did not have to adhere to such conventions. Thus, I needed access to data that reflected how people on the street talked about men and women, focusing on how ordinary individuals portrayed them. After considering various

sources for data collection, I chose the YouTube social media outlet because it provided the most substantial data.

To build the corpus, I began by selecting YouTube videos that addressed issues related to men and women. Using a keyword approach, I identified videos that could provide the relevant information, such as those related to “Femicide.” Figure 20 displays a frame from one of these videos.



Figure 20. YouTube video about femicide

Each video selected for this study offered a glimpse into the public's views on the topic, as viewers often reacted and commented on the content. One benefit of using YouTube as a corpus-building source is its anonymity, which encourages users to express opinions they may not share in a public forum. The platform is known for its contentious and controversial nature, and the comments on the selected videos for this research project were no exception. The videos, which dealt with issues surrounding femicide, provided valuable insight into how men and women are linguistically represented. Figure 20 displays a selection of comments made by users on the topic of femicide.

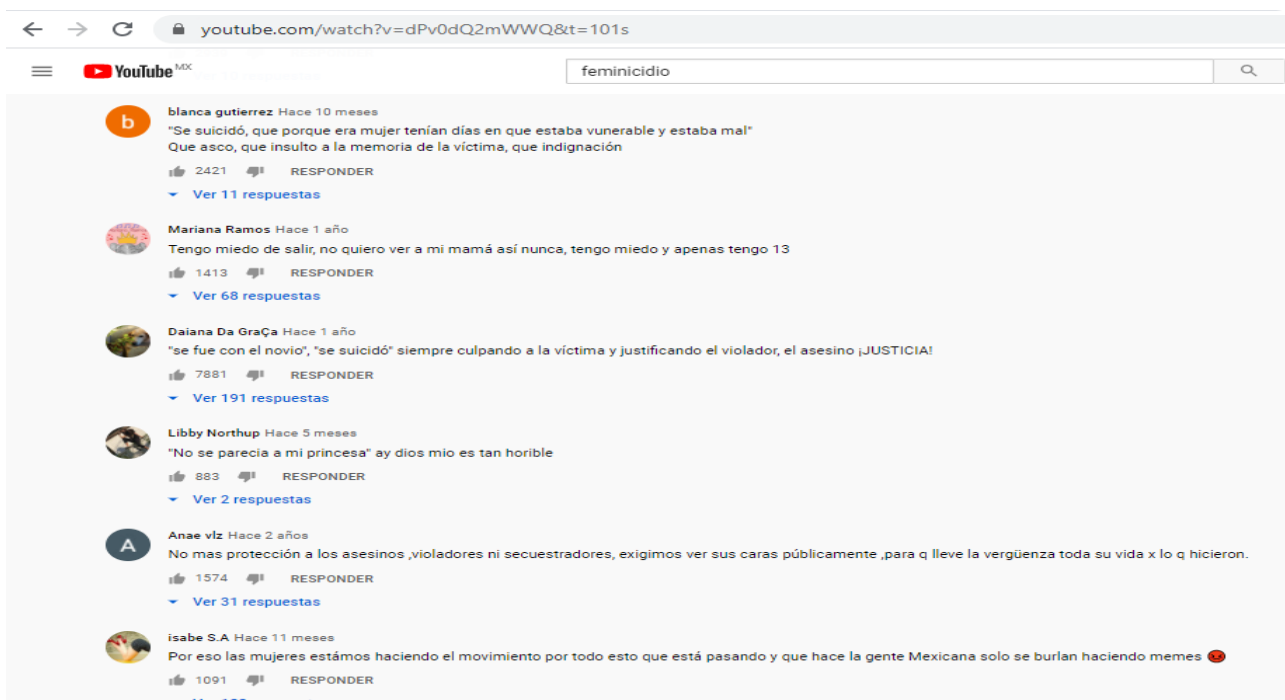


Figure 21. YouTube users' comments on femicide

Upon thorough examination of the information collected from the initial videos, it became apparent that those which focused primarily on women as the main subject should be separated into a distinct corpus. As a result, two corpora were compiled: one addressing topics concerning women and another centered on men. Later in the process, a third corpus was created, focused on the LGBT community. These three corpora were labeled as Viomujdis, Viogendis, and LGBTdis, respectively. Table 20 displays the keywords and the word count for each corpus.

Name of corpora	Topic	# of words per corpus
Viomujdis	Acoso callejero, acoso sexual, femicidio, feminismo, trata de personas, lenguaje machista.	731,286
Viogendis	corrupción, narcotráfico, homicidio, migración, secuestro, bullying.	684,994

LGBTdis	lésbico, gay, transexual, bisexual, matrimonio del mismo sexo, marcha orgullo LGBT	425,105
---------	--	---------

Table 20. YouTube Corpora.

When treating CMC sources as a digital genre (Miller & Kelly, 2016), it becomes apparent that users in these digital environments often do not conform to standard writing conventions. Instead, new writing literacies have emerged and are becoming increasingly common in these settings. As a result, the data required preprocessing before analysis, with the main focus being on correcting spelling errors. It should be noted that the comments could have benefited from more thorough editing or even rewriting, but due to the large amount of data, such an approach would have been impractical. Table 21 provides examples of these comments.

	Corpus: Viomujdis	Corpus: Viogendis	Corpus: LGBTdis
Samples of the YouTube comments	<p>Nada más que un buen correctivo bien aplicado en el hocico para que cambien de actitud estas pinches viejas.....</p> <p>Solo quieren llamar la atención para que no se sientan ofendidas, Si andan siempre con el resentimiento.</p> <p>El mundo estaría mejor si todas estas bestias asquerosas se murieran, mucha gente habla de encerrarlos y de castigarlos y no sé qué masmas, no señores!</p> <p>Que coraje, que impotencia, a este país se lo está cargando la mierda, empezando con que quien supuestamente debe impartir la justicia</p>	<p>(la gaviota) (la gaviota) es una gran hipócrita y falsa por eso se la suena el perro de su esposo que como compro esa casa también la compro a ella.</p> <p>Noooo por favor, no dejen que se queden aquí en México.</p> <p>Hay que alzar la voz para hacer un muro con la frontera sur!</p> <p>Sinceramente si a aumentando un chingo la delincuencia esperemos y el buen Obrador si de resultados que vamos de mal en peor.</p> <p>Lárguense invasores a su país ya no sean una carga para los mexicanos regrésense a su país los viejillos seniles de López obrador Sánchez cordero...</p>	<p>rompí un mandamiento y me enamoré de mi mejor amiga</p> <p>"no sé si suscribirme o no porque cada vez que la veo, dudo más de mi sexualidad :c</p> <p>arrepíentanse, Jesucristo nunca se casó y fue crucificado muriendo por nosotros, su alma fue siempre pura porque se resistió al pecado, ustedes también hagan lo propio y carguen su cruz</p> <p>gracias de todo corazón!!!</p> <p>porque todos somos iguales, nos merecemos más que respeto, amor, justicia, igualdad, lealtad y sobre todo más unión entre nosotros por la comunidad lgbtttiq!!! sin importarnos lo de los demás!!!</p>

Table 21. Comments made by YouTube users

After compiling the corpora, the next step was to conduct a keyword analysis to identify words related to topics about men, women, and the LGBT community. In a keyword analysis,

a corpus is compared to a reference corpus. Therefore, the corpora were analyzed using the AntConc query tool. Each of the corpora was compared to the other two, resulting in six independent comparisons. Keywords are defined statistically as words whose frequency is unusually high compared to a reference corpus. To calculate keywords, Rayson (2013) outlines a three-stage procedure. The first stage involves computing a word frequency list for each text or corpus and counting the total number of words. The second stage compares the frequency lists using a keyness statistic measure to determine the relative frequency of each word in the two texts or corpora. The third and final stage orders the words according to their keyness value, with higher keyness values indicating more relevant keywords. All three stages were performed using the AntConc query tool. Figure 21 shows the Viogendis corpus as the experimental corpus and Viomujdis as the reference corpus, uploaded into AntConc to generate the list of keywords.

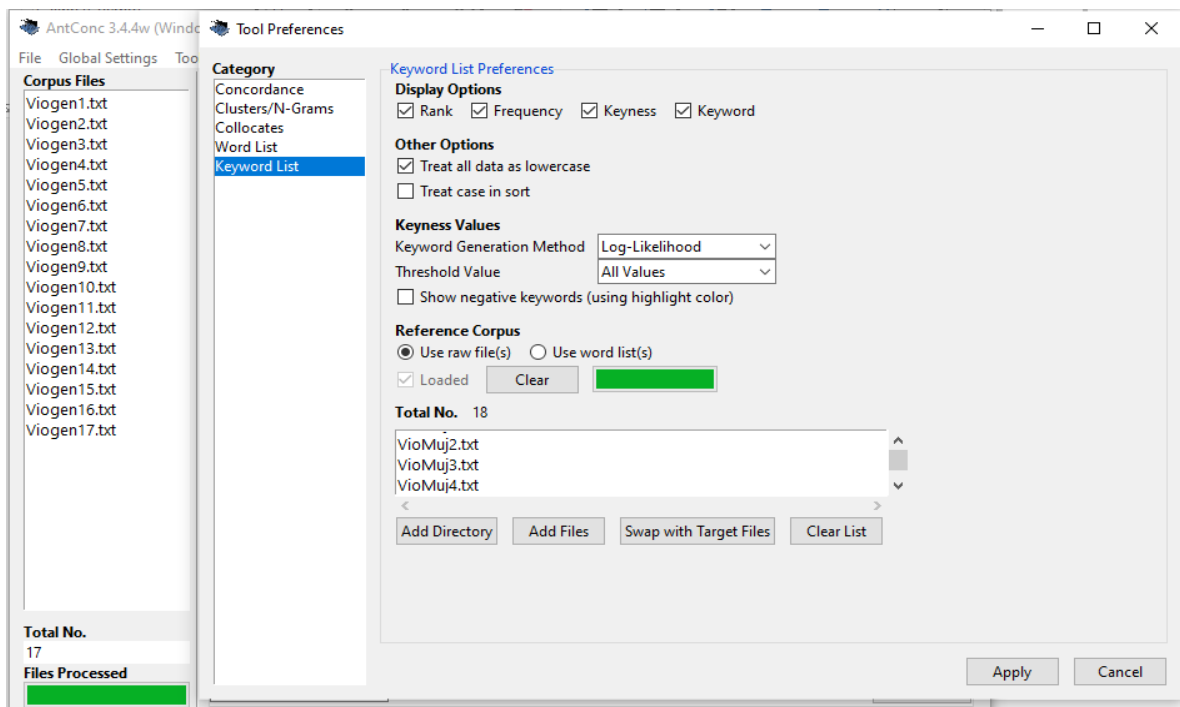


Figure 22. AntConc interface. Parameters set to generate the keyword list

The results of comparing the Viogendis and Viomujdis corpora are displayed in Figure 23, which presents a list of keywords along with their keyness measures.

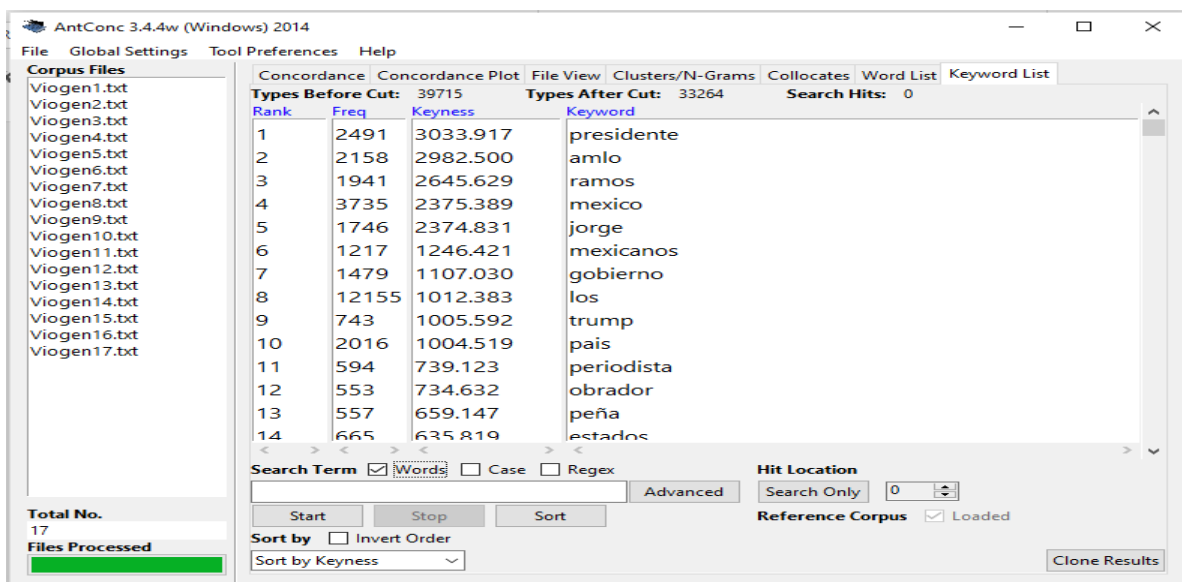


Figure 23. Keyword list obtained from comparing Viogendis and Viomujdis

As previously mentioned, each of the three corpora was individually compared with the other two, resulting in a total of six comparisons. Table 22 presents a selection of the results obtained from these comparisons.

KEYWORD ANALYSIS											
Keyness	Viogendis	Keyness	LGBTdis	Keyness	Viomujdis	Keyness	LGBTdis	Keyness	Viogendis	Keyness	Viomujdis
2193.7	presidente	1887.231	bisexual	2233.468	mujeres	1762.54	bisexual	3045.662	presidente	4668.573	mujeres
2147.5	ðy	1669.384	me	2056.176	ðy	1707.13	dios	2992.908	amlo	4522.338	mujer
2007.8	mexico	1531.563	gay	1506.969	hombres	1423.17	gay	2646.74	ramos	3930.515	hombres
1985.5	amlo	1310.806	matrimonio	1034.485	feminismo	1285.46	matrimonio	2419.929	mexico	3276.292	hombre
1797.2	ramos	1239.234	soy	1033.752	igualdad	1156.29	homosexuales	2383.231	jorge	2005.68	igualdad
1451.9	jorge	1217.146	homosexual	967.833	xd	813.001	rix	1251.937	mexicanos	1808.471	genero
1201.2	pais	1212.728	dios	910.159	lenguaje	762.599	homosexual	1113.174	gobierno	1648.68	feminismo
929.81	gobierno	1062.809	mujer	898.442	jajaja	752.407	homosexualid.	1035.776	los	1299.678	lenguaje
878.41	mexicanos	882.306	sexo	893.412	las	751.753	comunidad	1011.851	pais	1257.676	una
625.22	jajaja	836.897	señora	885.023	mujer	694.844	amor	1009.164	trump	1246.209	las
596.17	trump	803.166	amor	784.758	feminista	669.466	gays	741.935	periodista	1184.991	feminista
582.81	corrupcion	786.17	homosexual	703.049	ryan	665.032	iglesia	737.282	obrador	1121.155	feministas
508.57	periodista	779.927	rix	665.539	feministas	664.825	soy	661.757	peña	1018.03	ryan
484.44	obrador	765.816	hombre	602.873	hombre	568.314	religion	638.78	estados	1013.711	me
430.07	unidos	759.437	yo	575.222	jajajaja	519.273	biblia	635.786	meses	895.224	v
402.15	mexicano	742.607	homosexual	505.212	acoso	499.716	homofobica	600.145	unidos	888.463	machista

Table 22. Keyword analysis carried out for each corpus

This section provides an overview of the process for compiling the Viogendis, Viomujdis, and the LGBTdis corpora, along with a brief description of the preprocessing steps. Furthermore, I have explained how the keyword analysis was conducted by comparing each corpus with the other two to identify the words with the highest keyness measure. In the

next section, I will delve into the details of how these keywords were utilized in the automatic classification experiments.

4.3 Automatic Text Classification

After compiling the corpora and identifying the keywords with a keyness score above 3, the next step was to conduct automatic text classification experiments. Two experiments were carried out with different methods: The first experiment used the Violentómetro and keywords for text classification with a weight frequency scheme, while the second experiment used keywords for comment classification with a Boolean scheme and a string to word vector filter. It is worth noting that the term “text” refers to the comments from each video. In the first experiment, the texts were represented in a vector space model (VSM) to identify the frequency of each keyword within each text. This process, called indexing, reduces the complexity of the texts and makes them easier to handle. Table 23 displays the VSM used to record the frequency of occurrence of feature terms (keywords) in each document (text).

	Vector	Matrices	
	Feature 1 (keyword)	Feature 2	Feature 3
Text 1(Viomujdis 1)	0		
Text 2(Viomujdis 2)	2		
Text 3	1		
Text 4	16		

Table 23. Vector Space Model

The VSM in question represents 18 texts from Viomujdis, 17 texts from Viogendis, and 16 texts from LGBTdis, along with the frequency of each feature in each text. It’s worth noting that three sub-experiments were conducted using different features to determine which ones

would enhance the accuracy of the classifiers. Table 24 displays the various features utilized in each of the sub-experiments.

	1st experiment (Violentómetro- UAEM)	2nd experiment (29 keywords)	3rd experiment (242 keywords)
Keywords to classify the texts	asesin*, viola*, abus*, amenaz*, manose*, control*, menti*, intimidar*, humill*, golpe*, cachetea*, ofend*	Dios, respet*, acept*, derecho*, discrimina*, iguald*, acosa*, defend*, merez*, agred*, prostitu*, denunci* mata*, bend*, provoca*, soy*, biblia*, pecado*. Odio*, mandamiento* mujer*, hombre, inclusivo*, maltrat*, muert*, poder, culp*, critic*, *amlo	bullying, asil*, armar*, ayotzinapa, caravana*, catolic*, chairo*, corrup*, cree*, deporta*, fifi*, crim*, impunidad*, mediocre*, mafia*, politic*, racis*, pendej*, bisexual, creyente*, gomorra, prejuicio, etc.

Table 24. Features in each one of the three sub-experiments

The first sub-experiment utilized the categories (words/verbs) from the Violentómetro as features for evaluating the classifier, as can be observed in Table 24. It is worth noting that in all three sub-experiments, each feature was searched in the texts using a regular expression (), which is a condensed representation of a word that allows for the identification of related words. For instance, the Spanish regular expression of asesin* would match words such as asesinar, asesina, asesinó, and so on.

For the last two sub-experiments, keywords with a keyness measure above 3 were selected as features. However, since there were numerous words that met this criterion, the selection process took into account both the keywords with the highest keyness and the researcher's intuition regarding the words that could yield better results. Table 25 displays the VSM with 29 selected features, while Table 26 displays a VSM with 243 different features.

1	Class	Dios	Respet*	Acept*	Derecho*	Discrimina*	Iguald*	Acosa*	Defend*	Merec*	Agred*	Prostitu*	Denunci*	Mata*	
14	LGTBdis		138	233	31	48	18	8	2	11	11	6	1	0	7
15	LGTBdis		150	96	29	36	22	3	0	3	5	2	1	0	0
16	LGTBdis		4	29	27	2	5	2	0	1	2	0	0	0	5
17	Viogendis		22	2	13	22	0	1	0	8	3	3	18	5	12
18	Viogendis		271	32	7	35	0	0	2	16	10	0	0	13	88
19	Viogendis		12	30	5	13	211	1	14	60	6	17	1	10	8
20	Viogendis		17	27	27	5	0	0	0	28	0	0	10	2	19
21	Viogendis		28	13	2	70	0	0	0	41	5	3	0	3	128
22	Viogendis		69	127	8	44	1	2	2	14	14	9	3	10	36
23	Viogendis		27	28	8	7	1	0	0	23	12	6	2	34	25
24	Viogendis		53	80	6	71	1	0	0	43	14	5	2	0	46
25	Viogendis		16	21	13	7	6	2	0	13	8	1	2	11	21
26	Viogendis		50	94	9	59	1	0	0	31	13	11	1	3	33
27	Viogendis		99	24	7	12	5	0	0	4	31	0	0	5	18
28	Viogendis		29	53	8	49	2	1	0	21	19	4	2	6	25
29	Viogendis		27	53	8	7	1	2	0	11	8	8	3	0	42
30	Viogendis		20	49	1	9	0	2	0	15	5	0	3	3	32
31	Viogendis		35	40	10	28	1	0	0	19	6	6	5	2	49
32	Viogendis		26	28	5	8	0	1	0	22	10	1	0	5	27
33	Viogendis		35	40	10	28	1	0	0	19	6	6	5	2	49
34	Viomujdis		8	95	3	29	6	9	562	6	7	33	4	8	5
35	Viomujdis		76	27	10	4	0	0	1	0	29	0	94	20	40
36	Viomujdis		17	134	9	40	8	36	126	4	19	15	13	2	2
37	Viomujdis		25	48	5	41	0	0	5	29	34	36	2	44	68

Table 25. Sample of VSM with 30 features

1	Class	acab*	ancian*	armar*	arregl*	asesin*	asil*	ayotzinapa	baj*	bala*	bullying	caravana*	carcel*	catolic*	chairo*	
2	LGTB		8	0	0	0	0	0	1	0	0	0	1	2	0	
3	LGTB		10	3	1	0	1	0	1	3	0	1	0	1	16	0
4	LGTB		33	1	0	0	7	0	0	4	3	0	0	1	29	1
5	LGTB		9	1	0	3	14	0	0	8	0	0	0	0	86	0
6	LGTB		10	1	1	0	2	0	0	2	0	2	0	0	1	1
7	LGTB		15	0	2	0	5	0	0	6	0	0	1	0	5	0
8	LGTB		5	0	1	0	1	0	0	0	0	1	0	0	0	0
9	LGTB		10	0	1	0	2	0	0	4	1	0	0	0	5	0
16	LGTB		7	0	0	1	9	0	1	4	0	1	0	0	15	1
17	Viogen		17	0	0	2	5	0	11	11	1	0	0	4	2	1
18	Viogen		22	0	5	6	4	19	0	12	39	0	40	8	0	1
19	Viogen		16	1	2	4	8	4	0	21	4	2	27	3	0	17
29	Viogen		59	6	8	9	39	0	0	19	32	0	2	12	1	13
30	Viogen		20	1	0	3	34	0	0	23	1	0	9	14	0	6
31	Viogen		16	1	1	2	22	0	0	15	2	0	0	25	233	0
32	Viogen		13	2	2	8	3	31	0	16	39	0	50	4	0	0
33	Viogen		12	0	0	6	6	10	0	23	1	1	33	10	0	3
34	VioMuj		8	2	0	7	2	0	0	18	0	0	0	6	1	1
35	VioMuj		19	0	3	0	1	0	0	11	0	0	0	1	0	0
36	VioMuj		21	0	3	0	1	0	0	13	0	0	0	2	0	0
37	VioMuj		13	0	0	3	6	0	0	6	1	0	0	2	0	1
38	VioMuj		21	2	1	3	4	0	0	7	0	0	0	3	1	0
39	VioMuj		16	1	3	1	6	0	0	16	0	7	0	10	1	0
40	VioMuj		5	0	2	3	0	0	0	3	1	1	0	12	0	1

Table 26. Sample of VSM with 242 features

The second major experiment for text classification utilized a Boolean weighting scheme with a string to word vector filter. Unlike the previous experiment, this one focused on classifying individual comments within each text based on the presence or absence of features rather than their frequency. This required a more thorough preprocessing of the data, as all words in each comment were treated as features. Since there were nearly 100,000 comments in the three corpora, the number had to be reduced to 30,000, and eventually to 7,500 comments (2,500 per corpus) due to limitations in the software and equipment. Figure 24

displays a selection of comments that were classified according to their corresponding corpora, which were renamed V-Mujer, V-General, and V-LGBT for this experiment.

Instances of comments that were classified
<ul style="list-style-type: none"> • "Como se van a mesclar con la gente normal si eso es aberración ante dios." (V-LGBT) • "Cinthya M. que carajos, mejor ni respondo, yo no creo en dios" (V-LGBT) • "No se trata de burlarse de dios que se justifican mediante eso" (V-LGBT) • "Ella al hablar de la palabra de dios se cre superior a todo lo demás que hay en el mundo, cualquier religión o decisión" (V-LGBT) • "Que pesar tan grande ojala que los culpables sean detenido y ese hombre jamás salga", (V-General) • "si soy sincero me busco una vida en prisión y creo 4 cuerpos de los culpables los 3 que el sabe que fueron y el que lo estafo obvia que. valiera la pena un cuarto muchos juguetes y con que mantenerlos vivos", (V-General) • "en México les faltan valores a la gente culpa lo tienen el gobierno y la gente uno como padre no les enseña valores y los hijos andan en la calle y los padres les vale no asen nada y el gobierno les faltan pantalones porque no tienen huevos roban y todavía dicen vamos a ser un México mejor te dan puro palo" (V-General) • "Pobre sr ojala que a los culpables los alcance el karma y sufran mucho mas de lo q sufrio su hija y q sufre usted", (V-General) • "Poco a poco me estoy dando cuenta de que el feminismo no tiene sentido :/." (V-Mujer) • "Puede que si la mayoría miren mal al feminismo no es por nuestra culpa, sino por la vuestra, por desvirtuar el termino al ser tan retrasadas (algunas, las que mas ruido hacen)" (V-Mujer) • "Que va del feminismo a machete al machote creo no es igual" (V-Mujer) • "estimo que esto no le contribuye mucho al feminismo la verdad. En fin." (V-Mujer)

Figure 24. Classified comments according to their classes.

Displayed in Figure 25 is the Weka interface, which was used to upload the file containing the comments. As shown, the total number of comments to be classified is 7,493, and the number of attributes (features) considered in this classification is 1,751. It should be noted that the number of features was intentionally reduced to ensure that the amount of information analyzed would not exceed the computing capacity of the equipment.

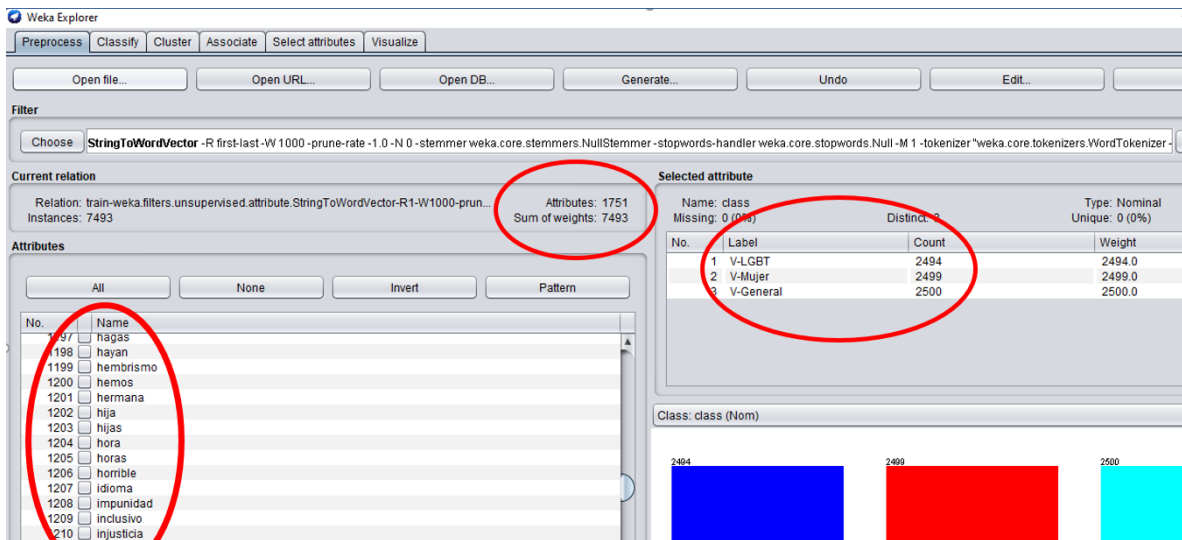


Figure 25. Weka interface showing instances (comments) and attributes (features)

As mentioned earlier, the string to word vector filter operates on a Boolean scheme by searching for each feature in each comment, and classifying the comments based on their presence or absence. Figure 26 provides a visual representation of how the string to word vector filter is executed.

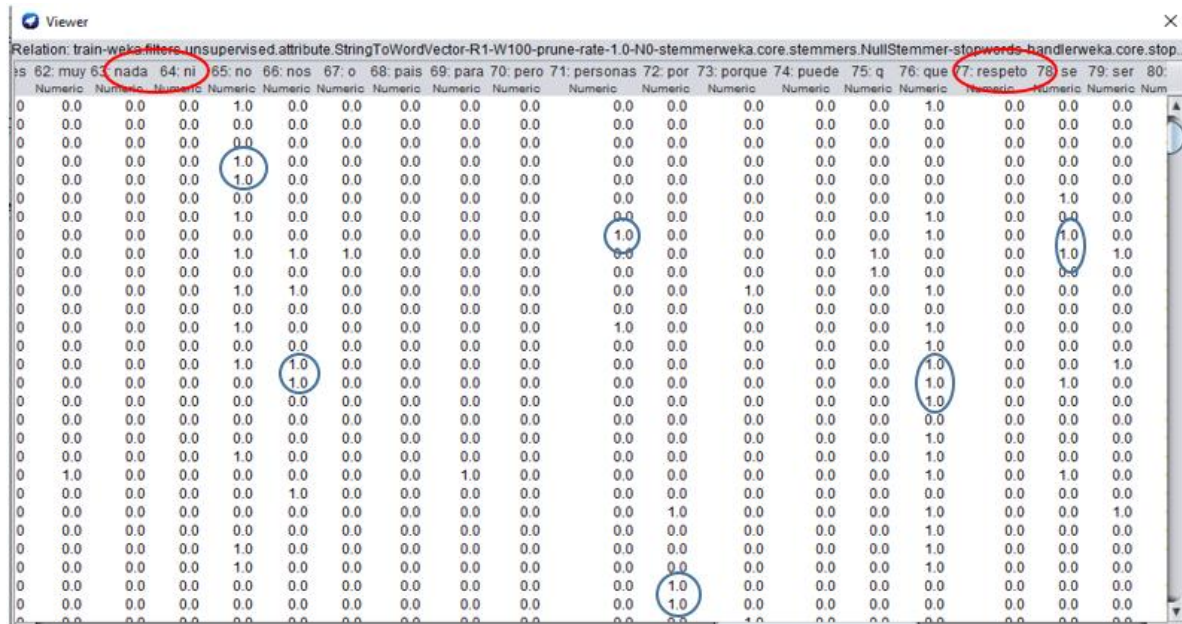


Figure 26. Boolean scheme

In the second major experiment, it is noteworthy that there were two sub-experiments. The first sub-experiment involved the 1,751 features that included the identified keywords, while in the second sub-experiment, the keywords were deliberately excluded to evaluate their impact on the classifiers' accuracy. It is worth mentioning that the identification of the keywords was done manually. A selection of the keywords removed in the second sub-experiment is presented in Table 27.

Keywords used in the string to word vector experiment			
Keywords related to religious terms	Keywords related to identities and other phenomena	Keywords related to political terms	Other verbs, adjectives, and nouns.
Dios Iglesia Matrimonio Mandamiento Pecado Biblia Cristiana Religión Casarse Familia Cree*	Ideología LGBT Feminismo Lesbiana Equidad Feminazi Bisexuales Homofóbica Identidad Patriarcado Igualdad Gay Soy Discriminar Derecho Inclusivo Amor	Política Corrupción Calderón. Invasores Muros Justicia Pobre Políticos Salarial Amlo Fifis Prian Migrantes Peje Pejezombies Violencia Victima Caravana Chairos Chayoteros	Acept* Defend* Maltrat* Viola* Respet* Acoso* Culpable Hombre* Mujer* Put* Agresión Gorda Victima Muerto* Rata*

Table 27. Keywords removed to assess their weight in the text classification

The removal of keywords in the second sub-experiment had a significant impact on the classifiers, demonstrating the importance of these keywords. The subsequent chapter will provide an in-depth analysis of the results from both the VSM representations and the String to word vector experiment.

The current chapter has detailed the identification of adjectival and verbal collocations in the NOW corpus and the employment of two classification frameworks to group these collocations. Furthermore, it outlines the compilation of the YouTube corpus and the extraction of keywords from it. The chapter also delves into the development of various text

classification experiments and the significance of using keywords as features to enhance algorithmic accuracy.

In the upcoming chapter, I will present the results of the collocational analysis of both adjectives and verbs. Additionally, I will discuss the outcomes of the text classification experiments and the performance of the keywords as features in these experiments.

5 Results and Analysis

This chapter comprises two main sections: a collocation analysis and an automatic text classification (ATC) experiment. As previously mentioned, the first stage involved the extraction of adjectival and verbal collocates for the lemmas HOMBRE ‘MAN’ and MUJER ‘WOMAN’ from the NOW corpus. These collocates were selected based on their high Mutual Information (MI) score of above 3, indicating that the lemmas and their respective adjectives or verbs frequently appeared next to each other in the corpus.

The second section presents the results of the text classification experiments and revisits the YouTube corpus previously discussed in earlier chapters. In the ATC experiments, keywords or features were used to perform automatic text classification tasks on the corpus. It is worth noting that this research study adopts an interdisciplinary approach, drawing on techniques, tools, and analysis from the fields of Corpus Linguistics and Natural Language Processing. The use of techniques and tools from both disciplines informed the data collection and analysis processes of the two corpora, making this research study relevant in both fields. As a result, different data analyses were necessary for both the collocational analysis and the ATC experiments.

Overall, this chapter provides a comprehensive overview of the collocation analysis and ATC experiments conducted in this research study, showcasing the interdisciplinary nature of the research and the variety of data analyses employed.

5.1 Results of Experiments of the NOW Corpus

The purpose of the collocational analysis is to examine how men and women are represented in the NOW Corpus, focusing on adjectival and verbal collocations. From the corpus, we extracted 1,586 adjectival collocations with a Mutual Information score above 3. As previously mentioned, these collocations were classified according to Supersenses (Tsvetkov et al. 2014). Of the 1,586 collocations, 834 were associated with the lemma WOMAN, with 435 collocating with the term *mujer* ‘woman’ and 399 with *mujeres* ‘women’. The remaining 752 adjectives collocated with the lemma MAN, with 506 collocating with the term *hombre* ‘man’ and 246 with *hombres* ‘men’. Table 28 displays the results of the lemmas MAN and WOMAN, but only those found to the right of the lemmas.

It is essential to note that the Supersenses classification enabled us to group the adjectives into different categories, which facilitated further analysis. We made a significant effort to identify semantic fields in which we could further group the adjectives. While we were able to group some adjectives, finding similarities among others was not possible.

Furthermore, it is important to clarify that the data analysis presented in the following paragraphs, which utilizes constellation networks, is not an in-depth analysis of all the adjectives. Rather, it offers some insights into how corpus linguistics techniques and data representation software can be used to analyze how women and men are represented in a corpus.

Hombre –Mujer (1D)	Behavior	Body	Feeling	Mind	Miscellaneous	Perception	Quantity	Social	Spatial	Substance	Temporal	Total general
Hombre	173	66	42	66	13	3	3	57	16	6	9	454
Mujer	91	61	24	28	6	1		124	12	2	11	360
Total General	264	127	66	94	19	4	3	181	28	8	20	814

Table 28. Results obtained for each one of the categories in the Supersenses classification

To further analyze the results, I will delve into a few of the categories and elaborate on the findings. First, I will expand on the behavior category’s findings. As shown in the previous

table, more adjectives collocated with the term man than with woman. Throughout this chapter, I will present constellation networks to illustrate the findings. It is essential to note that the thickness of the line connecting the adjective and lemma represents the MI score between the collocate and lemma, indicating the statistical significance of the adjectives with an MI score above 3.

It is crucial to remember that the MI score measures to what extent the occurrences of one word determine the occurrences of another word. Therefore, MI highlights the exclusivity of the collocation's relationships. Additionally, the collocations used in this analysis were obtained from the NOW Corpus, which collected information from newspaper articles on the web from various Spanish-speaking countries in Latin America and Spain. It is essential to consider the nuances that occur when working with different varieties within the Spanish language.

5.1.1 Adjectives

In this section, I will elaborate on the findings concerning the Behavior category. Given the considerable number of adjectives involved, I will only present those with a MI score exceeding 6.5. Figure 27 illustrates the collocating adjectives for both lemmas and their relevance to the adjectives.

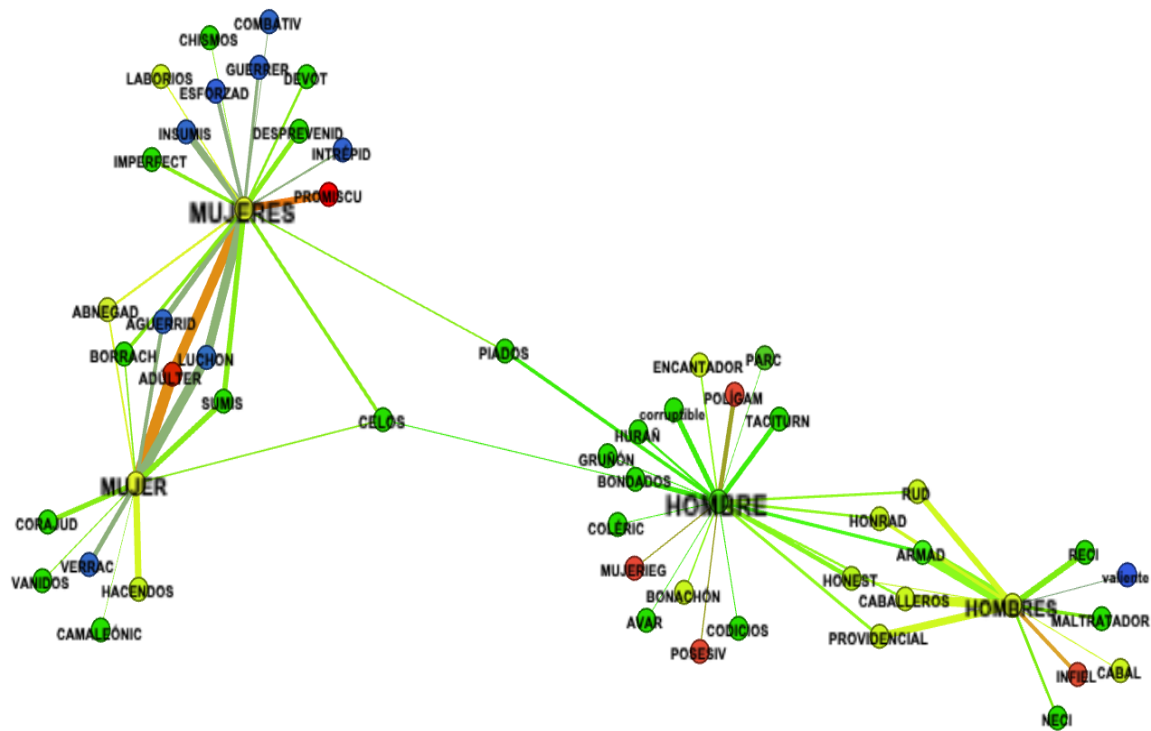


Figure 27. Constellation network of the collocations of the Behavior category

The blue semantic field in the constellation above is associated with struggling, and it contains adjectives like ‘combative’, ‘hard-working’, ‘warrior’, ‘fearless’, ‘unsubmissive’, ‘tough’, ‘feisty’, and ‘intelligent’ which are only used to describe women. Among these, ‘unsubmissive’ has one of the highest MI scores, indicating its exclusive association with women. This semantic field suggests that women experience more struggle than men, which is not only evident in their daily lives but also discussed in newspaper articles. The existence of such struggle implies a power imbalance, where women react to situations where power is distributed unequally. This power struggle is present in various social contexts, such as work, school, and relationships, where male dominance is the norm. If we assume that such inequality is reflected in the data, then there is not enough evidence to suggest that men experience the same degree of struggle as women. These adjectives could also imply that women are resisting this inequality, which is reflected in their representation in the NOW corpus.

It is pertinent to revisit the concept of “markedness” to understand the meaning behind marked and unmarked terms. The addition of a linguistic particle, such as an adjective, alters the meaning of a marked term, whereas an unmarked term conveys a meaning that is commonly understood. In the case of the adjectives within the semantic field of struggling, they only appeared with the noun “woman”. This indicates that women are marked with adjectives, suggesting that it should not be taken for granted that women possess qualities like fearlessness, unmissiveness, and combativeness. Conversely, the absence of these adjectives with “man” could imply that these traits are considered default characteristics of men.

Let’s turn our attention now to the adjectives marked in red. The adjectives *promiscua* ‘promiscuous’ and *adúltera* ‘adulterous’ only appeared with women, while the adjectives *polígamo* ‘polygamous’, *mujeriego* ‘womanizer*’, *posesivo* ‘possessive’, and *infiel* ‘unfaithful’ only appeared with men. These adjectives are part of the sexual conduct semantic field. Of all these adjectives, *adúltera* and *adúlteras* had a high MI score of 10 and 9, respectively, with only one other adjective in the entire corpus scoring above 10. This finding indicates that women are more likely to be associated with adultery and be described as such, while the adjective *unfaithful* only collocated with men.

It is important to note the distinction between the denotative meanings of the term “adultery” and “unfaithfulness” in legal contexts. Adultery refers to having sexual relationships with someone while being married to someone else, and is considered a legal cause for seeking divorce. Unfaithfulness, on the other hand, is a vague and subjective term in legal contexts, and could refer to emotional relationships or flirting with someone of the opposite or same sex, without necessarily implying sexual relationships.

Based on the above, it can be argued that women are more harshly represented in the NOW corpus, as the choice of words used to describe women's sexual behavior seems to criminalize their actions, whereas similar behavior by men is not similarly vilified.

The constellation network features a range of adjectives that present a challenge in terms of categorization. However, it is worth noting that some adjectives only collocate with either men or women. For instance, the adjectives *devotas* ‘devoted’ and *abnegada(s)* ‘self-

sacrificing’ are associated exclusively with women. It is interesting to reflect on these adjectives, which have been used historically to represent women who dedicate their lives to the well-being of others. Unfortunately, in some contexts, the use of such adjectives in a negative sense equates these women with individuals who forcibly surrender their life expectations to serve or please someone else. The fact that these adjectives only collocate with women could imply that there are certain social customs or expectations that are specific to women.

In the constellation network, we observe that certain adjectives have both a positive and negative polarity. For instance, adjectives like *maltratador(es)* ‘abusive’, *necio* ‘foolish’, *cólericos* ‘choleric’, *codiciosos* ‘greedy’, and *avariciosos* ‘greedy’ only collocate with men, whereas adjectives like *corajuda* ‘short-tempered’ and *chismosas* ‘gossipy’ only collocate with women. I grouped these adjectives into a semantic field that I call self-restrain, as they share the connotation of lacking self-restraint. These collocation patterns reveal the associations and connotations of words, and thus, the assumptions they embody for both women and men. One assumption we can make is that men can mistreat people physically, emotionally, and/or psychologically, as in the case of the adjective *abusive*, while women are more likely to gossip about others without reaching the point of physical mistreatment, as in the case of the adjective *gossipy*.

The following constellation network shows the adjectives that were categorized under “Body.” As we can see, certain adjectives only collocate with either men or women, while in some cases, certain adjectives only appeared in either singular or plural form. However, in many instances, some adjectives appeared with both lemmas.

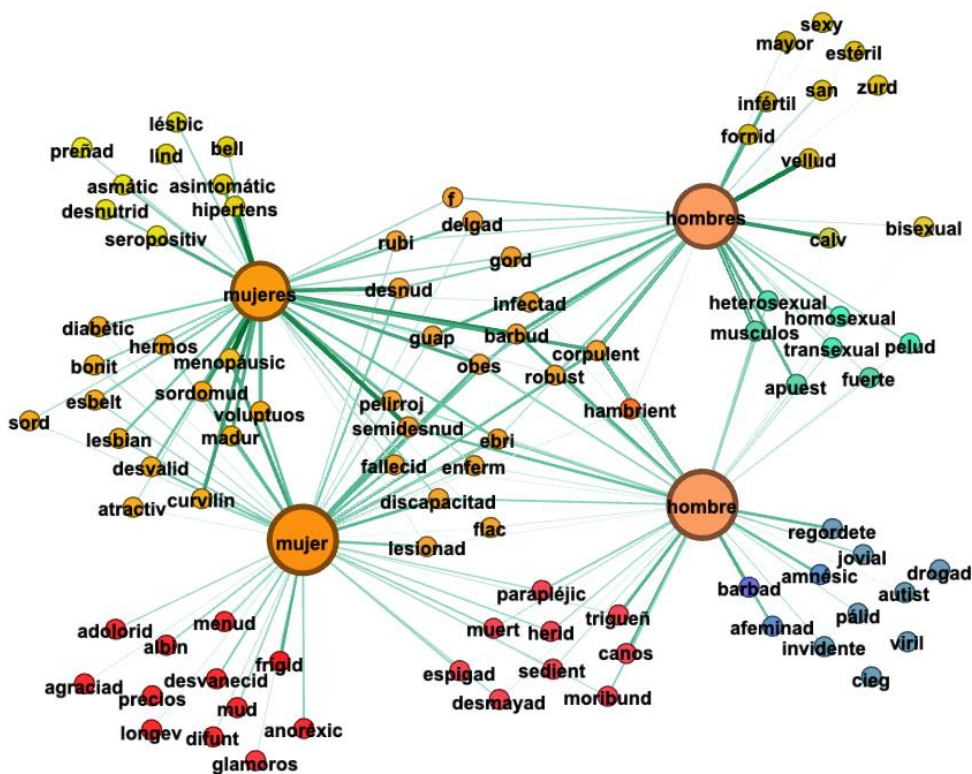


Figure 28. Constellation network of the collocations of the Body category

To begin with, the collocations were categorized into semantic fields, as mentioned earlier. However, only a few adjectives were identified under the semantic field of age. The adjectives *longeva* ‘long-lived’ and *madura(s)* ‘mature’ were found to collocate with woman, whereas *jovial* ‘youthful’ and *mayores* ‘senior’ collocated with man. Surprisingly, no adjective with a high MI score related to youth was found to collocate with woman. This finding deviates from the results of similar research in the past, which reported a substantial number of collocations pertaining to age and aging with both lemmas. It is plausible that if adjectives with a lower MI score were also considered, more adjectives related to age and aging might have been discovered.

Several adjectives in this constellation were also placed in a semantic field category labeled Health. The adjectives *adolorida* ‘sore’, *anoréxica* ‘anorexic’, and *muda* ‘mute’ collocated with woman and *hipertensa* ‘hypertense’, *asmática* ‘asthmatic’, *desnutrida* ‘malnourished’,

and *asintomática* ‘asymptomatic’ collocated with women; *diabética* ‘diabetic’, *menopáusica* ‘menopausal’, *sordomuda* ‘deaf-mute’, *sorda* ‘deaf’, and *desvalida* ‘helpless’ collocated with both forms. Regarding the form men, only *infértil* ‘infertile’ and *estéril* ‘sterile’ collocated with it, and *pálido* ‘pale’, *autista* ‘autistic’, *ciego* ‘blind’, and *invidente* ‘blind’ collocated with man. It is important to note the connotations of some of these adjectives; for instance, some of the adjectives that collocated with woman imply more serious health issues. It is also important to pay attention to the adjectives *sterile* and *infertile* which have some subtle differences; however, both men and women can be *sterile* and *infertile* but these adjectives only collocated with man. As far as adjectives related to health issues, there were twice as many adjectives that collocated with woman than with man. The following adjectives collocated with both lemmas: *parapléjico(a)* ‘paraplegic’, *herida(o)* ‘wounded’, *desmayado(a)* ‘fainted’, *moribunda(o)* ‘dying’, *fallecido(a)* ‘deceased’, *obesa(o)* ‘obese’, *lesionado(a)* ‘injured’, *enferma(o)* ‘sick’, *discapacitado(a)* ‘disabled’, and *infectada(o)* ‘infected’.

A group of adjectives in this constellation falls within the semantic field of physical appearance, and some of these adjectives cluster within sub-groups. For example, certain adjectives in the sub-group labeled “look” pertain to people’s physical appearance. Adjectives such as *agraciada* ‘gifted’, *preciosa* ‘beautiful’, and *glamorosa* ‘glamorous’ collocate with “woman,” while *bellas* and *lindas* ‘pretty’ collocate with “women,” and *hermosa(s)* and *bonita(s)* with both genders. *Viril* ‘manly’ and *barbado* ‘bearded’ collocate with “man,” *velludos* ‘hairy’ and *sexys* ‘sexy’ collocate with “men,” and *apuesto(s)* ‘handsome’ and *peludo(s)* ‘hairy’ collocate with both genders. The only adjective that collocates with both singular and plural forms of both genders is *guapa(o)* ‘good-looking’. Almost all of the adjectives have positive polarity, referring to pleasant appearances of both men and women.

Another sub-group within the physical appearance category is labeled body type. The adjective *anorexic* collocates with “woman,” while *esbelta(s)* ‘slender’, *voluptuosa(s)* ‘voluptuous’, *atractiva(s)* ‘attractive’, and *curvilínea(s)* ‘curvy’ collocate with both “woman” and “women.” Only *regordete* ‘chubby’ collocates with “man,” *fornidos* ‘well-built’ with “men,” and *musculoso (s)* ‘muscular’ and *fuerte (s)* ‘strong’ with both genders.

Adjectives such as *delgada* (o) ‘thin’, *gordo(a)* ‘fat’, *obesa(o)* ‘obese’, *corpulento(a)* ‘corpulent’, *robusta* (o) ‘sturdy’, *flaco(a)* ‘skinny’, and *espigada(o)* ‘tall’ collocate with both singular and plural forms of both genders.

The last sub-group in the physical appearance category is labeled “sexuality.” *Frígida* ‘frigid’ is the only adjective that collocates with “woman,” while *lésbicas* ‘lesbic’ collocates with “women,” and “lesbian(s)” with both genders. *Afeminado* ‘effeminate’ collocates with “man,” *bisexuales* with “men,” and *heterosexual(es)* ‘heterosexual’, *homosexual(es)* ‘homosexual’, and *transexual(es)* ‘transsexual’ collocate with both singular and plural forms of both genders. These types of analyses allow us to investigate not only gender differences in the representation of sexual orientation, as seen with the adjectives in the sexuality semantic field, but also the word choices people make.

The network that we'll be analyzing next is the constellation of adjectives under the Mind category. In this category, 66 adjectives collocated with the lemma MAN, while only 28 adjectives collocated with the lemma WOMAN. The difference in the number of adjectives associated with each lemma is quite striking.

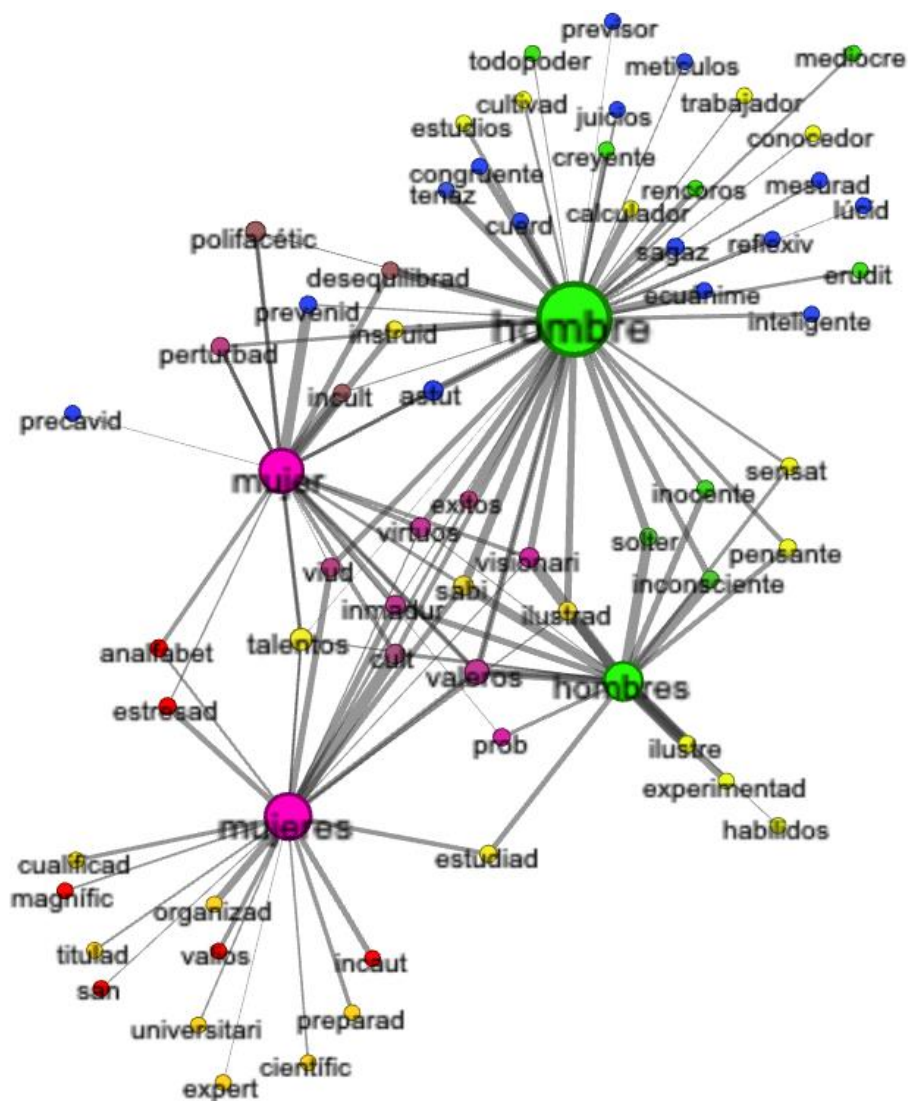


Figure 29. Constellation network of the collocations of the Mind category

The constellation network analyzed includes a semantic field labeled as Instruction ‘nodes in yellow), which is more prominent in the adjectives exclusively collocating with women and men. Adjectives like *cualificad* ‘qualified’, *titulad* ‘graduated’, *organizad* ‘organized’, *preparad* ‘skilled’, *universitari* ‘university student’, *científica* ‘scientist’, *expert* ‘expert’, and *estudiada* ‘educated’ collocate with women, while *sensat* ‘judicious’, *pensante* ‘thinking’, *habildoso* ‘skillful’, *experimentado* ‘experienced’, and *ilustre* ‘illustrious’ collocate with men. For adjectives that collocate with women, formal instruction seems to be implied to become qualified, expert, or a scientist. In contrast, adjectives that collocate with

men suggest natural ability for recipients to be labeled as judicious, skillful, illustrious, or experienced. While natural ability is present in becoming expert, organized, and/or prepared for women, it is not explicitly stated like in the case of men. Only *estudiada* ‘educated’ collocated with both lemmas. The representation of women and men in the network reveals a tendency in the NOW corpus to specify women's level of instruction, which is not present when representing men. This observation highlights not only how women and men are represented but also how they are not.

Another semantic field that emerged is Mental Acuteness (nodes in blue), which encompasses adjectives that connote people’s abilities to focus, recall, and reason, among other things. This semantic field became even more apparent when examining the adjectives that exclusively collocated with the singular form of both lemmas. The adjective *precavida* ‘cautious’ only collocated with woman, while *tenaz* ‘tenacious’, *cuerdo* ‘sane’, *congruente* ‘congruent’, *juicioso* ‘judicious’, *previsor* ‘far-sighted’, *meticuloso* ‘meticulous’, *sagaz* ‘sagacious’, *ecuánime* ‘unbiased’, *inteligente* ‘intelligent’, *reflexive* ‘thoughtful’, *lúcido* ‘lucid’, and *mesurado* ‘prudent’ all collocated with man. The adjectives *astute(o)* ‘shrewd’ and *prevenida(o)* ‘far-sighted’ collocated with both lemmas, although they were more strongly associated with women.

This semantic field is notable for the significant difference in the number of adjectives that collocate with each lemma. There are significantly more adjectives that collocate with the lemma “man” than with “woman.” This indicates that the NOW corpus strongly associates men with mental sharpness and wisdom, while this association is not as strong for women.

This analysis has focused on gender representation, examining commonly identified stereotypes for both women and men. While both genders are often described using similar physical adjectives, women are more frequently associated with terms relating to health, while men are linked to sexual identity orientations and reproductive health issues. Additionally, the NOW corpus suggests that women are commonly associated with the idea of academic instruction, which is not emphasized when discussing men. Some adjectives, such as “adulterous” for women and “unfaithful” for men, appear exclusively with one gender, and may require further investigation.

Critics of Corpus Linguistics argue that it is too quantitative and primarily concerned with identifying linguistic patterns by frequency and collocation in large data sets. However, this critique overlooks the benefits of these techniques, which allow researchers to expand the scope of their research by identifying linguistic patterns in vast amounts of data. Without these tools, researchers would be limited to analyzing data collected through interviews or observations. While there are limitations to working with small data sets, large corpora can also contribute to qualitative research. By analyzing the adjectival features of collocates, researchers can identify detailed similarities and differences between genders. This wealth of data offers an opportunity to gain a deeper understanding of the way in which adjectives are associated with different genders.

Moving on to the Social category, we find that more adjectives collocate with women than men. However, due to the high number of adjectives, only those with a mutual information score of over 4 are included in the constellation network shown in Figure 30. Within this network, several adjectives are grouped together in a semantic field labeled "Social Prominence." Interestingly, the adjectives *desempleada* "unemployed", *drogadicta* 'drug-addicted', *asalariada* 'salaried employee', *laica* 'lay', and *estupenda* 'wonderful.' Red nodes only collocate with women in their singular form, while *ejemplar* 'exemplary', *respetable* 'respectable', *excepcional* 'unique', *influyente* 'influential', *intachable* 'flawless', *todopoderoso* 'all-mighty', *trabajador* 'hard-working', *convicto* 'convict', *austero* 'austere', and *pobre* 'poor' collocate with men.

These associations suggest that men are more strongly associated with personal qualities and social prominence, while for women, only one adjective, "estupenda," is associated with such positive qualities.

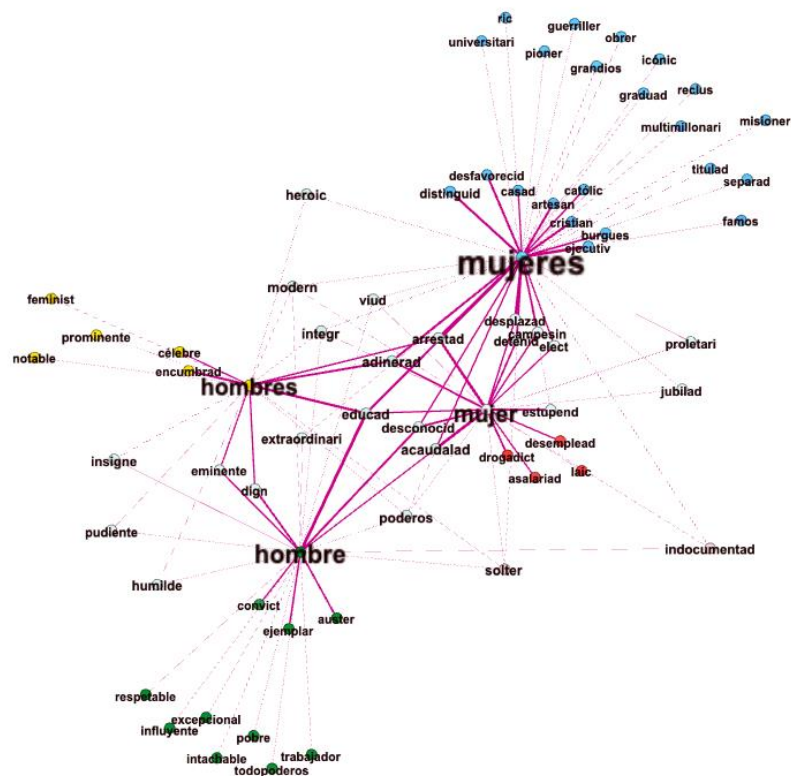


Figure 30. Constellation network of the collocations of the Social category

Moving on to the same semantic field and the adjectives that collocated with both lemmas, but in their plural form, we found that men were more likely to be associated with adjectives such as *célebre* ‘famous’, *encumbrado* ‘exalted’, *prominente* ‘prominent’, and *notable* ‘remarkable’, whereas women were more likely to be associated with adjectives such as *distinguidas* ‘distinguished’, *pioneras* ‘pioneers’, *multimillonarias* ‘multi-millionaire’, *famosas* ‘famous’, *ejecutivas* ‘executive’, *ricas* ‘wealthy’, and *icónicas* ‘iconic’. This suggests that in the NOW corpus, women are more often represented as being prominent, but only as a collective, not as individuals. Although this research does not focus on gender analysis or discourse analysis, it is worth noting this finding, as it implies that women are subject to a process of deindividuation when represented through language. This means that how women are represented as a community seems to overshadow how each woman is represented individually, which inadvertently marginalizes their identities when socially represented through language.

In this constellation network, it is noteworthy that certain adjectives such as *separada* ‘separated’, *casada* ‘married’, *cristiana* ‘Christian’, and *católica* ‘Catholic’ exclusively collocated with women, while *viuda(o)* ‘widow/widower’ and *soltera(o)* ‘single’ collocated with both genders. This finding aligns with previous research indicating that women are more frequently associated with adjectives related to marital status. Interestingly, there was no adjective related to marital status or religious affiliation that exclusively collocated with men in either their singular or plural forms.

Through the analysis of adjectival collocations in the previous four constellation networks, we can see how men and women are socially represented in the NOW corpus. Women tend to be associated with struggles and health issues, marital status, and religious affiliations. In contrast, men strongly collocate with fertility and sexuality issues, as well as mental sharpness and societal prominence. Additionally, both genders appear to experience a deindividuation process that prioritizes group identity over individual identity, resulting in stereotypes being placed upon members of the male and female communities without regard for their individual identities. This finding suggests a masculine bias towards considering men as the norm, with women being described in terms of their markedness.

While this analysis could be enriched by incorporating insights from other disciplines, such as discourse analysis and gender studies, the scope of this research is limited to demonstrating how tools and techniques from Corpus Linguistics can inform various fields. In the following section, I will present the verbal collocations and expand on the findings.

5.1.2 Verbs

A total of 3,200 adjectival and verbal collocations with a MI above 3 were extracted from the NOW corpus, as discussed in Chapter 5. The verbal collocations were classified according to the ADESSE verb classification, which is an ongoing project that addresses the different semantic classes to which a verb may belong, and some verbs may fall into more than one class. In this section, I will present some constellation networks to illustrate how collocation analysis can be used to uncover gender representations in language. The verbs were further classified into semantic fields when possible, similar to the adjectives. Figure

Some of these verbs were further grouped in a semantic field labeled Emotional Vocal Expression, which includes the verbs to grumble and to exclaim that collocated with man, and the verbs to appeal, to chant, to insist, to debate, to demand, and to clamor, which carry the connotation, at least in Spanish, of an intense or emotional verbal or vocal expression.

Among the listed verbs, "to perjure" has the highest MI score and only collocates with women. In contrast, verbs such as *to postulate*, *to redefine*, *to acknowledge*, *to deliberate*, *to testify*, *to notify*, and *to formulate* imply a standard method of communicating information and also exclusively collocate with women. Meanwhile, some verbs such as *discutir* 'to argue,' *mentir* 'to lie,' *amenazar* 'to threaten,' and *gritar* 'to yell' are more associated with men, while *protestar* 'to protest,' *callar* 'to be silent,' and *alegar* 'to argue' are more prominent with women. Notably, "to be silent" only collocates with women in phrases such as "women remain silent" 'las mujeres callan' or "we remain silent" (*mujeres callamos*), indicating that gender representations are often placed on the entire community rather than individual members, leading to deindividuation.

This finding raises the concept of "othering," which highlights the division between "us" and "them." The linguistic representation of gender further embeds this idea for both men and women. While constellation networks present data in an accessible interface, identifying phenomena such as the one described above requires reviewing the raw data. However, constellation networks have limitations as they only display lemmatized collocates (adjectives/verbs) and do not offer a comprehensive view of the data.

The aim of this research is to demonstrate how Corpus Linguistics techniques, specifically collocations, can inform qualitative studies. While this thesis does not delve deeply into gender studies or discourse analysis, I will dedicate a few paragraphs to exploring the concept of "othering" and its relation to gender representation. According to Jensen (2011), "othering" assumes that people who lack power are often relegated to being "the other" in discourse, which reinforces the legitimacy and superiority of those in power and shapes the identity of subordinates (p. 65). When we consider the notion that men are the norm and women are the "other," men become the ones who define and describe women, further perpetuating the objectification and otherization of women. A collocational analysis can help

researchers uncover and analyze the ways in which women are otherized and objectified, and it could be argued that the process of differentiation and demarcation highlighted by the collocational analysis described in this thesis reduces both women and men to stereotypes.

In terms of the specific data, the following constellation network displays the collocations of the Life category. In comparison to other categories, the Life category yielded fewer collocations, and most of the verbs in this category collocated with both lemmas.

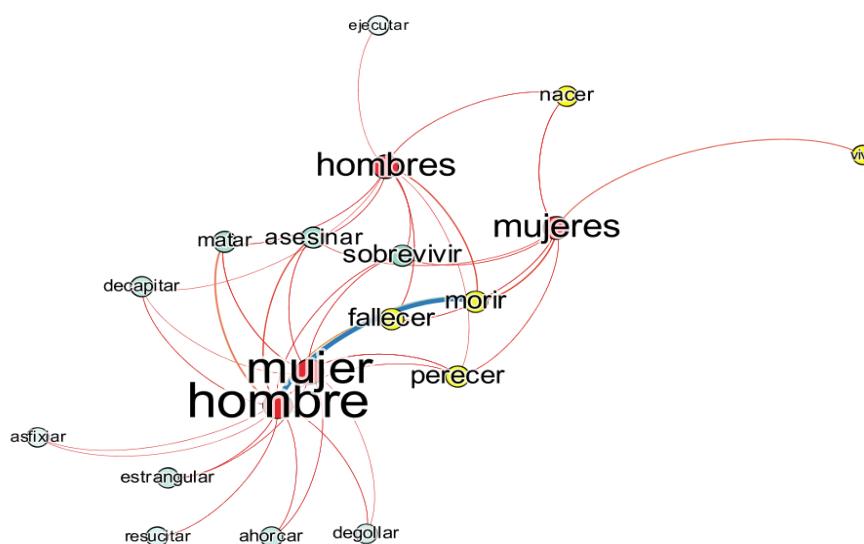


Figure 32. Constellation network of the category of Life (ADESSE)

In this analysis, the verb vivir ‘to live’ predominantly collocated with women, while ejecutar ‘to execute’ and resucitar ‘to resurrect’ collocated with men. Of these four verbs, only "ejecutar" had a negative connotation. Meanwhile, the verbs matar ‘to kill’, decapitar ‘to decapitate’, asesinar ‘to murder’, estrangular ‘to strangle’, degollar ‘to behead’, asfixiar ‘to asphyxiate’, and ahorcar ‘to choke’ all had negative connotations and collocated with both

genders, but were more commonly associated with men. For instance, “matar” had a higher MI score with men (7.21) than with women (5.22), and “degollar” had an MI score of 8.33 and 5.98 with men and women, respectively. These results align with other studies that suggest men are more often associated with violent acts.

Regarding non-violent verbs, morir ‘to die’ strongly collocated with men, while fallecer ‘to pass away’ was more commonly associated with women. However, because the focus of this analysis was on gendered subjects, it was not possible to explore how these verbs would behave if analyzed as objects. This would have allowed for an examination of not only who commits violent acts, but also who is on the receiving end. Nevertheless, these findings demonstrate the potential of collocational analysis in gender representation research.

Next, I will present the results from the Competition category in a constellation network. These verbs imply competition not only in sports, but also in various life contexts. The ADESSE classification not only categorizes verbs into classes but also considers their valences and the agents involved when the verbs are used with different meanings. For example, in the competition class, ADESSE considers not only the competitor(s) who participate in a competition, but also the competition itself and the antagonist(s). During a competition, competitors usually try to prevail over somebody or something. This collocational analysis of verbs, however, does not address the competitions or the antagonist(s) involved. Such analyses could be carried out using Corpus Linguistics techniques such as collocations or concordancers. Figure 33 shows the verbs related to competition. As observed, only a few verbs collocated exclusively with either man or woman. However, what makes these results interesting is the fact that some of them are antonyms.

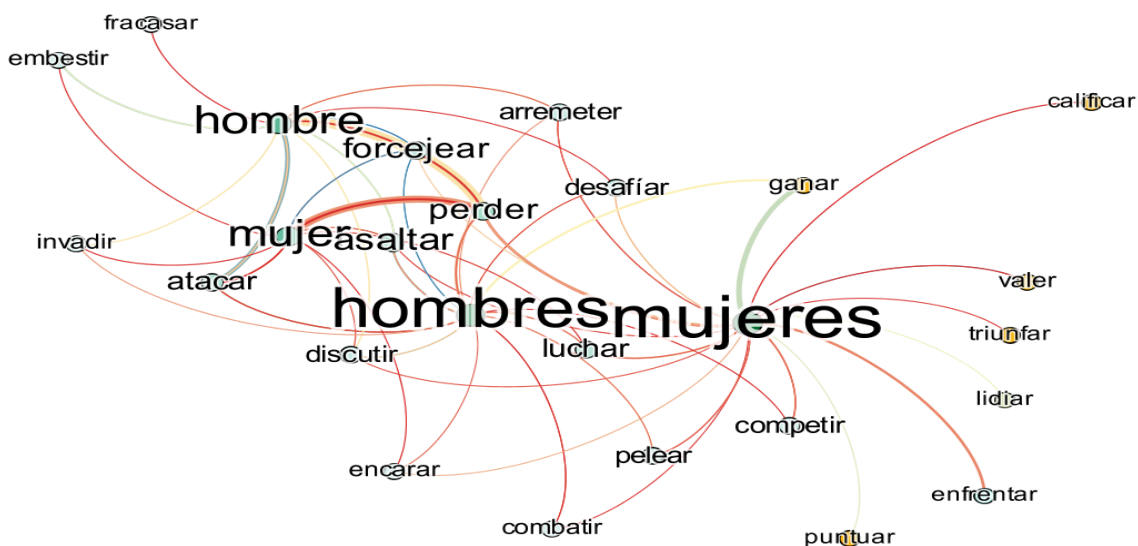


Figure 33. Constellation network of the category of Competition (ADESSE)

The verb fracasar ‘to fail’ only collocates with man, while triunfar ‘to succeed’, calificar ‘to qualify’, valer ‘to be worth’, lidiar ‘to deal with’, and enfrentar ‘to cope’ only collocate with woman. Among these verbs, lidiar and enfrentar imply facing struggles, and no such verbs collocate with man. Notably, fracasar collocates exclusively with man, and perder ‘to lose’ collocates with both genders. However, fracasar has a more negative connotation than "perder" in Spanish. Some verbs that collocate with both genders are luchar ‘to struggle’, competir ‘to compete’, desafiar ‘to challenge’, and encarar ‘to face’, which all pattern more strongly with woman and have no negative polarity, except for combatir ‘to combat’. In contrast, verbs like asaltar ‘to rob’, atacar ‘to attack’, invadir ‘to invade’, embestir ‘to charge’, and alegar ‘to argue’ pattern more strongly with man and all have negative polarity.

In the previous paragraphs, I demonstrated that verbs associated with emotional vocal expressions and formality tend to collocate more strongly with women, while actions

denoting violent behavior tend to collocate more strongly with men, which is consistent with previous research in the English language. The importance of adjectival and verbal collocation lies in the fact that while there has been extensive research on gender representation in English, studies on the topic in Spanish have been limited. In the following section, I will present the results of the automatic text classification and explain how the features have contributed to the accuracy of the algorithms.

5.2 Results of Machine Learning Experiments

In this section, I will present the results of the automatic text classification experiments. It is worth noting that these experiments were conducted using corpora created from content on the YouTube social network. Specifically, I built three different corpora, each labeled as *Viomujdis*, *Viogendis*, and *LGBTdis*, which focused on topics related to women, men, and the LGBT community, respectively. More information about these topics can be found in the methodology chapter.

It is important to provide some context on how this research study progressed. In the first stage, I analyzed the NOW corpus using collocations to investigate gender representations of women and men through adjectives and verbs. Although I aimed to expand our analysis, working with fourth-generation concordances (like the NOW corpus) presented a major disadvantage, as users can only run searches and not access the corpus. To address this limitation, I decided to build our own corpus, which allowed me to investigate whether my previous findings from the NOW corpus would hold true and to conduct more comprehensive analyses beyond collocations.

During the construction of the YouTube corpus, we encountered unexpected issues, the most significant of which was related to the nature of the corpus itself. Unlike the NOW corpus, which comprises edited newspaper articles, the YouTube corpus contains user comments that often disregard writing conventions. As explained in the earlier chapters of this thesis, the comments were riddled with spelling mistakes, making collocational analyses difficult. Additionally, unlike the NOW corpus, which allowed us to investigate the lemmas MAN and

WOMAN, conducting a similar search in the YouTube corpus was futile due to the varied ways that men and women were referred to in the comments. For example, men were referred to as “típo” or “viejo,” and women as “Viejas” or “señoras,” among other terms. These issues arose because social networks do not adhere to traditionally accepted research conventions. Nonetheless, we aimed to maintain the focus on gender representation and divided the YouTube corpus into two sub-corpora: one that addressed topics related to men and another that addressed topics related to women, as discussed in the previous chapter. Later, we added a third corpus that focused on topics related to the LGBT community.

Once these corpora were constructed, I performed a keywords analysis to identify terms that related to women, men, or the LGBT community. It is important to note that keywords refer to the primary topics of the texts. With these results in hand, I employed techniques from the NLP field to conduct automatic text classification. A detailed explanation of the keyword analysis process is provided in the earlier chapter.

In the upcoming sections, I will present the results of our various automatic text classification experiments. It is worth highlighting that this research study's major contribution to Natural Language Processing was our use of keywords in these experiments.

5.2.1 Preliminary Experiment of “Violentómetro”

Based on the findings obtained with the NOW corpus which allowed me to analyze gender representations through collocations, in this second stage, I utilized keywords to classify the same texts from which these keywords were derived. For the sake of clarity, it is worth clarifying that by texts I refer to the comments expressed in each one of the YouTube videos.

Before testing the keywords to assess if these could accurately classify the texts, I carried out a preliminary classification in which I utilized the verbs used in the Violentómetro. Since the objective was to classify texts through the use of keywords that relate to women and men (see chapter 5), I sought to explore how the Violentómetro classification would perform when using its information to classify the texts. Some of the verbs that I utilized in this first

experiment were: asesinar ‘to kill’, violar ‘to rape’, and humillar ‘to abase’ among others; it is important to note that some of these verbs were also found in the NOW corpus and to kill appeared with both lemmas but the highest MI score (asesinó) appeared with man. The verb violar ‘to rape’ appeared only with man and the verb manosear ‘to grope’ collocated more strongly with man whereas the verb humillar ‘to abase’ only collocated with woman. Based on the above which seeks to explain how this research study progressed from a collocation analysis to a keyword analysis whose results were extrapolated to automatic text classification, Table 29 shows part of the vector space model (VSM) which represents the frequency of each one of the 13 features (verbs used in the Violentómetro) in the two classes (18 Viomujdis texts and 17 Viogendis texts).

Class	asesinar	violar	abusar	amenazar	manosear	golpear	controlar	intimidar	humillar
Viomujdis	0	1	1	0	1	0	1	0	0
Viomujdis	0	3	2	0	0	0	4	0	0
Viomujdis	0	3	0	0	1	2	0	4	3
Viomujdis	1	4	0	1	0	1	0	1	1
Viomujdis	4	5	2	0	0	1	2	0	0
Viomujdis	0	4	0	0	1	1	1	0	0
Viomujdis	1	3	0	0	0	1	0	0	1
Viomujdis	0	2	1	0	0	1	1	0	0
Viomujdis	0	3	1	0	0	0	2	0	0
Viomujdis	0	1	0	0	0	2	1	0	0
Viogendis	1	0	0	0	0	1	0	0	0
Viogendis	7	1	0	1	0	1	1	1	0
Viogendis	1	0	0	0	0	0	0	0	0
Viogendis	0	0	0	0	0	1	3	0	0
Viogendis	1	0	0	0	0	0	4	3	0
Viogendis	2	1	0	0	0	1	2	1	0
Viogendis	0	7	9	0	0	0	2	0	1
Viogendis	0	6	1	0	0	2	5	0	2
Viogendis	0	3	0	0	0	0	8	0	0
Viogendis	0	1	0	0	0	0	2	0	0
Viogendis	0	0	1	0	0	0	1	0	0

Table 29. Vector space model representing the frequency of features in two classes. (Violentómetro)

After identifying the frequency of each feature in the texts using the AntConc Corpus linguistics tool, the VSM was processed using machine learning software to evaluate how accurately the features could classify the texts. In this initial attempt at classification, we

utilized Naïve Bayes, Sequential Minimal Information implementation of Support Vector Machines (SVMs), and J48 decision tree with 10-Fold cross-validation. The results obtained are as follows:

Weighting scheme	Naïve Bayes	Support Vector Machines (SVMs)	J48
Frequency	62%	74%	54%

Table 30. Results obtained in the experiment. Features taken from the Violentómetro

The SVM algorithm produced the most favorable results, indicating that the verbs used as features were somewhat significant and relevant for text classification. It is essential to note that features are always extracted from the same texts in text classification. Hence, the results were acceptable, considering that the features used in this experiment did not come from the corpora. Consequently, I initiated a feature engineering process to identify the most accurate features that could enhance the algorithm's performance in text classification.

To improve the existing results, I utilized the regular expression (*) technique that expands the search of other strings (words) through a sequence of characters. For instance, the regular expression of viol* can find words such as violó, violaron, violan, etc. within a dataset. Additionally, I included the LGBTdis corpus (16 texts) to the automatic classification task, making the task more challenging for the algorithm. Table 31 displays the VSM with the three classes and the features. However, due to the VSM's size, a shortened version is presented.

Class	Asesin*	Viola*	Abus*	Amenaz*	Golpe*	Manose*	Control*	Intimidar*	humill*
LGBTdis	0	4	0	0	0	1	4	1	0
LGBTdis	2	11	3	0	2	0	4	0	1
LGBTdis	2	1	4	0	3	0	0	0	0
LGBTdis	0	1	1	2	0	0	2	0	0
LGBTdis	1	1	1	0	4	0	1	0	1
LGBTdis	3	2	1	0	1	0	1	0	1
LGBTdis	3	5	3	0	1	0	1	0	0
Viogendis	5	3	3	8	6	0	8	0	0
Viogendis	95	11	6	6	9	0	9	1	3

Viogendis	3	1	18	1	29	0	0	0	6
Viogendis	25	2	2	1	5	0	4	0	4
Viogendis	39	8	4	4	2	0	10	3	2
Viogendis	34	12	7	4	2	0	12	1	0
Viogendis	22	120	64	5	5	1	5	0	2
Viomujdis	2	45	11	2	5	8	6	1	2
Viomujdis	7	30	34	13	7	0	12	0	5
Viomujdis	4	20	9	3	9	4	0	4	13
Viomujdis	57	125	16	8	15	0	0	1	5
Viomujdis	60	112	19	7	13	0	5	0	1
Viomujdis	59	66	20	3	41	1	20	0	9
Viomujdis	130	79	24	3	21	0	1	1	3

Table 31. Vector space model representing the frequency of features in three classes. (Violentómetro)

The expanded search through regular expressions is evident in the increased frequency of each feature in each class, as shown in the updated VSM. The results obtained from running the same algorithms with these adjusted features are displayed in Table 32.

Weighting scheme	Naïve Bayes	Support Vector Machines (SVMs)	J48
Frequency	74%	82%	64%

Table 32. Results obtained in the experiment once Features were adjusted (Violentómetro)

The SVMs proved to be the most accurate algorithm for classifying the classes, achieving an accuracy of 82%. Interestingly, all the algorithms performed better once the features were adjusted using regular expressions. This first experiment, which used the information from the Violentómetro, provided a baseline for comparison with the keywords obtained from the same corpora in the next section. In the next section, I will discuss how the keywords extracted from the YouTube corpora were used as features to test the accuracy in the ATC experiments.

5.2.2 Automatic Text Classification (Video Files)

In this first major experiment, the goal was to use keywords as features for classifying texts. However, with hundreds of keywords to choose from, the task at hand was to select the most effective ones for the classification process. To achieve this, I identified the keywords with

the highest keyness and used my intuition to select 10 features (keywords) per class. The selected features were regular expressions of words such as *God, respect, to accept, rights, to discriminate, equality, to harass, to defend, to deserve, to assault, prostitute, to denounce, to kill, to bless, sin, bible,* and *Amlo* (acronym referring to the Mexican president), among others.

Given the vast amount of information, a partial view of the VSM is shown in Table 33. As previously mentioned, the selection of these features was based on a combination of the keywords' keyness and my own judgement.

Class	Dios	Derecho*	Discrimina*	Iguald*	Acosa*	Mata*	Soy	Biblia	Pecado*	Odi*	Poder	Amlo
LGTBdis	22	22	11	0	2	5	151	4	6	12	6	0
LGTBdis	173	52	22	5	5	8	51	17	23	19	9	0
LGTBdis	153	23	43	1	1	10	60	31	18	30	8	0
LGTBdis	289	36	11	6	0	11	11	23	23	28	10	0
LGTBdis	267	3	15	4	3	3	65	29	4	27	0	0
LGTBdis	326	20	12	5	1	8	112	42	75	30	9	0
LGTBdis	232	46	19	2	0	13	75	55	29	32	6	0
LGTBdis	4	2	5	2	0	5	531	1	2	16	4	0
Viogendis	22	22	0	1	0	12	25	4	2	12	38	27
Viogendis	271	35	0	0	2	88	9	4	7	2	62	31
Viogendis	12	13	211	1	14	8	84	0	0	14	6	0
Viogendis	17	5	0	0	0	19	26	0	0	11	49	277
Viogendis	28	70	0	0	0	128	12	1	1	2	33	98
Viogendis	69	44	1	2	2	36	16	0	1	3	48	205
Viogendis	27	7	1	0	0	25	32	36	22	15	38	5
Viomujdis	8	29	6	9	562	5	49	0	4	4	12	0
Viomujdis	76	4	0	0	1	40	47	3	3	4	24	0
Viomujdis	17	40	8	36	126	2	35	0	0	30	9	0
Viomujdis	25	41	0	0	5	68	17	1	0	5	20	7
Viomujdis	160	36	1	1	0	56	20	0	1	5	21	2
Viomujdis	29	45	6	41	17	104	51	1	0	17	16	0
Viomujdis	101	45	5	42	11	152	26	0	0	34	27	3
Viomujdis	8	63	16	238	3	23	72	1	0	27	20	0

Table 33. Vector space model representing the frequency of features in the three classes (keywords)

In this experiment, we aimed to evaluate the relevance of the selected keywords to each text. It is important to note that the texts revolve around topics related to issues that affect women, men, and members of the LGBT community, and these keywords directly relate to the topics

covered in each text. To test the accuracy of classification, the VSM was once again run using the same algorithms, and the results are presented in Table 34.

Keywords as features			
Weighting scheme	Naïve Bayes	Support Vector Machines (SVMs)	J48
Frequency	98%	96%	84%

Table 34. Results obtained in the experiment with keywords as features (30 features)

The results of the previous experiment show that the accuracy of the algorithms increased significantly when accurate features were used, which can be identified through keyness since they signal aboutness of the texts. To verify if other keywords would also yield similar results, we designed another VSM with 243 words, selecting the keywords with the highest keyness. These keywords included carcel ‘prison’, catolic* ‘Catholic*’, cristian* ‘Christian*’, chairo*, corrupt* ‘corrupt’, impunidad ‘impunity’, and racis* ‘racism’, among others.

Keywords as features			
Weighting scheme	Naïve Bayes	Support Vector Machines (SVMs)	J48
Frequency	98%	98%	86%

Table 35. Results obtained in the experiment with keywords as features (243 features)

The previous experiments showed that the selected keywords are effective features for automatic text classification, as they directly relate to the topics discussed in the texts. To further test their performance, I decided to investigate whether these features would be equally effective for classifying comments instead of texts. For this purpose, a new VSM was created using a set of comments from the same sources, and the same set of 243 keywords was selected based on their keyness.

The classifiers were then trained and tested using the new VSM, and the results were consistent with the previous experiments. All classifiers maintained a high level of accuracy, indicating that the selected keywords are indeed reliable indicators of the aboutness of the

comments as well. These findings suggest that the same set of features can be used for classifying both texts and comments, which could potentially reduce the time and effort required for feature selection in future experiments.

5.2.3 Text Automatic Classification (users' comments)

The second major experiment aimed to classify the comments within each text by using all words in each sentence as features, instead of just the keywords. To achieve this, a string-to-word vector filter was utilized, which converted each comment (string) into a vector of words, where each word in the string became a feature. This approach is known as Bag of Words (BoW) and is commonly used in the field, as mentioned in previous research. The filter employed a Boolean weighting scheme to test the classifier based on the presence or absence of features in a string. In other words, if an algorithm detected certain features in a comment (string), then it would classify the sentence based on the previously identified classes, allowing the algorithm to learn from these findings.

To clarify, we have been using three classes (LGBTdis, Viogendis, Viomujdis), each consisting of 16, 17, and 18 texts, respectively. In this experiment, all the comments within each text were reclassified, meaning that all comments within the LGBT texts were labeled as V-LGBT, those within the Viogendis texts were labeled as V-General, and those within the Viomujdis were classified as V-Mujer. As stated in chapter 5, this second major experiment involved the automatic classification of 7,500 comments, 2,500 per class. The classification was conducted using the same algorithms employed in the previous experiments and 10-fold cross validation.

"J ahora todo es la palabra de dios ",V-LGBT

"Que pena la homosexualidad no es natural porque Dios creo hombre y mujer, no creo un hombre para tener relaciones sexuales, si dios no existía todo esto estaría permitido, pero no es natural acaso es natural que un hombre tenga relaciones sexuales con un mono es natural, pero todos somos libres ya sé que no es lo mismo pero no es natural hasta científicamente no puede haber creación, un hombre con una niña de años teniendo relaciones sexuales está bien? Hasta si la niña lo quiere, pero todos somos libres, no es algo Natural", V- LGBT

"La palabra de dios la doña ",V- LGBT

" , Muchos grandes proyectos que hizo cuando fue jefe de gobierno, tal como la licencia permanente, el segundo piso, el tramo que no se cobra y muchos otros y ahora que es mi presidente AMLO, hay la lleva efectuando cambios verdaderos por el bien de Mexico." V-General

"Jajaja jajaja lo que mas les purga a toda esa lacra es su sonrisa Sr presidente animo mi chingonazo AMLO",V-General

"Sus amistades no son de mi agrado presidente, pero sigo siendo amlover", V-General

" ,Y según equidad de genero, Que hijos DP me cagan i",V-Mujer

" ,La equidad de género no existe.",V-Mujer

" ,Aquí se ve claramente como la sociedad es doble moral si defendemos a una mujer cuando es agredida, pero cuando vemos que agreden a un hombre hasta se rien es por esto que a veces no tomo en enserio la equidad de género", V-Mujer

" ,ahi esta su equidad de género",V-Mujer

Figure. 34 Samples of comments being classified

Figure 34 displays a selection of the 7,500 comments that were automatically classified in this experiment. Each comment was treated as a string, and every word in the string was considered a feature. By analyzing which words were present in certain classes and which were not, the classifier learned to associate words with each class. As for the results, Table 36 summarizes the outcomes of this experiment. Despite the considerable increase in the number of features used in the classification, the algorithms maintained their good performance across the three classes, as demonstrated by the results obtained with the three different algorithms.

String to Word Vector			
Weighting scheme	Naïve Bayes	Support Vector Machines (SVMs)	J48
Boolean	92%	91%	85%

Table 36. Results obtained when classifying comments. (1,756 features)

Before determining the significance of keywords in the algorithm’s accuracy, one final step remained. Among the 1,756 features, 203 were previously identified as keywords. To assess their impact, we conducted another experiment where we excluded these features. The results of both experiments are summarized in Table 37.

Algorithm	10-Fold cross validation	66%: 34% (training-test data)	Without Keywords
Naïve Bayes Multinomial	92%	91%	77%
SVM	91%	91%	74%
J48	85%	83%	64%
ZeroR	33%	33%	33%

Table 37. Results obtained when the keywords were excluded

Table 37 depicts the results of the experiment, which involved 10-fold cross-validation and a random split of data into 66% training and 34% testing. By partitioning the data and training the algorithm on the training dataset, while evaluating it against the test dataset, we ensured the independence of the results from the training data set. Notably, the Naïve Bayes algorithm produced the most accurate results, with minimal variation between partitioning techniques. However, Table 37 also includes results obtained after excluding the keywords (features). This modification decreased the accuracy of the Naïve Bayes, SMO, and J48 algorithms by 15%, 17%, and 19%, respectively. Therefore, the inclusion of keywords appears to be a useful technique in the NLP field, which helps to evaluate the precision of classification tasks.

In this section, I explained the process of identifying features, specifically keywords, from the YouTube corpora. I also outlined the steps taken to enhance the accuracy of the algorithms across different ATC tasks. Traditionally, the Natural Language Processing (NLP) field has relied on standard features such as character n-gram, token n-grams, bag-of-words, and embedding to evaluate their models and determine the one that provides the best results. Some researchers have also integrated linguistic features such as morphology, pragmatics, figures of speech, punctuation, and symbols to evaluate their models.

By adopting a technique from Corpus linguistics and applying it to the NLP field, this study has provided valuable insights. The results indicate that keywords are a useful feature for text classification tasks, thereby expanding the list of features that NLP researchers can consider to improve the accuracy of their models. As such, this finding has the potential to advance the field of NLP and help researchers achieve more accurate results in text classification tasks.

This chapter presents two major sections. The first section focuses on the collocational analysis of the lemmas MAN and WOMAN. Through this analysis, it was revealed that certain adjectives and verbs collocated exclusively with these lemmas, highlighting gender representations for both men and women. Furthermore, the findings suggested that deindividuation and othering processes were apparent for both genders.

The second section discusses text classification experiments, where keywords were employed as features to enhance the accuracy of the different classification tasks. Specifically, keywords with high keyness scores were selected to function as features, and these proved to be effective across the experiments. Additionally, the chapter describes another text classification experiment that relied on a string to word vector filter. The results of this experiment indicate that keywords can help algorithms discriminate among almost 7,500 comments.

Overall, this chapter demonstrates that keywords, a technique derived from Corpus Linguistics, can be adapted to the NLP field for text classification tasks. These findings can contribute to the advancement of the NLP field and improve the accuracy of text classification tasks.

6 Discussion

The primary objective of this thesis was to address a series of research questions and confirm or refute the hypotheses proposed. This chapter presents a synthesis of the results and provides answers to the research questions, which are presented in the following order:

1. **How can corpora built from online social networks help reveal gender representations?**
 - a) Data obtained from social networks can provide unique insights that are not typically available through traditional data collection methods, allowing marginalized communities to have a platform to express their opinions.

2. **To what extent is keyword analysis an effective feature selection method in machine learning, and how can its efficiency be measured?**
 - b) *Keyword analysis, as applied in corpus linguistics, can serve as a feasible feature selection method for those without technical expertise in traditional machine learning feature selection methods.*
 - c) *As keywords reflect the main focus of a text, their use in algorithms enhances or at least maintains the accuracy of ATC tasks.*

3. **What are the differences between traditional corpora and those built from online resources?**
 - d) *Differences between traditional and online CMC corpora can be attributed to differences in the corpus nature, sample selection, and content quality.*

4. **What are the possibilities and complexities involved in constructing a corpus from the web?**

e) Collecting linguistic data from the web provides access to a wealth of specialized, current, and detailed material on various topics.

Gender representations

Gender representation has been a central theme in my research from the beginning. To compile the corpus for my study, I gathered comments from YouTube on topics such as femicide, sexual harassment, gay rights, sexist language, and drug trafficking, among others. As I progressed with the automatic text classification tasks and attempted to operationalize keywords, the linguistic data consistently revealed how women, men, and members of the LGBT community were represented in language. While hate speech classification is a well-developed area of research in machine learning, including misogyny, xenophobia, and immigration (Poletto et al., 2020; Anzovino, 2018), I felt it was necessary to take my research, which focused on automatic text classification, a step further. Thus, I conducted a collocational analysis to explore how men and women were linguistically represented in the YouTube corpus.

As I have highlighted in this thesis, the nature of the YouTube corpus made it difficult to conduct the collocational analysis, but I was able to carry out the analysis using the NOW corpus. In the following paragraphs, I will answer the first research question and its hypothesis.

How can corpora built from online social networks help reveal gender representations?

a) Data obtained from social networks can provide unique insights that are not typically available through traditional data collection methods, allowing marginalized communities to have a platform to express their opinions.

In the previous chapter, I discussed the results of an analysis conducted on the NOW corpus, in which 1,586 adjectival collocations for the lemmas MAN and WOMAN, with an MI above 3, were extracted and classified according to the Supersenses classification. To showcase the patterns of adjectives that co-occur with both lemmas, I used constellation networks. Out of the 1,586 adjectives, 814 were chosen for analysis, and it is worth noting that more adjectives collocated with man than with woman.

In the analysis, I identified adjectives that exclusively patterned with either man or woman and other cases where adjectives collocated more strongly with either the singular or plural form. An interesting finding was that the adjective *adúltera* ‘adulteress’ only collocated with woman, while *infiel* ‘unfaithful’ patterned with man. It is worth noting that these two adjectives are sometimes used interchangeably, despite having different legal meanings. ‘*Adúltera*’ had the highest MI score among all the adjectives analyzed in this collocational analysis, which suggests that every time this word appears in the NOW corpus, it is highly likely to appear exclusively with the word ‘woman.’ In a legal context, adultery is grounds for seeking divorce and is defined as having sexual relationships with someone while being married to someone else. In contrast, “unfaithful” is a vague and subjective term in legal contexts.

I argued that this particular example shows that women are more harshly represented in language. The data from the NOW corpus suggests that men are unfaithful, while women are adulteresses. The choice of words used to represent women, as in the case of the NOW corpus, could go beyond mere linguistic representation and seem to criminalize women’s behavior, but not men’s behavior. This could be an erroneous generalization, but analyzing this kind of data can uncover how women and men are linguistically represented.

Table 38 presents some of the findings of the collocational analysis, and it is worth noting that the adjectives with an asterisk (*) can appear with both the singular and plural forms of the lemmas.

	MAN	WOMAN
Behavior (struggling)		insumisa, combativa, aguerrida, intrépida, esforzada, verraca, luchona
	infiel, posesivo, polígamo	adultera*, promiscua
		abnegada*, devotas
	maltratador*, necio, colérico*, codiciosos, avariciosos	corajuda, chismosas
Body (age)	jovial, mayores	longeva, madura*
(health)	infértil, estéril, pálido, autista, ciego, invidente	adolorida, anoréxica, muda, hipertensas, asmáticas, desnutridas, asintomáticas, menopaúsicas, sordomudas, sordas, desvalidas
(physical appearance)	viril, barbado, velludos, sexies, apuesto*, peludo*	agraciada, preciosa, glamorosa, bellas, lindas, hermosa*, linda*
(body-type)	regordete, fornidos, musculoso*, fuerte*	anoréxica, esbelta*, voluptuosa*, atractiva*, curvilínea*
(sexuality)	bisexuales, heterosexual*, homosexual*, transexual*	frígida, lésbicas, lesbiana*
Mind (instruction/plural)	sensatos, pensantes, habilidosos, experimentados, ilustres	cualificadas, tituladas, organizadas, preparadas, universitarias, científicas, expertas, estudiadas
(Mental acuteness/singular)	tenaz, cuerdo, congruente, juicioso, previsor, meticoloso, sagaz, ecuánime, inteligente, reflexivo, lúcido, mesurado	Precavida
Social prominence	ejemplar, respetable, excepcional, influyente, intachable, poderoso, trabajador, convicto, austero, pobre	desempleada, drogadicta, asalariada, laica, estupenda
	celebres, encumbrados, prominentes, notables	distinguidas, pioneras, multimillonarias, famosas, ejecutivas, ricas, icónicas
Marital status/religion	soltero, viudo	separada, casada, soltera, viuda, cristiana, católica

Table 38. Adjectival collocation of MAN and WOMAN

Table 41 presents a selection of verbal collocations extracted from the NOW corpus, classified into three groups. It is worth noting that several verbs are associated with both lemmas, and there are various categories among them. The purpose of exploring adjectival and verbal collocations was to assess how automatic text classification of hate speech could be extended to a comprehensive analysis of gender portrayals.

	MAN	WOMAN
Communication	to greet, mention, thanks, faint, swear, grumble, testify, exclaim	to postulate, appeal, chat, redefine, insist, debate, acknowledge, deliberate, demand, testify, assure, perjure, blow the whistle, notify, formulate, predict, claim
Life	to execute, to resurrect	to live
	to kill, decapitate, murder, strangle, behead, asphyxiate, choke (<i>also collocated with WOMAN but were much more salient with MAN</i>)	
Competition	to fail, lose	to succeed, qualify, to be worth, to deal with, cope, lose

Table 39. Verbal collocations of the lemmas MAN and WOMAN

In the previous section, I discussed the intricacies of working with data from the web, particularly from social media platforms. I noted that traditional sources, such as books and newspapers, tend to be filtered and edited for correctness and appropriateness, often presenting only a partial truth about gender portrayals. However, the online world, including social media, offers a unique avenue for exploring gender representations. Unlike traditional sources, social media platforms allow ordinary individuals and marginalized communities to participate, making them vital spaces for cultural expression. Online spaces enable users to develop and showcase their identities, share their values, engage with others, negotiate meaning, and encounter diverse cultures.

Social media platforms also provide anonymity and the opportunity for antagonistic discourse, which can lead to conversations and practices that may be avoided in face-to-face interactions. These platforms offer spaces where controversial topics, such as gender inequality, femicide, and LGBTQ+ rights, can be discussed, enabling engagement with diverse communities across societies. Based on these factors, online corpora, particularly those from social media, can help reveal gender representations as people are more likely to express themselves without fear of criticism.

Keywords as features in automatic text classification tasks

Automatic text classification presents a major challenge of selecting appropriate features to achieve accurate classification. With thousands of words in a dataset, it is crucial to identify linguistic units that yield optimal results. Commonly used linguistic features include parts of speech, stylometry, n-grams, syntax, and sociolinguistic features, among others (Garcia-Díaz et al., 2021; Pang & Lee, 2008). Feature selection methods such as bag of words, TF-IDF, Mutual Information, and Best Terms are widely employed to identify relevant features (Deng et al., 2019).

In this study, I will summarize the detailed results presented in the previous chapter, and answer the research question while expanding on the hypotheses.

To what extent is keyword analysis an effective feature selection method in machine learning, and how can its efficiency be measured?

b) Keyword analysis, as applied in corpus linguistics, can serve as a feasible feature selection method for those without technical expertise in traditional machine learning feature selection methods.

c) As keywords reflect the main focus of a text, their use in algorithms enhances or at least maintains the accuracy of ATC tasks.

The primary objective of this study was to investigate the effectiveness of using keywords obtained through traditional corpus linguistics procedures as features in automatic text classification (ATC) tasks. The ATC tasks involved a frequency weighting scheme and a set of 48 texts, comprising 18 Viomujdis texts, 17 Viogendis texts, and 16 LGBT texts.

To establish a baseline for comparison, I initially conducted a preliminary ATC task using information (verbs) from the violentómetro as features. Subsequently, I carried out two more ATC tasks, employing keywords extracted from the same texts as features. It is worth noting that the texts in the corpora pertain to issues concerning women, men, and the LGBT community. The goal of using keywords as features was to determine their efficacy in ATC.

Table 38 illustrates the results for each ATC task and the keywords utilized in each of them. As previously mentioned, different algorithms were employed in the ATC tasks, but only the best-performing algorithms for each task are presented in Table 38.

	1st ATC task (Violentómetro- UAEM)	2nd ATC task (30 keywords)	3rd ATC task (242 keywords)
Keywords	asesin*, viola*, abus*, amenaz*, manose*, control*, menti*, intimidar*, humill*, golpe*, cachetea*, ofend*	Dios, respet*, acept*, derecho*, discrimina*, iguald*, acosa*, merec*, agred*, prostitu*, mata*, bend*, provoca*, soy*, biblia*, mujer*, hombre,*amlo	bullying, asil*, armar*, ayotzinapa, caravana*, catolic*, chairo*, corrup*, cree*, deporta*, fifi*, crim*, impunidad*, mediocre*, mafia*, politic*, racis*, pendej*, bisexual, creyente*, gomorra, prejuicio, etc.
Algorithms with the best performance	SMO	Naïve Bayes	Naïve Bayes
Accuracy	74% (2 corpus without regular expressions) / 82% (3 corpus with regular expressions)	98%	98%

Table 40. Accuracy of the best algorithms

The results indicate a significant improvement in the accuracy of the algorithms once the keywords were included in the last two ATC tasks. It is important to note that the keywords were obtained using traditional corpus linguistics procedures, which are crucial in the natural language processing field from which the classification tasks originate. The positive results

obtained in the first ATC tasks confirm the effectiveness of the CL traditional procedures in yielding acceptable results in the ATC tasks.

To further test the efficacy of the keywords, a new and different experiment was conducted, which involved categorizing almost 7,500 comments in three classes (V-Mujer, V-General, V-LGBT) using a Boolean scheme. Unlike the first experiment, the ATC task was run with both a 10-fold Cross-validation and with the data being split 66%: 34% as training and testing, respectively. In this new experiment, a string to word vector filter was used, which meant that all the words in the comments became features (keywords) and could be automatically taken into account in the classification tasks.

The results in Table 41 show that the Naïve Bayes Multinomial algorithm achieved the highest accuracy of 92% when a 10-fold Cross-validation was used. However, the most significant finding was that the precision of the algorithm decreased by 15% when the keywords were removed.

In summary, the results demonstrate that incorporating keywords obtained through traditional corpus linguistics procedures can significantly improve the accuracy of the algorithms in ATC tasks. The second experiment further confirms the importance of keywords in achieving high accuracy and highlights the negative impact of removing them.

Algorithm	10-Fold Cross-validation	66%: 34% (training-test data)	Without keywords
Naïve Bayes Multinomial	92%	91%	77% (76%)

Table 41. Results of the automatic text classification with the string to word vector filter

The latest ATC task provides further evidence that the incorporation of keywords improves the accuracy of ATC tasks. The use of keywords in this context is significant as it refers to the aboutness of both the texts and comments, thereby functioning as operationalized features in corpus linguistics. While other linguistic features, such as figures of speech, pragmatics, morphology, grammar and spelling mistakes, parts of speech, punctuation, and symbols are

often employed in ATC tasks, the use of topic features (i.e., keywords) in my study effectively captured the topic in both the texts and comments.

The relevance of keywords as topic features lies in their ability to be traced back to specific topics. For instance, words like *puta* (bitch) or *inclusivo* (inclusive) in gender-inclusive language are highly likely to be associated with women’s affairs. Therefore, in the second major experiment, by converting comments into feature vectors, the most salient and critical features (i.e., keywords) were made available, which had already been identified, but their relative importance had not been determined until this point.

Table 42 displays the keywords with more information gain, which indicates that these words were the most relevant when classifying the nearly 7,500 comments according to the three pre-established classes. While some of the keywords may appear self-evident, a more thorough and qualitative analysis would be necessary to comprehend the significance of others. The table includes only 32 features, but a more exhaustive list could help to elaborate on those that are offensive, clearly related to women, men, or the LGBT community, or those that are grammatically feminine, such as *las*, *una*, *señora*.

In conclusion, the results underscore the importance of keywords in ATC tasks and demonstrate their effectiveness as topic features in capturing the aboutness of both texts and comments.

Information Gain Ranking					
1	0.103781	1404 AMLO	17	0.032137	834 respeto
2	0.07845	521 igualdad	18	0.031663	990 violencia
3	0.077355	1180 <i>feminismo</i>	19	0.029622	170 biblia
4	0.074765	645 <i>mujeres</i>	20	0.027864	1509 amlo
5	0.070765	332 dios	21	0.027849	1226 lenguaje
6	0.061722	898 <i>soy</i>	22	0.027767	1329 <i>puta</i>
7	0.054608	494 hombres	23	0.027073	493 hombre
8	0.053228	644 <i>mujer</i>	24	0.026572	1122 corrupción
9	0.044589	1683 presidente	25	0.026189	551 <i>las</i>
10	0.041107	1221 justicia	26	0.023235	62 México
11	0.038916	1082 <i>acoso</i>	27	0.022519	816 religión
12	0.037952	438 género	28	0.02094	434 <i>gays</i>
13	0.036661	433 <i>gay</i>	29	0.020555	873 <i>señora</i>

Information Gain Ranking					
14	0.035091	599	matrimonio	30	0.020532 951 una
15	0.032765	500	homosexuales	31	0.02014 1250 maltrato
16	0.032527	1208	inclusivo	32	0.019915 441 gobierno

Table 42. Ranking of the attributes (keywords) with higher information gain

The results of the classification tasks clearly demonstrate the effectiveness of using keywords to improve accuracy, with a significant drop in accuracy noted when the keywords were not included. In response to the first research question, it is evident that the Naïve Bayes Multinomial model achieved an impressive overall accuracy of 98%, outperforming the baseline. This further emphasizes the importance of using feature selection methods to improve classification tasks, which in turn can optimize network resources and enhance precision.

In addressing the second research question, my research demonstrates the feasibility of using keyword analysis as a feature selection method in machine learning. While my approach may be less automated than traditional machine learning methods, it offers a novel way to select features that draws from traditional corpus linguistic techniques. My study underscores the potential for further interdisciplinary collaboration between corpus linguistics and machine learning, which can yield valuable insights and tools to improve ATC tasks.

My research study makes several valuable contributions. First, it proposes the use of keywords as potential features for ATC tasks, as operationalized in corpus linguistics. This approach can be applied to the automatic detection and classification of hate speech in the Spanish language, contributing to the advancement of this field.

Additionally, my research study involves the compilation of a corpus to conduct the ATC tasks and evaluate the effectiveness of the keywords. This contributes to the ongoing development of web-based corpus construction, which is an important topic in corpus linguistics.

Finally, I hope that my work will encourage other corpus linguists to establish interdisciplinary connections and collaborate with other fields, broadening the possibilities for research and expanding our understanding of language.

Corpora and the web

The internet offers a vast amount of linguistic data that can be utilized for various types of linguistic analyses. For language researchers, the web is a valuable source of information, providing access to the kind of language they are interested in and allowing them to access data that may not be available in edited or written texts. This linguistic data has been utilized to construct a wide range of corpora, but has also revealed discrepancies regarding the criteria that corpora should meet. In this section, I will address the following research questions and hypotheses.

3. What are the differences between traditional corpora and those built from online resources?

d) Differences between traditional and online CMC corpora can be attributed to differences in the corpus nature, sample selection, and content quality.

4. What are the possibilities and complexities involved in constructing a corpus from the web?

e) Collecting linguistic data from the web provides access to a wealth of specialized, current, and detailed material on various topics.

Corpora built from online resources differ from traditional corpora in two major ways: first, they do not adhere to the prescriptive requirements of corpus construction, and second, the nature of the language in online corpora is different. In this section, I will explore these differences and attempt to answer the research questions and hypotheses.

When it comes to corpus construction from the web, there are some challenges that need to be addressed. For instance, online corpora do not necessarily adhere to the tenets established by traditional corpus linguistics, which emphasize the need for a balanced and representative corpus. According to scholars such as Biber (1993) and McEnery & Wilson (1996), a corpus should include all text types in similar amounts to make generalizations about the language.

However, to ensure that a corpus truly represents a language, other issues need to be addressed. For example, Kilgarriff and Grefenstette (2003) argue that we must answer several theoretical questions with practical implications.

:

- Is a language event an event of speaking or writing, or one of reading or hearing? (productions and reception).
- Do speech events and written events have the same status? (speech and text).
- Does passing (and possibly subliminal reading) a roadside advertisement constitutes a reading event? (Background noise)
- In text domains, is republishing news a new writing event?

Answering these theoretical questions would be a difficult and potentially futile task. Such a definition presents significant obstacles to constructing specialized corpora, which could hinder researchers from addressing specific linguistic analyses. However, in many other disciplines such as business, medicine, and engineering, the use of corpora to expand research interests has a long history, and they were not all too concerned with fitting the requisites established for a corpus. For my research study, which required a corpus to carry out automatic classification of misogynistic language, it would not have been possible if I had constrained myself to fulfilling such requirements. Thus, it is clear that defining what corpora are has been given more attention than considering if such a corpus could be useful for a particular task. Therefore, for my research, the corpus did not need to represent all Spanish language as a whole, it just needed to be a sample of misogynistic language. This assertion brings me to the argument that a corpus also needs to be balanced.

A balanced corpus aims to accurately represent a specific language in a given population. For instance, if one wanted to examine the representation of gender in language, as was the case in this research, one would need to look beyond social media and include personal interactions, letters, newspapers, and other sources. However, gathering such a corpus would be expensive and time-consuming, potentially discouraging researchers from such an undertaking. Specialized corpora, such as those focused on business, require data from a variety of sources such as reports, meetings, and newsfeeds, making compilation difficult.

In my research, I compiled a corpus to perform automatic text classification, focusing on language used in the YouTube social media platform. This approach provided me with the freedom to research how gender is discussed in social media and use this information for classification tasks. However, unlike published texts that adhere to standard language conventions, web texts (such as YouTube comments) are often produced with little regard for correctness. As a result, the YouTube corpus used in my research contained spelling mistakes, grammatical errors, lexical issues, and punctuation errors, which posed challenges for certain analyses. Nevertheless, the corpus was still suitable for word frequency and concordance analysis, and could be expanded for other linguistic studies.

Regarding the requirement that a corpus be a standard reference, this poses a significant difference between a corpus compiled from the web and one from traditional sources. While published texts are edited and conform to language standards, web texts, including YouTube comments, are produced without such concerns. Although this limitation prevented me from conducting collocational analysis in my research, I was still able to use the corpus for ATC tasks. Therefore, it is important to recognize that a corpus's usefulness should be assessed based on the task it was compiled for, rather than rigidly adhering to established standards.

When it comes to corpora, the BNC and the Brown corpus, among others, are often considered as the reference or de facto standard for balance and representativeness in compiling specialized corpora. However, there is a clear dissonance between traditional corpora and specialized corpora, which often lack these aforementioned characteristics. This presents an opportunity to redefine what a corpus is and what characteristics it should have. As the corpus approach to language develops, there will likely be more scrutiny placed on how a corpus should be operationalized.

In the following paragraphs, I will discuss the second major difference between traditional corpora and corpora built from the web, particularly from social network outlets. This difference concerns the nature of the language represented in both types of corpora.

To elaborate on this difference, it is important to revisit the concept that a corpus should be a standard reference. The standard variety of a language, which is a somewhat elusive term, is generally a written form of the language that has undergone some degree of regularization

or codification. It is widely recognized as the prestigious variety of the language and serves as the high variety. This type of language is commonly used in academic, government, and educational contexts, as noted by Holmes (2013). Corpora such as the BNC, Brown corpus, NOW corpus, the CREA corpus, and the CORPES XXI aim to represent this standard language and serve as reference corpora.

The CREA and CORPES XXI corpora contain a mixture of written and spoken text, including edited books, novels, plays, academic essays, newspapers, magazines, transcriptions from radio and television, speeches, and conversations. The language present in these sources is usually a polished version of the language, filtered to conform not only to written conventions but also to expected social norms for each type of source. This ensures that the language is appropriate for the intended readership.

In contrast, corpora compiled from the web, particularly from social media outlets, often include a more unfiltered version of language. Web texts, such as YouTube comments, are typically produced with little attention to grammatical correctness or adherence to standard language conventions. The YouTube corpus used in my research, for example, contained numerous spelling mistakes, grammatical errors, lexical issues, and punctuation errors. While this made it unsuitable for collocational analysis, it was still useful for automatic text classification tasks, word frequency analyses, and concordance analyses, and could be expanded for other types of linguistic analysis. As the corpus approach to language continues to evolve, more attention is likely to be placed on how corpora should be constructed to accommodate these different types of language.

In contrast to traditional corpora like the ones mentioned above, web corpora, and particularly those built from social networks such as YouTube, do not undergo any filtering process. The language used on these platforms is not always reflective of the standard variety of a language, as individuals from various backgrounds participate in these conversations and bring their own linguistic idiosyncrasies that may not align with what is considered standard. The parameters set for compiling a corpus according to a standard reference may exclude certain vernacular languages from analysis. It is important to note that vernacular language refers to the most colloquial variety of a person's linguistic repertoire, which is typically used

at home and with close friends. The YouTube corpus contains numerous examples of these varieties of language. Please, see the following instances.

pinches huevones q no se quieren levantar lo atacan ahora los patos le tiran a las escopetas jajaja

No mames wey. en qué mundo vives o en qué país?

It is worth noting that the underlined words in these examples are unlikely to be found in a traditional corpus. Such words are often considered non-standard or inappropriate in the context of a standardized language. This emphasis on a standardized language in traditional corpora overlooks the fact that there are many local varieties of language that may be prevalent on the internet and social media platforms. These local varieties may be the primary linguistic repertoire of people who use the web, and they are often overlooked by traditional corpus builders.

As the internet has become a platform for a diverse range of voices, it has given a voice to those who were previously unheard. This has resulted in the emergence of new forms of language, which may not conform to a standardized language. As such, it is important to recognize the value of these forms of language and include them in linguistic analyses, rather than privileging a standardized form of language.

The comparison between traditional corpora and those built from the web should not involve prioritizing one linguistic variety over another. Instead, these corpora should be seen as complementary sources. It is important to allow for vernacular varieties to be included in corpora, as this would permit scholars to analyze different regional varieties that may not be easily accessible otherwise. Online communities provide a unique opportunity to study sociolects and linguistic variation. In fact, new varieties are being created and developed in these communities, and the online sociolect continues to evolve as more people participate in online conversations.

In the previous paragraphs, I discussed the differences between traditional corpora and those built from the web, with a focus on the YouTube corpus. However, it is important to note that replicability is a key consideration in corpus linguistics. While online corpora may offer

advantages such as broader coverage and dynamic content, the availability of web data can be unpredictable, potentially compromising replicability.

One possible solution is to transfer online corpora offline, which allows for greater control, accessibility, and level of analysis (Hundt et al., 2007). Offline corpora can be curated to include only relevant texts, which enables researchers to have a deeper understanding of the contents. Accessing offline corpora also allows researchers to use software with which they are more familiar. Additionally, offline corpora can be annotated, which enhances the range of analyses that can be conducted.

For my own research study, I used the YouTube corpus to carry out ATC tasks and keyword analyses. Having greater control over the corpus allowed me to better understand the possibilities and limitations of the corpus, which informed my analytical approach.

The second research question of this section aims to explore the possibilities and intricacies of compiling a corpus from the web. Over the past 15 years, the use of data derived from internet use and computer-mediated communication (CMC) has skyrocketed, leading to a rapid increase in the number of online communities where internet users discuss a wide range of topics (Horrigan, 2001; Nie, Hillygus, & Erbring, 2002). CMC environments such as Twitter, Facebook, YouTube, email, instant messaging, chat rooms, discussion forums, blogs, online classes, and video conferencing offer diverse opportunities for linguistic analysis that extend beyond corpus building. It is worth emphasizing that we are not only gaining access to linguistic data, but also exploring new communication methods.

One advantage of digital communication technologies is their ability to connect people across great distances or within hard-to-reach communities. Additionally, they offer researchers access to topics that may be difficult to study through traditional methods. For example, certain conversations and discussions only occur within specific computer-mediated communication (CMC) sources. The compilation of the YouTube corpus, which explores issues such as femicide and sexual harassment affecting women, as well as drug trafficking and gay rights for men and members of the LGBT community, provides a clear illustration of this phenomenon. These conversations are made possible by the existence of CMC outlets, which fill a gap in public discourse on these topics. In addition, these CMC outlets provide a

platform not only for individuals and communities to voice their opinions and grievances, but also for people to connect and form social relationships. While there are discussion forums led by scholars, journalists, and experts on these topics, everyday people are often excluded from these conversations. However, through the YouTube corpus, we can see comments and perspectives from women and members of the LGBT community, who are often marginalized and may be hesitant to express their views in face-to-face interactions. By accessing information shared in online communities, researchers can investigate a range of topics using traditional and progressive approaches to language analysis. For example, Page, Unger, Zappavigna, and Barton (2014) outline several areas of inquiry for web-based language research:

- Linguistic practices: what people do with language, the regular behaviors that develop within particular communities, and how language is used to perform particular identities (for instance, linguists might analyze how a forum community uses narratives/stories to enhance group cohesion, or how Facebook friends code-switch between different languages to signal their linguistic identities).
- Texts/utterances: collections of words, clauses, and sentences arranged deliberately in a structure with a clear communicative function. When a certain type of text becomes easily recognizable, this is often referred to as a genre, e.g., a comment thread on a newspaper site. This level of language is also sometimes referred to as discourse.
- Clauses and sentences: strings of words arranged in a structure, often described as syntax or grammar.
- Lexemes or words: units of meaning consisting of one or more morphemes, like eggs.
- Morphemes: the smallest units of meaning, e.g., egg, which calls up a certain concept in our minds, or ‘-s’ to indicate plurality.
- Phonemes: individual sounds/signs that make up spoken or signed words; and graphemes, e.g., letters or characters in writing.

In addition to analyzing the texts themselves, researchers can also investigate the contexts in which these texts are produced, extending their inquiry beyond language alone.

- Participants: the people who take part in the interaction and their relationship to others in the group.
- Imagined context: the projected contexts created cognitively by participants based on their world knowledge and the cues provided in CMC. This can include the projected audience that the participant addresses and the community they assume they are part of.

- Extra-situational context: the offline social practices in which the participants are involved, which might be shaped by cultural values relating to demographic factors such as age, gender, ethnic or national identity, and specific values relating to their involvement in particular communities (such as friendship groups, educational cohorts, hobby or interest groups, members of the same workplace, fan communities, and so on).
- Behavioral context: the physical situation in which the participants interact via social media (e.g., where and when the social media interaction takes place, what devices are used, and so on).
- Textual context: sometimes referred to as co-text, the textual context can include the surrounding interactions (the text published in preceding and subsequent posts or comments); semi-automated information such as timestamps, location-based information like ‘check ins’; screen layout, and resources.
- Generic context: the social media site in which the communication takes place including the site’s stated purpose, rules, and norms for conduct. These are often stated explicitly (such as Wikipedia’s core content policies, or can emerge from the participants’ activities which recognize certain forms of interaction as appropriate or not).

(p.31-33)

In addition to the proliferation of web-based data, the emergence of new genres and text types has significantly broadened the scope of research within Linguistics. For example, mobile phone text messages represent a unique area of study, where researchers have examined how the constrained format and cumbersome input technology have encouraged the use of short sentences, abbreviations, and graphic symbols (commonly known as emoticons or smileys) (Lindquist, 2009, p. 224).

As highlighted in Chapter 2, CMC outlets have been extensively utilized in various areas of Linguistics, including sociolinguistics, discourse analysis, and language learning, among other disciplines (Androutsopoulos, 2006; Herring, 2013; Kern et al., 2016). In fact, within some of these areas, such as discourse analysis and language learning, subfields such as computer-mediated discourse analysis (CMDA) and computer-assisted language learning (CALL) have emerged as independent disciplines in which CMC outlets play a central role. Moreover, exploring not only written but also spoken language further expands the possibilities for research.

When dealing with web data, accessing the information can pose a challenge. The sheer volume of linguistic data available can be daunting, especially for those who are unfamiliar

with information retrieval procedures. However, there are ways to construct a corpus through manual queries and downloads. While this process may not be as advanced as the techniques advocated in corpus linguistics and natural language processing, the quality of the results depends on the methodological rigor employed during the corpus construction, data analysis, and research planning.

With careful planning, a well-constructed corpus can serve multiple linguistic analyses. It is important to maintain control over the inclusion of data in the corpus to determine the scope of analysis that is possible. Another option for building corpora is to use toolkits such as BootCat. This toolkit, which operates through a web interface, employs an iterative procedure to create specialized corpora and terms from the web. It requires only a small list of "seeds" as input, which are terms that are typical of the domain of interest (Baroni & Bernardini, 2004). BootCat guides the user through the entire process, from the introduction of the seeds to the final moment when the corpus is created, retrieved, and formatted as a text. While BootCat has proven to be most effective when searching the web in English, acceptable results have also been obtained in other languages. These toolkits offer an alternative to manually querying the web and enable the creation of larger corpora.

Crawling the web is another method for building corpora. Specialized software allows users to crawl websites and pages, fetching them automatically and processing them to index their content. This enables the content to be searched using query tools like WordSmith, MonoConc, and AntConc. Crawling also allows users to remove duplicates and other non-linguistic material, and some crawls even allow for tokenization, lemmatization, and part-of-speech tagging of the corpus (Baroni & Kilgarriff, 2006). Crawling has become a popular way to compile corpora and can result in billions of words. It can also be a valuable tool for building corpora in languages with limited resources corpus (Cho et al., 1998; Clarke et al., 2005; Thelwall, 2005; Liu & Curran, 2006). However, this process requires technical expertise with the command line, which can be a challenge for some researchers. Despite this setback, the benefits of overcoming this challenge can be significant, enabling researchers to expand their studies and uncover new insights.

Working with data from the web, whether it be for constructing a corpus or not, entails various intricacies that can compromise the reliability and validity of the research process. Issues of replicability, representativeness, and balancedness, as previously mentioned, are among the challenges. Moreover, using web data poses methodological concerns, such as the lack of control over social network participants who may use false identities, making it difficult to generalize results to other populations (Andrews, Nonnecke, & Preece, 2003). Additionally, information can be taken down from the web, hindering replication of research studies. Despite these challenges, the opportunities afforded by the web and computer-mediated communication (CMC) outlets for corpus and linguistic analysis are undeniable. The linguistic data available on the web has expanded the possibilities for various linguistic analyses, limited only by the researchers' imagination.

While working with web data can present challenges, including the lack of technical knowledge required to access information and potential dissonances with research conventions, the benefits are worth considering. The use of web data provides access to specialized, detailed, and up-to-date information that can help expand the scope of research interests.

7 Conclusion

In my thesis, I investigated the collocational behavior of the lemmas MAN and WOMAN in an online corpus, as well as the effectiveness of using keywords as features in automatic text classification tasks. Through two collocational analyses and various classification tasks on texts and strings of vectors, I found that the use of keywords significantly improved the accuracy of the classification tasks with different algorithms. To perform these classification tasks, I compiled a corpus from the YouTube social network and extracted the keywords, which was a major accomplishment of this study. Additionally, a key topic throughout the automatic classification tasks was the linguistic representation of women and men in corpora. Based on these findings, I have arrived at several results and conclusions.

- Online social networks provide an opportunity to investigate the linguistic representation of women and men.
- These platforms allow ordinary people and marginalized communities to participate in discussions and negotiate meanings, identities, and discourses.
- Keywords, as defined by Corpus Linguistics, can serve as attributes in automatic text classification.
- In our investigation, using keywords as topic features with a frequency weight scheme achieved 98% accuracy in automatic text classification.
- Classifying almost 7,500 comments resulted in 92% accuracy when using keywords, but the accuracy decreased by an average of 17% when keywords were not used.
- While feature selection methods in machine learning are well established, our study found that a keyword analysis was effective in identifying accurate topic features.
- Corpus linguistics tools can aid in corpus construction and analysis for automatic text classification.
- The World Wide Web and social networks offer valuable resources for corpus compilation.

- However, corpora from the web may not conform to traditional conventions, which can affect representativeness and balance.
- Replicability can also be a challenge with online corpora, as data can be removed at any time.
- Despite these challenges, online corpora provide access to specialized, detailed, unfiltered, and up-to-date data.

Implications

The primary motivation for this research was to establish a connection between corpus linguistics and machine learning. The aim was to investigate whether corpus linguistics could contribute to the development of techniques in machine learning. It is hoped that the findings of this investigation will stimulate interdisciplinary dialogue not only among subfields within linguistics but also with other major fields.

Limitations of the study

Throughout this investigation, I encountered several limitations that affected the scope and generalizability of my findings. Firstly, my limited experience in the field of machine learning hindered my ability to compare the performance of the keyword analysis with other feature selection methods. This limitation could be addressed in future research by involving collaborators with expertise in machine learning to ensure a more comprehensive analysis.

Secondly, my focus on topic features (keywords) alone may have limited the extent to which I could draw meaningful conclusions about the performance of other features that could have been extracted from the corpus. Future studies could explore the comparative effectiveness of other feature selection methods and assess their suitability for different research contexts.

Finally, the size of the corpus and limitations of the computing equipment available to me limited the scale of my analysis. Although the classification process involved almost 7,500 comments, the corpus contained nearly 30,000 comments, and I was unable to test the

performance of the keywords with the full corpus. This limitation highlights the need for more robust computing infrastructure and software to support the processing and analysis of large-scale corpora in future studies.

Future research

There are several important issues that require further investigation in the future. Firstly, more research needs to be conducted on the use of keywords in feature selection compared to other similar methods in machine learning. Secondly, the YouTube corpus used in this study requires further processing and expansion for other linguistic analyses. Finally, there is a pressing need to revisit the issue of gender representation, but this requires a cleaned corpus for accurate and reliable analysis.

References

- Andrews, D., Nonnecke, B., & Preece, J. (2003). Electronic survey methodology: A case study in reaching hard-to-involve Internet users. *International Journal of Human-Computer Interaction*, 16(2), 185-210. https://doi.org/10.1207/S15327590IJHC1602_04
- Androutsopoulos, J. (2006). Introduction: Sociolinguistics and computer-mediated communication. *Journal of Sociolinguistics*, 10(4), 419-438. <https://doi.org/10.1111/j.1467-9841.2006.00286.x>
- Anthony, L. (2005). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. In T. K. Nakamura, K. Nagao, & T. Tokunaga (Eds.), *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning* (pp. 7-13). Waseda University.
- Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, 30(2), 141-161. <https://doi.org/10.17250/KHISLI.30.2.201308.001>
- Anzovino, M., Fersini, E., & Rosso, P. (2018). Automatic identification and classification of misogynistic language on twitter. In M. Silberstein, F. Atigui, E. Kornysheva, E. Métais, & F. Meziane (Eds.), *Natural language processing and information systems* (pp. 57-64). Springer. https://doi.org/10.1007/978-3-319-91947-8_6
- Baker, P. (2004). Querying Keywords: Questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics*, 32(4), 346-359. <https://doi.org/10.1177/0075424204269894>
- Baker, P. (2010). *Sociolinguistics and corpus linguistics*. Edinburgh University Press.
- Baker, P. (2012). Corpora and gender studies. En K. Hyland, M. H. Chau, & M. Handford (Eds.), *Corpus applications in applied linguistics* (pp. 100-116). Bloomsbury Publishing.
- Baker, P. (2013). Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English. *Gender and Language*, 1(1), 1-28.

<https://doi.org/10.1558/genl.v1.i1.17188>

- Barlow, M. (1999). MonoConc 1.5 and ParaConc. *International Journal of Corpus Linguistics*, 4(1), 173-184. <https://doi.org/10.1075/ijcl.4.1.09bar>
- Baroni, M., & Bernardini, S. (2004, May). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, 1313-1316.
- Baroni, M., & Kilgarriff, A. (2006). Large linguistically-processed web corpora for multiple languages. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics. Posters & Demonstrations*, 87-90.
- Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., & Sanguinetti, M. (2019). SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in twitter. *Proceedings of the 13th International Workshop on Semantic Evaluation*, 54-63. <https://doi.org/10.18653/v1/S19-2007>
- Baxter, J. (2003). *Positioning gender in discourse: A feminist methodology*. Springer.
- Bergh, G., & Zanchetta, E. (2008). Web Linguistics. En A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 309-327). W. de Gruyter.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.
- Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing*, 8(4), 243-257.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman.
- Biber, D., & Reppen, R. (Eds.). (2015). *The Cambridge Handbook of English Corpus Linguistics*. Cambridge University Press.

- Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.
- Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139-173.
<https://doi.org/10.1075/ijcl.20.2.01bre>
- Burgess, J., & Green, J. (2013). *YouTube: Online video and participatory culture*. John Wiley & Sons.
- Butler, J. (1990). *Gender trouble: Feminism and the subversion of identity*. Routledge.
- Caldas-Coulthard, C. R., & Moon, R. (2010). Curvy, hunky, kinky: Using corpora as tools for critical analysis. *Discourse & Society*, 21(2), 99-133. <https://doi.org/10.1177/0957926509353843>
- Canós, J. S. (2018). Misogyny identification through SVM at IberEval 2018. IberEval@SEPLN. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages*, 229-233.
- Chandler, B. (1989). *Longman mini concordancer: Text study software for teachers and students: IBM PC and compatibles*. Longman.
- Cho, J., Garcia-Molina, H., & Page, L. (1998). Efficient crawling through URL ordering. In *WWW7 Proceedings of the seventh international conference on World Wide Web 7 archive*, 30,161–172.
- Clarke, C., Craswell, N., & Soboroff, I. (2005). The TREC terabyte retrieval track. *SIGIR Forum*, 39(1),25.
- Clear, J. (1993). From firth principles: Computational tools for the study of collocations. En M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair* (pp. 271-292.). John Benjamins Publishing.
- Coates, J. (2015). *Women, men, and language: A sociolinguistic account of gender differences in language*. Routledge.
- Crawford, M. (1995). *Talking difference: On gender and language*. SAGE.

- Culpeper, J. (2009). Keyness: Words, parts-of-speech, and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*. *International Journal of Corpus Linguistics*, 14(1), 29-59.
<https://doi.org/10.1075/ijcl.14.1.03cul>
- Dalal, M. K., & Zaveri, M. A. (2011). Automatic text classification: A technical review. *International Journal of Computer Applications*, 28(2), 37-40.
- Davis, M. (2020). <https://www.english-corpora.org/>. Accessed on October 10, 2020.
- Deng, X., Li, Y., Weng, J., & Zhang, J. (2019). Feature selection for text classification: A review. *Multimedia Tools and Applications*, 78(3), 3797-3816. <https://doi.org/10.1007/s11042-018-6639-9>
- Dixon, R. M. W., & Aikhenvald, A. I. (2004). *Adjective classes: A cross-linguistic typology*. Oxford University Press.
- Evert, S. (2008). Corpora and collocations. En A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 1212-1248). W. de Gruyter.
- Fernández Anta, A., Morere, P., Chiroque, L. F., & Santos, A. (2012, September). Techniques for sentiment analysis and topic detection of Spanish tweets: preliminary report. In *Spanish Society for Natural Language Processing Conference*.
- Fersini, E., Rosso, P., & Anzovino, M. (2018). Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages*, 214-228.
- Fishman, P. M. (1978). The work women do. *Social problem*, 25(4), 397-406.
- Fletcher, W. H. (2004). Making the web more useful as a source for linguistic corpora. *Applied Corpus Linguistics*, 191-205. https://doi.org/10.1163/9789004333772_011
- Flowerdew, J., & Richardson, J. E. (Eds.). (2017). *The routledge handbook of critical discourse analysis*.

- Frank, E., Hall, M. A., & Witten, I. H. (2016). *The WEKA Workbench. Online Appendix for «Data Mining: Practical Machine Learning Tools and Techniques (4th ed.)*. Morgan Kaufmann.
- Gabrielatos, C. (2018). Keyness analysis: Nature, metrics, techniques. En C. Taylor & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 225-258). Routledge.
- García-Díaz, J. A., Cánovas-García, M., Colomo-Palacios, R., & Valencia-García, R. (2021). Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114, 506-518. <https://doi.org/10.1016/j.future.2020.08.032>
- García-Miguel, J., & Albertuz, F. (2005). Verbs, semantic classes and semantic roles in the ADESSE project. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, 50-55. University of Vigo.
- Gesuato, S. (2003). The company women and men keep: What collocations can reveal about culture. *Proceedings of the Corpus Linguistics 2003 Conference*, 253-262.
- Gries, S. T. (2013). 50-something years of work on collocations: What is or should be next *International Journal of Corpus Linguistics*, 18(1), 137-166. <https://doi.org/10.1075/ijcl.18.1.09gri>
- Halliday, M. A. K. (1993). Quantitative studies and probabilities in grammar. En M. Hoey (Ed.), *Data, description, discourse: Papers on the english language in honour of John McH Sinclair on his sixtieth birthday* (pp. 1-25). HarperCollins.
- Hardaker, C. (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research*, 6(2), 215-242. <https://doi.org/10.1515/jplr.2010.011>
- Hardie, A. (2020). *CQPweb [Computer Software]*. Available from <http://cwb.sourceforge.net/index.php>
- Herring, S. C. (2013). Discourse in web 2.0: Familiar, reconfigured, and emergent. En Deborah Tannen &

- A. M. Trester (Eds.), *Discourse 2.0: Language and new media* (pp. 1-26). Georgetown University Press.
- Hillard, D., Purpura, S., & Wilkerson, J. (2008). Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, 4(4), 31-46.
<https://doi.org/10.1080/19331680801975367>
- Holmes, J. (2013). *An introduction to sociolinguistics*. Routledge.
- Horrigan, J. B. (2001). *Online communities: Networks that nurture long-distance relationships and local ties*. Pew Internet and American Life Project.
- Hundt, M., Nesselhauf, N., & Biewer, C. (Eds.). (2007). *Corpus linguistics and the web*. Rodopi.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.
- Hyland, K., Chau, M. H., & Handford, M. (2012). *Corpus applications in applied linguistics*. Bloomsbury Publishing.
- Jensen, S. Q. (2011). Othering, identity formation, and agency. *Qualitative Studies*, 2(2), 63-78.
- Jones, R., & Ghani, R. (2000). Automatically Building a Corpus for a Minority Language from the Web. In *Proceedings of the Student Workshop of the 38th Annual Meeting of the Association for Computational Linguistics*, 38, 29-36.
- Kaye, G. (1990). A corpus builder and real-time concordance browser for an IBM PC. En J. Aarts & W. Meijs (Eds.), *Theory and practice in corpus linguistics* (pp. 137-162). Rodopi.
- Kendall, S., & Tannen, D. (2015). Discourse and gender. In D. Tannen, H. E. Hamilton, & D. Schiffrin (Eds.), *The handbook of discourse analysis* (2nd ed., pp. 639-660). Wiley Blackwell.
- Kennedy, G. (2014). *An Introduction to corpus linguistics*. Routledge.

- Kern, R., Ware, P., & Warschauer, M. (2016). Computer mediated communication and language learning. En G. Hall (Ed.), *The routledge handbook of english language teaching* (pp. 542-555). Routledge.
- Liu, V., & Curran, J. (2006). Web text corpus for natural language processing. *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 233–240.
- Kilgarriff, A. (2020). *Sketch Engine [Computer Software]*. Accessed from <https://www.sketchengine.eu/>
- Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3), 333-347. <https://doi.org/10.1162/089120103322711569>
- Korde, V., & Mahender, C. N. (2012). Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications*, 3(2), 85-99. <https://doi.org/10.5121/ijaia.2012.3208>
- Lakoff, R. T. (1975). *Language and woman(s) Place*. Harper & Row. Publishers.
- Lazar, M. M. (2014). Feminist critical discourse analysis. En S. Ehrlich, M. Meyerhoff, & J. Holmes (Eds.), *The handbook of language, gender, and sexuality* (2nd ed., pp. 180-200). Wiley Blackwell.
- Leech, G., & Fallon, R. (1992). Computer corpora—what do they tell us about culture. *ICAME journal*, 16, 29-50.
- Leech, G. (2007). New resources, or just better old ones? En M. Hundt, N. Nesselhauf and C. Biewer (eds.) *Corpus linguistics and the web* (pp. 134–9). Rodopi.
- Lew, R. (2012). The Web as corpus versus traditional corpora: Their relative utility for linguists and language learners. En P. Baker (Ed.), *Contemporary corpus linguistics* (pp. 289-301). Continuum.
- Lindquist, H. (2009). *Corpus linguistics and the description of english*. Edinburgh University Press.
- Litosseliti, L. (2014). *Gender and language theory and practice*. Routledge.
- Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4), 491-502.

- Maltz, D. N., & Borker, R. A. (2012). A cultural approach to male-female miscommunication. En L. Monaghan, J. E. Goodman, & J. M. Robinson (Eds.), *A Cultural approach to interpersonal communication: Essential readings* (pp. 168-185). John Wiley & Sons.
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.
- Marsland, S. (2014). *Machine learning: An algorithmic perspective* (2nd ed.). Chapman and Hall/CRC.
- Mautner, G. (2007). Mining large corpora for social information: The case of elderly. *Language in Society*, 36(1), 51-72.
- McEnery, A., & Baker, P. (2015). *Corpora and discourse studies: Integrating discourse and corpora*. Springer.
- McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- McEnery, T., & Wilson, A. (1996). *Corpus linguistics*. Edinburgh University Press
- McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction*. Edinburgh University Press.
- Mihalcea, R., & Moldovan, D. I. (1999). A Method for word sense disambiguation of unrestricted text. *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, 152–158. <https://doi.org/10.3115/1034678.1034709>
- Miller, C. R., & Kelly, A. R. (2016). *Emerging genres in new media environments*. Springer.
- Moon, R. (2014). From gorgeous to grumpy: Adjectives, age, and gender. *Gender & Language*, 8(1). <https://journals.equinoxpub.com/GL/article/view/14715>
- Nie, N., Hillygus, S. & Erbring, L. (2002). Internet use, interpersonal relations and sociability: Findings from a detailed time diary study. In B. Wellman (Ed.), *The Internet in everyday life* (pp. 215–243). Blackwell Publishers.

- Page, R., Unger, J. W., Zappavigna, M., & Barton, D. (2014). *Researching language and social media: A student guide*. Routledge.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundation and trends in information Retrieval*, 8, 1-135. DOI: f10.1561/1500000001
- Pearce, M. (2008). Investigating the collocational behaviour of man and woman in the BNC using Sketch Engine. *Corpora*, 3(1), 1-29. <https://doi.org/10.3366/E174950320800004X>
- Peters, I., & Peters, W. (2000). The treatment of adjectives in SIMPLE: Theoretical observations. *In LREC. European Language Resources Association*.
- Phillips, M. (1989). *Lexical structure of text*. University of Birmingham.
- Pihlaja, S. (2014). *Antagonism on youtube: Metaphor in online discourse*. Bloomsbury Publishing.
- Plaza-del-Arco, F. M., Molina-González, M. D., Martín, M., & Ureña-López, L. A. (2019). SINAI at SemEval-2019 Task 5: Ensemble learning to detect hate speech against immigrants and women in english and spanish tweets. *In Proceedings of the 13th International Workshop on Semantic Evaluation*, 476–479. <https://doi.org/10.18653/v1/S19-2084>
- Pojanapunya, P., & Todd, R. W. (2018). Log-likelihood and odds ratio: Keynes statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14(1), 133-167. <https://doi.org/10.1515/cllt-2015-0030>
- Poland, B. (2016). *Haters: Harassment, abuse, and violence online*. U of Nebraska Press.
- Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 1-47.
- Pustejovsky, J., & Stubbs, A. (2012). *Natural language annotation for machine learning: A guide to corpus-building for applications*. O'Reilly Media.

- Raun, T. (2016). *Out Online: Trans self-representation and community building on YouTube*. Routledge.
- Rayson, P. (2013). Corpus analysis of key words. En C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1260-1267). John Wiley & Sons.
- Rayson, P. (2015). Computational tools and methods for corpus compilation and analysis. En D. Biber & R. Reppen (Eds.), *The Cambridge handbook of english corpus linguistics* (pp. 32-49). Cambridge University Press.
- Rayson, P., Leech, G. N., & Hodges, M. (1997). Social differentiation in the use of english vocabulary: Some analyses of the conversational component of the british national corpus. *International Journal of Corpus Linguistics*, 2(1), 133-152. <https://doi.org/10.1075/ijcl.2.1.07ray>
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv Preprint ArXiv:1908.10084*. <https://arxiv.org/abs/1908.10084v1>
- Renouf, A. (2007). Corpus development 25 years on: From super-corpus to cyber-corpus. En R. Facchinetti (Ed.), *Corpus linguistics 25 Years on* (pp. 27-49). Rodopi.
- Romaine, S. (2000). *Language in society: An introduction to sociolinguistics*. OUP Oxford.
- Romero, D. (2010). *El adjetivo en el nuevo testamento: Clasificación semántica [Published doctoral dissertation]*. Universidad de Córdoba. <http://helvia.uco.es/xmlui/handle/10396/3535>
- Schmidt, A., & Wiegand, M. (2017). A Survey on hate speech detection using natural language processing. *In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10. <https://doi.org/10.18653/v1/W17-1101>
- Scott, M. (1996). *Wordsmith tools*. Oxford University Press.
- Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. John Benjamins Publishing.

- Sebastiani, F. (2005). Text categorization. Encyclopedia of database technologies and applications. *IGI Global*, 683-687. https://doi.org/10.1007/978-0-387-39940-9_414
- Sigley, R., & Holmes, J. (2002). Looking at girls in Corpora of English. *Journal of English Linguistics*, 30(2), 138-157. <https://doi.org/10.1177/007242030002004>
- Sinclair, J. (1991). *Corpus, concordance, Collocation*. Oxford University Press.
- Spender, D. (1980). *Man made language*. Routledge & Kegan Paul.
- Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2), 209-243. <https://doi.org/10.1075/ijcl.8.2.03ste>
- Stubbs, M. (2010). Three concepts of keywords. En M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 21-42). John Benjamins Publishing.
- Sunderland, J. (2004). *Gendered discourses*. Springer.
- Swift, K., & Miller, C. (2001). *The Handbook of nonsexist writing: For writers, editors and speakers* (2nd edition). iUniverse. (Original work published in 1981).
- Tannen, D. (1990). *You just don't understand*. Ballentine Books.
- Tannen, D. (1994). *Gender and discourse*. Oxford University Press.
- Teubert, W. (2005). *My version of corpus linguistics*. John Benjamins Publishing Company.
- Thelwall, M. (2005). Creating and using web corpora. *International Journal of Corpus Linguistics*, 10(4), 517-541.
- Theobald, O. (2018). *Machine learning for absolute beginners: A plain english introduction*. Independently Published.

- Thomas, M. (2013). Otto Jespersen and “The Woman”, then and now. *Historiographia Linguistica*, 40(3), 377-408. <https://doi.org/10.1075/hl.40.3.03tho>
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins Publishing.
- Tsvetkov, Y., Schneider, N., Hovy, D., Bhatia, A., Faruqui, M., & Dyer, C. (2014). *Augmenting english adjective senses with supersenses*. University of Copenhagen.
- Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). *Practical natural language processing: A comprehensive guide to building real-world NLP systems*. O'Reilly Media.
- Van den Bosh, A. (2008). Machine learning. En A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 855-873). W. de Gruyter.
- Weatherall, A. (2005). *Gender, language and discourse*. Routledge.
- Weisser, M. (2016). *Practical corpus linguistics: An introduction to corpus-based language analysis*. John Wiley & Sons.
- Xiao, R. (2015). Collocations. En D. Biber & R. Reppen (Eds.), *The Cambridge handbook of corpus linguistics* (pp. 106-124). Cambridge University Press.
- Yang, J., Liu, Y., Zhu, X., Liu, Z., & Zhang, X. (2012). A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. *Information Processing & Management*, 48(4), 741-754.
- Zimmerman, D. H., & West, C. (1975). Sex roles, interruptions and silences in conversations. En B. Thorne & N. Heinle (Eds.), *Language and sex: Differences and dominance* (pp. 105-119).