Using keywords in the automatic classification of language of gender violence

Héctor Castro Mosqueda°, Antonio Rico-Sulayes*
°Escuela Normal Superior Oficial de Guanajuato, *Universidad de las Américas Puebla

This paper employs lexical analysis tools, quantitative processing methods, and natural language processing procedures to analyze language samples and identify lexical items that support automatic topic detection in natural language processing. This paper discusses how keyword extraction, a technique from corpus linguistics, can be employed in obtaining features that improve automatic classification; in particular, this research is concerned with extracting keywords from a corpus obtained from social networks. The corpus consists of 1,841,385 words and is subdivided into three sub-corpora that have been categorized according to the topic of the comments in each one of them. These three topics are violence against women, violence against the LGBT community, and violence in general. The corpus has been obtained by scraping comments from YouTube videos that address issues such as street harassment, femicide, feminist movements, drug trafficking, forced disappearances, equal marriage, among others. The topic detection tasks performed with the corpus extracted from the social media showed that the keywords rendered a 98% accuracy when classifying the collection of comments from 51 videos, as one of the three categories mentioned above, and 92% when classifying almost 7,500 comments individually. When keywords were removed from the classification task and all words were used to perform the classification task, accuracy dropped by an average of 17%. These results support the argument for keyword relevance in automatic topic detection.

Keywords: Corpus Linguistics; Automatic Text Classification; Sexist Language Detection.

1. Introduction

This work attempts to use tools and techniques from Corpus Linguistics (CL) to inform topic classification tasks in the Natural Language Processing (NLP) field. Topic classification (TC) is a Machine Learning (ML) task that is a branch of Artificial Intelligence in NLP. TC is a learning task that assigns a given document

CHIMERA. Romance Corpora and Linguistic Studies 10 (2023), 19-43. ISSN 2386-2629

to a class in a set of categories based on its content and extracted features. In TC tasks, features such as word- and character-sequences (so-called n-grams), POS, morphology, and pragmatic linguistic features, among others, are used to evaluate the classification accuracy. ML is about extracting knowledge from data, which allows a system to learn information from the data to apply it to several tasks. As one of these tasks, TC has different useful applications such as content management, spam filtering, opinion, and sentiment analysis, improving result ordering in search engines, ranking or grouping results, online reviews of products, among other applications (Sebastiani 2005; Dalal & Zaveri 2011; Vajjala *et al.* 2020). In addition, an important amount of research in TC has also focused on investigating hate speech to carry out classification tasks. In such experiments, data addressing topics such as ethnicity, immigration, gender identities, and misogynistic language have been employed to carry out classification tasks.

In TC tasks (also known as topic detection tasks), one of the main challenges is feature selection. To select efficient features, one needs to choose among thousands of words and other linguistic items, many of which may not only be non-informative but also render conflicting and/or poor results. In natural language processing, there is a wide variety of feature selection methods which are part of four major categories (filter model, wrapper model, embedded model, and hybrid model) (Deng *et al.* 2019; Yang *et al.* 2012; Liu & Yu 2005). Some of the most common feature selection methods to measure the goodness of features in TC tasks are the use of bag-of-words, TF-IDF, and information gain (IF). What we specifically propose in this study is the use of keyword extraction, the way it is performed in CL, as a feature selection method that supports topic classification tasks.

Although there has been extensive research addressing misogynistic language, to the best of our knowledge, there is no study that focuses on identifying lexical features that help distinguish verbal violence directed towards women versus men or members of the LGBT community in Spanish. Given the limited amount of research on gender violence in Spanish, as it has just been outlined, we hope that this study contributes to the literature bringing into play a variable, gender violence against the LGBT community, that has not been targeted so far.

2. Theoretical Framework

In this research study, tools and techniques from CL and NLP have been brought together. CL is a research area that focuses on a set of procedures or methods for studying the language; one of these procedures is the keywords that are used in this research study to classify gendered comments in automatic classification tasks.

2.1 Keywords

This investigation focuses on identifying keywords in corpora to be used in TC tasks. Keyword analysis is employed across applied linguistics and its uses vary from genre analysis to critically-oriented studies with different purposes, such as producing a general characterization of a genre or identifying text-specific ideological issues (Pojanapunya & Todd 2018). Baker (2004) emphasizes the popularity and adaptability of keyword analysis in CL, as it has been employed in a wide variety of studies. In this sense, we can note that keyword analysis can be employed to gain a descriptive account of different genres as well as to spot traces of discourse within language.

To identify the significant differences of keywords in different corpora, a keyness statistical measure is employed. In keyword analysis, "Keyness is a quality words may have in a given text or set of texts, suggesting that they are important, they reflect what the text is about, avoiding trivial and insignificant detail; what the text boils down to is its keyness" (Scott & Tribble 2006). To calculate the keyness, four elements are considered: 1) the frequency of a word in the target corpus, 2) its frequency in the comparative corpus, 3) the total number of words in the target corpus, and 4) the total number of words in the comparative corpus. The procedure to select keywords proceeds as follows. Firstly, a word frequency list is computed for each of the two texts or text collections that are to be compared. The word frequency records the different word forms (types) and how many times they occur (tokens); also, the total number of words in each text or collection is counted. Secondly, the two frequency lists are compared. For this, a keyness statistic measure is selected which compares the relative frequency of each word in the two sources; the larger the difference in relative frequencies, the larger the keyness value. Finally, the words are ordered according to their keyness value; namely, the higher the keyness, the more relevant the keyword is. Table 1 shows how each one of the frequencies in this type of analysis is obtained.

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

	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*

In Table 1, the value *a* and *b* refer to the frequency of the words in the two corpora, and the values *c* and *d* refer to the size of each corpus. Using these figures, an expected value is calculated for each word. The expected values are calculated using the following formula:

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

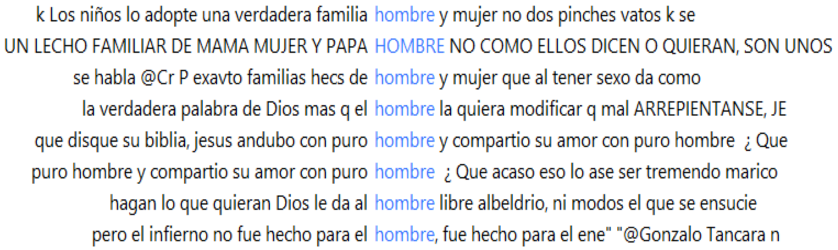
The expected values are the averages for each word adjusted for the corpus size. In the formula above, "N" refers to the total number of words, and "O" corresponds to the observed value. The expected values are represented in Table 1 as c and d. So, we calculate = *c* * (*a* + *b*) / (*c* + *d*) and = *d* * (*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 X (((*a* * *ln*(*a*/)) + (*b* * *ln* (*b* /E2))).

2.2 Tools

Corpora can comprise millions of words; therefore, to analyze such an amount of information we need computer software to search and retrieve information from each corpus. Concordancers are automated systems that compile and display instances of specific tokens in all the particular contexts where they occur in a corpus. See Figure 1 for an example of a concordance list.



```
        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 albeldrio, 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.

The capabilities of concordancers allow the users to analyze keywords, n-grams, collocations, as well as frequency lists. In this research study, the AntConc concordance tool, a third-generation concordancer, was employed to identify the keywords which would later be used in the TC tasks. Through AntConc, it was possible to compare a corpus against a reference corpus and statistically identify the keywords with higher statistical significance. This investigation also relied on the Waikato Environment for Knowledge Analysis (WEKA) software which allows its users to access tools from machine learning and data mining. WEKA users are provided with a series of learning algorithms for data preprocessing, manipulation, evaluation, and visualization. Some of the algorithms included in this workbench are algorithms for classification, regression, clustering, and attribute selection.

2.3 Violentómetro

The Violentómetro is a taxonomy designed by the National Polytechnic Institute in Mexico, through its Institutional Management Program with a Gender Perspective. This classification allows its users to identify how gender violence can be represented in any everyday situation. The Violentómetro has three different scales which describe events that range from behaviors involving verbal or psychological abuse to behaviors describing life-threatening events. As part of the TC tasks, in the first experiment, instead of using the features (keywords) obtained from the YouTube corpora, the verbs listed in the Violentómetro were used as features. This was done to establish a baseline. The verbs listed in the Violentómetro are employed to classify the level of danger of those events in which women may be involved. Since our YouTube corpora involved gender relationship dynamics, the Violentómetro classification was considered a good source of features to use for the first classification experiment.

2.4 Automatic classification

Automatic classification (AC) is a discipline that intersects ML and Information Retrieval (IR) and shares several characteristics with other tasks such as text mining. Some applications of AC technology are newswire filtering, opinion classification, patent classification, and webpage classifications, as these tasks rely on a topic approach for their classification procedures. This research study adopts a topic approach in classification tasks. Among the many applications that can benefit from AC, we can mention spam filtering, authorship attribution, author gender detection, and affective rating (Sebastiani 2005).

AC involves assigning a text document to a set of predefined classes automatically (Dalal & Zaveri 2011); such classification is usually done by relying on words or other textual features extracted from the same documents. Dalal and Zaveri (2011) describe a generic approach to automatic classification as follows:

I) Document pre-processing
II) Feature extraction/selection
III) Model selection
IV) Training and testing the classifier

In the first phase, *stop-words* (roughly equivalent to functional words, such as articles, prepositions, and auxiliary verbs) are eliminated because they are not considered useful for the classification task in machine learning. This happens because these words are not specific to topics and do not contribute to discriminating documents among their different classes (Kadhim, 2018; Birjali *et al.* 2021). Also, words are reduced to their root or base form in a so-called stemming process; for example, inflections for number and gender, in the case of nouns or adjectives in Spanish, or tenses and person for verbs are consolidated into a single word. In this phase, the size of documents is significantly reduced. If the data comes from web sources, this undergoes further pre-processing to eliminate web-derived content that may be too hard to process, such as URLs, hyperlinks, or hashtags. The second phase focuses on identifying important words in the documents. Such identification can rely on a statistical o semantic approach such as the TF-IDF (term frequency-inverse document frequency) model or the Latent Semantic Indexing (LSI) respectively. Other methods to process document contents are 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. Important words identified in this phase are commonly named features, attributes, or variables. In this same phase, each document is represented as a document vector; in other words, the representation of the documents is used to reduce the complexity of the documents and make them easier to handle, this is called *indexing preprocessing*. Table 2 shows a vector space model (VSM) representation that retains the information regarding the frequency of occurrences of the feature terms in each one of the texts (videos) for each one of the three classes (corpora). In this representation, each column is known as a vector and it stores the values for a given feature (a word or some other linguistic element) across all the documents in the collection. In contrast, each row stores the instances of all features for a given document. Each matrix can contain multiple vectors and instances and all of these are known as matrices. The representation in Table 2 adopts a multinomial model in which each vector retains the information regarding the frequency of each occurrence (feature) in every instance (document).

Table 2. Vector space model representation.

	Vector	**Matrices**	
Classes (texts)	**Feature 1 (keyword)**	**Feature 2**	**Feature 3**
V_Mujer (Text 1)	0		
V_Mujer (Text 2)	2		
V_General (Text 1)	1		
V_LGBT (Text 1)	16		

It is important to point out that the VSM has limitations and some of them are: high dimensionality of the representation, loss of correlation with adjacent words, and loss of semantic relationships that exist among the terms (features) in a document; to overcome such problems, term weighting methods are used to assign appropriate weights to the terms (Korde & Mahender 2012). The three most common weighting schemes are Boolean, Word Frequency, and TF-IDF.

In the Boolean scheme, a 1 is assigned to a_ik if this occurs in the document and a 0 if it does not.

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

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

In the word frequency scheme, the frequency of the word in each document is considered.

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

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

The TF-IDF scheme considers the frequency of the word in every document in each class. TD-IDF (Term Frequency X Inverse Document Frequency) evaluates how relevant a word is to a collection of documents; this metric is measured by multiplying the frequency of a word in a particular document times the inverse document frequency of the word across a whole set of documents.

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

In the formula above, f_{ik} stands for the frequency of the word *i* in document *k*, *N*, for the number of documents in each class, and n_i, for the number of documents in which the word *i* appears. For example, consider a document containing 100 words in which the word *powerful* appears 3 times. The term frequency for *powerful* is (3/100) = 0.03. Now, assume there are 10 million documents, and the word *powerful* appears in 1000 of these. Then, the inverse document frequency is calculated as log (10,000,000/1000) = 4. Thus, the TF-IDF weight is the product of these quantities: 0.03 * 4 = 0.12.

2.5 Language and the World Wide Web

Traditionally, scholars in CL have stressed that a corpus should be of finite size, in a machine-readable format, derived from a standard reference, and representative; however, in a great number of research studies, a corpus compiled from the web is used, which does not generally adhere to these conditions. Two approaches have been generally employed in CL when using the web as a source of data. Hundt, Nesselhauf, and Biewer (2007) describe these two approaches as follows:

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”).

Concerning the “Web for corpus building” approach, three advantages are identified: control, accessibility, and level of analysis. In regard to control, the researchers decide what kind of texts they include in their databases. This allows them to be more familiarized with the content of the corpus; regarding “accessibility”, once the corpus has been transferred offline, the researcher can use the standard software they are more accustomed to. Finally, concerning "level of analysis", offline corpora can be annotated and thus allow researchers to conduct a wider variety of analyses. Corpora can be used to describe the different linguistic behavior and forms used in socializing through digital discourse; the linguistic analysis can address lexical, syntactic, semantic, or discourse issues such as identity, politeness, rhetorical strategies, gender, power, and ideology among others. Based on the above, a web for corpus building approach seems to better serve corpus linguistics because the nature of linguistic analysis varies considerably; in other words, different kinds of data are needed for different linguistic analyses.

2.5.1 *Social Media and Language Research*

Covered under the umbrella term of Computer-Mediated Communication (CMC), social networks are sites designed to facilitate communication and strengthen social relationships. Some CMC environments are Twitter, Facebook, YouTube, electronic mail, instant messages, chats, discussion forums, blogs, and video conferencing among others. CMC and social networks in particular have opened an important avenue for linguists to exploit in their research. The corpora for this study were built from comments to YouTube videos that addressed topics about gender issues. The rationale for choosing videos was that users' comments about these kinds of videos would capture people's opinions regarding important gender issues.

YouTube has become a place for cultural participation (Burgess & Green 2013). Unlike Twitter and Facebook where social networking is based on personal profiling, on YouTube the video content is the main vehicle for communication which enables cultural participation by ordinary citizens who can express their identities, share their values, engage with others, negotiate meaning, and encounter cultural differences.

Very often antagonism and controversy arise in YouTube communities and this 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 develop new literacies, new cultural forms, and new social practices that are constructed, challenged, rejected, or adopted. Besides the antagonism that derives from online interactions, the anonymity that users benefit from has also attracted attention. In general, CMC sources offer a high degree of anonymity which may foster the effect of deindividuation that may lead the users to develop a sense of impunity, loss of self-awareness, and a likelihood of acting upon normally inhibited impulses (Hardaker 2010 in Pihlaja 2014). Furthermore, anonymity also presents users the opportunity to engage in conversations with people that will not otherwise occur due to the nature of the topics. In other words, YouTube presents a space for disenfranchised communities, a space that is usually not offered by those in power or by mainstream media.

3. Related work

As it was stated above, automatic classification can be employed in several tasks and some of these can be topic-oriented. For instance, there is a growing interest in applying tasks of classification and automatic detection to hateful online communication, sentiment analysis, and opinion mining. Events such as SemEval (Semantic Evaluation) and IberEval are natural language processing research workshops that focus on advancing NLP systems (see Basile *et al.* 2019; Fersini *et al.* 2018). In these events, detection and classification of misogynistic language (Canós, 2018; Anzovino *et al.* 2018; García-Díaz *et al.* 2021) and hate speech detection against women and immigrants (Plaza-del-Arco *et al.* 2019) are common currency. Research has also been conducted in Spanish that attempts to classify different types of texts. Fernández Anta et al., (2012) employed a corpus of Spanish tweets to conduct a comparative analysis of different approaches and classification techniques in topic detection tasks. Their focus was to examine whether common approaches that have been proved effective in topic classification in English are effective with Spanish data as well. They evaluated the use of n-grams, input data, lemmas/stems, correct words, word types, hashtags, author tags, and links in topic classification tasks. Among the different classes used in these tasks were music, economy, entertainment, politics, technology, sports, and literature. The results showed that none of these sources of features proved to be highly relevant in the Spanish tweet classification; in fact, the highest accuracy for topic classification was obtained with the use of n-grams, which reached a 58% accuracy with a Naïve Bayes classifier. In a similar study, a topic classification of tweets in Spanish by topic was conducted, but this particular study is not based on the bag-of-words paradigm, which is very much common in these kinds of tasks; instead, graphs generated from the texts were used and the graph similarity was employed to classify the texts by topic. The effectiveness of graphs, in terms of element relationship representation, and the extensive mathematical work in graph theory, have been successfully exploited for many tasks such as summarization and information retrieval (Cordobés *et al.* 2014: 31). This method is based on the assumption that well-connected nodes (e.g., terms or sentences), are especially suited to be represented in a graph; in other words, this method proposes a system where very short text classification is possible by using a vector classification model for which the features are not terms, but graph metrics. The basic principle under this method is that every piece of text (tweet) can be represented as a graph. Cordobés et al. (2014) hypothesized that by knowing how to build a graph for each tweet, graphs belonging to the same topic have a common representative structure (topic reference graph). For the text classification, they looked for the similarities between the graph generated for a given text and different topic reference graphs; then, their technique used graph similarity measures to detect the topic of a piece of text. Seven thousand tweets were used in the training phase and around 60,000 for testing, and features such as the page rank, number of hits, and graph density were used. This experiment showed that some tweets were more accurately classified than others. For example, the class "politics" achieved a 78% of correct classification; however, most of the other categories did not even achieve a 50% accuracy. The authors account for this lack of consistency by the fact that the number of training texts in some categories was rather scarce. They also considered that the actual design of the system could have also influenced the results. Vilares et al. (2015) also tackled topic detection tasks taking into consideration morphological, syntactic, and psychometric information to classify Twitter messages. Their findings showed that the use of n-grams, both in words and lemmas, outperformed features based on part-of-speech and psychological knowledge.

The research covered above shows that different feature selection models, as well as different representation schemes, have been used in topic-oriented text classification. This body of research also shows that topic detection has been used for a wide range of purposes such as monitoring broadcast news and alerts (Allan 2002), information filtering (Sriram 2010), sentiment classification (Bermingham & Smeaton 2010), trending topic classification in social networks (Lee *et al.* 2011), and more recently, topic classification of environmental education (Chang *et al.* 2021), topic discovery on Covid-19 online discussions (Jelodar *et al.* 2020) and people's emotion detection during Covid-19 social isolation (Jelodar *et al.* 2021).

4. Corpus building

To conduct automatic classification tasks and examine how the keywords operate when used as features, we compiled a corpus from YouTube comments. The comments from the videos addressed issues that pertain to men, women, or the LGBT community. We employed a topic approach to identify the videos that could yield the expected information. We acknowledge that such a selection is subjective, but this exercise helps us establish an initial approximation to such a topic classification task. The following table shows the different topics used in the corpus building stage.

Table 3. Topics used to identify the videos for corpus building.

Name of corpora	Topic	# of words per corpus
V_Mujer	Street harassment, sexual harassment, femicide, feminism, human trafficking, sexist language.	731,286
V_General	Corruption. Drug trafficking, homicide, migration, kidnapping, bullying.	684,994
V_LGBT	Lesbian, gay, transgender, same-sex marriage, LGBT pride march.	425,105

One advantage of using YouTube as a source of data to build a corpus is that this social media outlet offers a great sense of anonymity and this tends to encourage users to comment on what they would not otherwise say in a public sphere. The following table shows some comments for each of the corpora.

Table 4. Comments from the YouTube videos.

	Corpus: V_Mujer	Corpus: V_General	Corpus: V_LGBT
Samples of the YouTube comments	Nada más que un buen correctivo bien aplicado en el hocico para que cambien de actitud estas pinches viejas...... Solo quieren llamar la atención para que no se sientan ofendidas, Si andan siempre con el resentimiento.	Sinceramente si a aumentando un chingo la delincuencia esperemos y el buen Obrador si de resultados que vamos de mal en peor. 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…	rompí un mandamiento y me enamoré de mi mejor amiga. arrepiéntanse, 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

This corpus building process yielded three corpora, which were later compared. The V_mujer, the V_General, and the V-LGBT corpus contained the comments obtained from 18, 17, and 16 videos respectively.

5. Data preprocessing

Once the corpora were compiled, we conducted a keyword analysis to identify statistically the keyness of the most relevant words that relate to topics concerning men and women, as well as to the LGBT community. It is customary that in keyword analyses, a corpus is compared with a reference corpus to identify those lexical items that are unusually high in comparison with a reference corpus. Given the above, each one of the three corpora was compared individually to the other two corpora, and the keywords with a keyness score of above 3 were selected for the classification. The higher the keyness, the more relevant the keywords are; this is because words with high keyness scores tell us what is peculiar about the texts they belong to.

For the sake of clarity, it is important to emphasize that two different topic classification tasks were carried out. The first classification task involved three subtasks; in the first one, we used the words of the Violentómentro, and in the last two subtasks we considered the different number of features (keywords) for each subtask.

In our first task, we employed the VSM to represent the information regarding the frequency of occurrences of the features (keywords) in each class (see Table 2). In the VSM, 18 V_Mujer comment collections (one per video), 17 V_General comment collections, and 16 V_ LGBT comment collections were represented, as well as the frequency of each feature in each of the comment collections. Table 4 shows the different features we used for the three subtasks.

Table 4. Features in each one of the three corpora.

	(Violentómetro)	**(Subtask with 30 keywords*)**	**(Subtask with 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* merec*, 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.

It is important to note that each feature was searched in each one of the comment collections using the regular expression *., which allows a concordancer to retrieve various forms associated with a root or stem query. For example, the regular expression in Spanish of *asesin** could retrieve the words *asesinar*, *asesina*, *asesinó,* and so on.

For the second major classification task, we used the string to word vector filter in Weka. This function employs a Boolean weighting scheme, which tests the classifier based on the presence or absence of the features in each one of the comments instead of their frequency. Unlike the first task in which we used the features (keywords) to classify an entire comment collection (extracted from a given video), in the second task we used the features to classify individual comments. It is also important to mention that in this second task, not only did we use the features (keywords) we had identified but also all other words in each one of the comments.

This task required the texts to undergo a more in-depth preprocessing. Since we were now classifying individual comments and not comment collections, we needed to make sure that every comment was related to the categories (each one of the three corpora) used in the classification task. Originally, we retrieved almost 100,000 comments for the three corpora, from which we selected manually 7,500 comments, 2500 per corpus. Figure 2 shows some of the comments to be classified according to their classes.

Table 6. Sample comments according to their classes.

Instances of comments that were classified
"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 burlase de dios que se justifican mediante eso" (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 obvio aria 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 feminismos no tiene sentido :/." (V-Mujer)
"Puede que si la mayoria 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)

It is important to mention that in this second major task, there were two sub-tasks. The first one included the previously identified keywords plus additional words that functioned as features totaling 1,751 features being used. In the second sub-task, some keywords were removed to assess how much these keywords were contributing to the accuracy shown by the classifiers. Retrieving the keywords was done manually. Table 5 shows the keywords that were removed in the second subtask.

Table 7. Keywords removed to assess their weight in the text classification

Keywords used in the string to word vector subtask			
Keywords related to religious terms	**Keywords related to identities and other phenomena**	**Keywords related to political terms**	**Other verbs, adjectives, and nouns.**
Dios (God)	Ideología (ideology)	Política (politics)	Acept* (accept)
Iglesia (church)	LGBT (lgbt)	Corrupción (corruption)	Defend* (defend*)
Matrimonio (marriage)	Feminismo (feminism)	Calderón (former mexican president)	Maltrat* (abuse*)
Mandamiento (commandment)	Lesbiana (lesbian)	Invasores (invaders)	Viola* (rape*)
Pecado (sin)	Equidad (equity)	Muros (walls)	Respet* (respect*)
Biblia (bible)	Feminazi (feminazi*)	Justicia (justice)	Acoso* (harass*)
Cristiana (Christian)	Bisexuales (bisexuals)	Pobre (pobre)	Culpable (guilty)
Religión (religión)	Homofóbica (homophobic)	Políticos (politicians)	Hombre* (man/men)
Casarse (get marry)	Identidad (identity)	Salarial (wage)	Mujer* (woman/women)
Familia (family	Patriarcado (patriarchy)	Amlo (mexican president)	Puta* (bitch*)
Cree* (believe*)	Igualdad (equality)	Fifis [1]	Agresión (agression)
	Gay (gay)	Prian (political parties of PRI and PAN)	Gorda (fat woman)
	Soy (I am)	Migrantes (migrants)	Victima (victim)
	Discriminar (discriminate)	Peje[2]	Muerto* (dead)
	Derecho (right)		Rata* (crook/thief)

[1] Offensive term that refers to opponents of AMLO, the Mexican president to serve office from 2018-2024.

[2] Offensive term that refers to AMLO.

	Inclusivo (inclusive) Amor (love)	Pejezombies[3] Violencia (violence) Victima (victims) Caravana (migrant caravan) Chairos[4] Chayoteros[5]

6. Results and Analysis

As already mentioned, this study involved two major classification tasks. The first task involved three subtasks. In the first subtask, we sought to explore how the Violentómetro classification would perform when using its information to classify the texts. Some of the verbs that we utilized in this first subtask were: *asesinar* 'to kill', *violar* 'to rape', and *humillar* 'to abase', among others (see Table 4).

In this first approximation to this classification, the Naïve Bayes, the Sequential Minimal Information implementation of Support Vector Machines (SVMs), and the J48 decision tree with 10-Fold cross-validation were employed. In comprehensive research surveys of TC tasks, these algorithms have been found as some of the most common and most successful ones (Rico-Sulayes 2018). The following results were obtained:

Table 8. Results obtained in the experiment. Features taken from the Violentómetro

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

The SVM algorithm yielded the best results signaling that the verbs which were used as features were relevant in the classification of the comment collections (videos). It is important to keep in mind that in text classification, features are always extracted from the very texts they are supposed to classify; considering that the features used in this subtask were not taken from the comment collections (videos), the results were competitive. To improve the results already obtained, we relied on the regular expression (*.), which allows us to retrieve all varying

[3] Offensive term that refers to the followers of AMLO.

[4] Offensive term that refers to the followers of AMLO.

[5] Offensive term that refers to the followers of AMLO.

forms that share some suffix or stem sequence; for example, the regular expression of *viol** finds words such as *violó*, *violaron*, *violan*, etc. within a dataset. Table 7 shows the results obtained once the features were adjusted.

Table 7. Results obtained in the experiment once Features were adjusted (Violentómetro)

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

The SVMs were again the most accurate at classifying the classes with an 82% of accuracy; in fact, all the algorithms performed better once the features were adjusted. This first subtask in which the information of the Violentómetro was employed helped us establish a baseline to compare the results of the next subtasks. To accomplish the next subtask, we selected 10 features (keywords) per class, and the criteria for this selection involved taking into account keywords with the highest keyness and our intuition. In other words, once the keywords with the highest keyness were identified, we chose the ones we considered would perform better. In total, we used 30 features to classify the texts of the three classes. The features we selected were regular expressions of the following words: *Dios* 'God', *respeto* 'respect', *aceptar* 'to accept', *derechos* 'rights', *discriminar* 'to discriminate', *igualdad* 'equality', *acosar* 'to harass', *defender* 'to defend', *merecer* 'to deserve', *agredir* 'to assault', *prostituta* 'prostitute', *denunciar* 'to denounce', *matar* 'to kill', *bendecir* 'to bless', *pecado* 'sin', *biblia* 'bible' and *Amlo* (acronym referring to the Mexican president) among others. The VSM contains 51 vectors labeled as class; table 8 shows a partial view of the VSM.

Table 9. Vector space model representing the frequency of features in the three classes (keywords)

Class	Dios	Derecho*	Discrimina*	Iguald*	Acosa*	Mata*	Soy	Biblia	Pecado*	Odi*	Poder	Amlo
V-LGBT	22	22	11	0	2	5	151	4	6	12	6	0
V-LGBT	173	52	22	5	5	8	51	17	23	19	9	0
V-LGBT	153	23	43	1	1	10	60	31	18	30	8	0
V-LGBT	289	36	11	6	0	11	11	23	23	28	10	0
V-LGBT	267	3	15	4	3	3	65	29	4	27	0	0
V-LGBT	326	20	12	5	1	8	112	42	75	30	9	0

V-LGBT	232	46	19	2	0	13	75	55	29	32	6	0
V-LGBT	4	2	5	2	0	5	531	1	2	16	4	0
V-General	22	22	0	1	0	12	25	4	2	12	38	27
V-General	271	35	0	0	2	88	9	4	7	2	62	31
V-General	12	13	211	1	14	8	84	0	0	14	6	0
V-General	17	5	0	0	0	19	26	0	0	11	49	277
V-General	28	70	0	0	0	128	12	1	1	2	33	98
V-General	69	44	1	2	2	36	16	0	1	3	48	205
V-General	27	7	1	0	0	25	32	36	22	15	38	5
V-Mujer	8	29	6	9	562	5	49	0	4	4	12	0
V-Mujer	76	4	0	0	1	40	47	3	3	4	24	0
V-Mujer	17	40	8	36	126	2	35	0	0	30	9	0
V-Mujer	25	41	0	0	5	68	17	1	0	5	20	7
V-Mujer	160	36	1	1	0	56	20	0	1	5	21	2
V-Mujer	29	45	6	41	17	104	51	1	0	17	16	0
V-Mujer	101	45	5	42	11	152	26	0	0	34	27	3
V-Mujer	8	63	16	238	3	23	72	1	0	27	20	0

In this subtask, we tested how relevant these keywords were to each one of the texts. It is important to remember that the texts are related to issues that concern women, men, and members of the LGBT community; in other words, these keywords directly signal the "aboutness" of each one of the texts. The VSM was again run with the same algorithms previously used, and the results are shown in Table 9.

Table 10. Results obtained in the experiment with keywords as features (30 features)

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

In this task, the accuracy of the algorithms improved substantially; all the algorithms improved, but the Naïve Bayes had the best performance with an increase of 24% in its accuracy. Once more, what these results show is that accurate features can also be identified via keyness since these features signal the "aboutness" of the text from which they are extracted. To verify that other keywords would yield the same results, another VSM was designed but this time with 243 words (features); again, keywords with the highest keyness were selected. Keywords such as *carcel* 'prison', *catolic** 'Catholic*', *cristian** 'Cristian*', *chairo**, *corrupt** 'corrupt', *impunidad* 'impunity', and *racis** 'racis*' among others were added. Results are shown in Table 10.

Table 11. Results obtained in the experiment with keywords as features (243 features)

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

The results were similar to the previous ones; the accuracy of the algorithms was maintained with all the classifiers. What these results tell us is that the keywords are good indicators signalling the aboutness of the texts, and are good features to be used in automatic text classification. However, at this point, we decided to test how these features would perform if we used separately individual comments for the classification, instead of the whole set of comments under some given video.

The second major experiment, which targets this much more challenging goal, involved two subtasks that aimed at classifying the comments within each text, but now instead of using only the keywords as features, all the words in each sentence were used as features as well. To carry out both subtasks, a string to word vector filter was employed; what this filter does is that each comment (string) is converted into a vector of words in which each word in that string becomes a feature. Also, the word to string vector filter employs a Boolean weighting scheme that tests the classifier based on the absence or presence of features in a string. In other words, if an algorithm identifies that a certain feature is present in a comment (string) then this comment or sentence has a higher chance to be classified as the class that has that feature; what the algorithm does is that it classifies based on what it learns from the comments in the different classes.

For the sake of clarity is important to remember that as a point of departure we used three classes (V-LGBT, V-General, V-Mujer), and each one of them comprised 16, 17, and 18 comment collections respectively. Our second major experiment, however, involved the automatic classification of 7,500 individual comments, 2,500 per class. The classification was run with the same algorithms employed in the previous experiments and with 10-fold cross-validation. In the first subtask, each comment is identified as a string and each word in each string becomes a feature. What the classifier does is examining which words appear in some classes and which do not; based on this, the algorithm learns which words are associated with each class. Table 11 shows the results obtained in the first subtask. As it can be seen in the table, even though the number of features being used in the classification increased substantially, the algorithms still performed well with three different algorithms.

Table 12. Results obtained when classifying comments. (1,756 features)

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

Although the above results were encouraging regarding the value of the keywords for the classification, there was still one more procedure to carry out to identify if the keywords were responsible for the accuracy being demonstrated by the algorithms. In the second subtask, out of the 1,756 features, 203 words were keywords that had been previously identified. Given this, in our last experiment, the 203 keywords (features) were not included. Table 12 summarizes the results obtained with and without the keywords.

Table 13. Results obtained when the keywords were excluded

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%

The above table shows that this time the experiment was rerun with 10-fold cross-validation and with a random split data of 66% training and 34% testing. What this means is that the data was partitioned and the algorithm was trained on the training dataset and was evaluated against the test dataset. Such techniques guarantee that the results are independent of the training data set. The results were similar with both partitioning techniques; the Naïve Bayes algorithm showed the best results; however, the table also shows the results once the keywords (features) were not included. The Naïve Bayes decreased by 15% in accuracy, the SMO by

17%, and the J48 showed the most significant reduction in accuracy with a 19% decrease.

7. Discussion

A major issue in automatic classification is the selection of correct features which can allow for more accurate classification results. Considering that in a data set there can be thousands of words, the challenge is to identify those units of language that can yield the best outcome. There is a wide variety of linguistic features that are used as attributes; some of these are parts of speech features, stylometry features, n-grams, syntax, and sociolinguistic features among many others (García-Díaz *et al.* 2021; Pang & Lee 2008). At the same time, there is a wide variety of feature selection methods from which we can choose, such as the bag of words, TF-IDF, Mutual Information, and Best Terms among many others (Deng *et al.* 2019). The main purpose of this study was to carry out automatic classification tasks and evaluate if keywords obtained from traditional corpus linguistics procedures yielded good results when used as features. The results show that the accuracy of the algorithms improved once the keywords were included in the classification tasks. It is equally important to remember that the keywords were obtained via traditional corpus linguistics procedures since, in the natural language processing field, from which this classification task originates, they do have their feature selection methods. The results in the classification tasks confirm that keywords, which are obtained via CL traditional procedures, can yield acceptable results in the ATC tasks. Such an assertion was also confirmed in the last task in which a Boolean scheme and a string to word vector filter were used, and in which the accuracy of the model reached 92% with the Naïve Bayes classifier. Furthermore, the relevance of the keywords as features was evident since the accuracy of the classifier decreased by 15% when the keywords were removed. The classification tasks here presented have shown that keywords improved the accuracy of the tasks and such an effect was noted when the keywords were removed.

The fact that keywords refer to the “aboutness” of both the texts collections and individual comments themselves shows that keywords, as operationalized in corpus linguistics, can function as effective features for challenging tasks, such as individually classifying thousands of short social media posts. Table 13 shows the keywords with more information gain; what this means is that these words were the most relevant ones when classifying the almost 7,500 comments according to the three pre-established classes. It may be obvious that some keywords appear in this list, but it may require a more in-depth and qualitative analysis to understand the appearance of many others. In this table, only 32 features are shown from a more complete list. In future work, we intend to elaborate on the peculiarities of the different features which are related to women, men, and the LGBT community.

Table 14. Ranking of the attributes (keywords) with higher information gain

Information Gain Ranking					
1	0.103781	1404 AMLO (acronym used to identify the Mexican president Andrés Manuel López Obrador*)	**17**	0.032137	834 respeto (respect)
2	0.07845	521 igualdad (equality)	**18**	0.031663	990 violencia (violence)
3	0.077355	1180 feminismo (feminism)	**19**	0.029622	170 biblia (bible)
4	0.074765	645 mujeres (women)	**20**	0.027864	1509 amlo (*)
5	0.070765	332 dios (God)	**21**	0.027849	1226 lenguaje (language)
6	0.061722	898 soy (I am)	**22**	0.027767	1329 puta (bitch)
7	0.054608	494 hombres (men)	**23**	0.027073	493 hombre (man)
8	0.053228	644 mujer (woman)	**24**	0.026572	1122 corrupción (corruption)
9	0.044589	1683 presidente (president)	**25**	0.026189	551 las (the - feminine)
10	0.041107	1221 justicia (justice)	**26**	0.023235	62 México (Mexico)
11	0.038916	1082 acoso (harassment)	**27**	0.022519	816 religión (religion)
12	0.037952	438 género (gender)	**28**	0.02094	434 gays (gays)
13	0.036661	433 gay(gay)	**29**	0.020555	873 señora (lady)
14	0.035091	599 matrimonio (marriage)	**30**	0.020532	951 una (a - feminine)
15	0.032765	500 homosexuales (homesexuals)	**31**	0.02014	1250 maltrato (mistreatment)
16	0.032527	1208 inclusivo (inclusive)	**32**	0.019915	441 gobierno (government)

In this study, we have used a traditional technique from corpus linguistics and extrapolated it to machine learning to carry out classification tasks. We consider this important not only because it has proven to be a valuable technique in terms of its results, but also because it represents a way for linguists with some programming skills to join the discussion regarding text classification tasks. In this sense, an important contribution of this research is that it calls for more interdisciplinary work between corpus linguistics and machine learning. This study has shown that there are tools and techniques in corpus linguistics that can inform classification tasks in the machine learning field. We also consider that the contribution of this research study is manifold. First, it places keywords, as operationalized in corpus linguistics, as features to carry out topic detection tasks. This research also adds up to the development of topic detection. Furthermore, taking into account that a corpus was compiled to carry out classification tasks and test the efficacy of the keywords, our research study contributes to continuing the development of corpora construction from the web. Last but not least, our study promotes research on verbal violence that women, men, and members of the LGBT community experience in online communities.

References

Anzovino, M., Fersini, E., & Rosso, P. (2018). Automatic identification and classification of misogynistic language on twitter. In M. Silberztein, F. Atigui, E. Kornyshova, 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

Allan, J. (2002). Introduction to topic detection and tracking. In J. Allan (Ed.), In *Topic detection and tracking* (pp. 1-16). Springer.

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

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

Bermingham, A., & Smeaton, A. F. (2010, October). Classifying sentiment in microblogs: Is brevity an advantage? In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 1833-1836).

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226, 107134.

Burgess, J., & Green, J. (2013). *YouTube: Online Video and Participatory Culture*. John Wiley & Sons.

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.

Chang, I., Yu, T. K., Chang, Y. J., & Yu, T. Y. (2021). Applying text mining, clustering analysis, and latent dirichlet Allocation techniques for topic classification of environmental education journals. *Sustainability*, 13(19), 10856.

Cordobés, H., Fernández Anta, A., Chiroque, L. F., Pérez, F., Redondo, T., & Santos, A. (2014). Graph-based techniques for topic classification of tweets in Spanish*. International Journal of Interactive Multimedia and Artificial Intelligence*, 2(5), 31-38.

Dalal, M. K., & Zaveri, M. A. (2011). Automatic text classification: A technical review. *International Journal of Computer Applications*, 28(2), 37-40.

Deng, X., Li, Y., Weng, J., & Zhang, J. (2019). Feature selection for text classification: A review. *Multimedia Tools and Applications*, 78(3), 3797-3816.

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.

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

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

Hundt, M., Nesselhauf, N., & Biewer, C. (Eds.). (2007). *Corpus Linguistics and the Web*. Rodopi.

Jelodar, H., Wang, Y., Orji, R., & Huang, S. (2020). Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, 24(10), 2733-2742.

Jelodar, H., Orji, R., Matwin, S., Weerasinghe, S., Oyebode, O., & Wang, Y. (2021). Artificial intelligence for emotion-semantic trending and people emotion detection during covid-19 social isolation. DOI: https://doi.org/10.48550/arXiv.2101.06484

Kadhim, A. I. (2018). An evaluation of preprocessing techniques for text classification. *International Journal of Computer Science and Information Security (IJCSIS)*, 16(6), 22-32.

Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., & Choudhary, A. (2011, December). Twitter trending topic classification. In *2011 IEEE 11th International Conference on Data Mining Workshops*, 251-258. IEEE.

Liu, H., & Yu, L. (2005) Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans Knowl Data Eng*, 17(4):491–502.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval*, 8, 1-135. DOI: f10.1561/1500000001

Pihlaja, S. (2014). *Antagonism on YouTube: Metaphor in online discourse*. Bloomsbury Publishing.

Plaza-del-Arco, F. M., Molina-González, M. D., Martin, M., & Ureña-López, L. A. (2019). SINAI at SemEval-2019 Task 5: Ensemble learning to detect hate speech against inmigrants 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: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14(1), 133-167. https://doi.org/10.1515/cllt-2015-0030

Rico Sulayes, A. (2018). *Authorship attribution on crime-related social media: Research on the darknet in forensic linguistics*. Aracne.

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

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. John Benjamins Publishing.

Sriram, B. (2010). Short text classification in twitter to improve information filtering, unpublished Master's thesis, The Ohio State University.

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.

Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2015). A linguistic approach for determining the topics of Spanish Twitter messages. *Journal of Information Science*, 41(2), 127-145.

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.