Analyzing Twitter data to discover gender biases in Spanish politics

Javier Beltrán Jorba

27 June 2018

Director: Enrique Romero Merino

Department of Computer Science

Universitat Politècnica de Catalunya - BarcelonaTech

Co-Director: Aina Gallego Dobón

Institut Barcelona d'Estudis Internacionals

Ponente: Lluís Padró Cirera

Department of Computer Science

Universitat Politècnica de Catalunya - BarcelonaTech

Master in Artificial Intelligence

Facultat d'Informàtica de Barcelona (FIB)

Facultat de Matemàtiques (UB)

Escola Tècnica Superior d'Enginyeria (URV)

Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

Universitat de Barcelona (UB)

Universitat Rovira i Virgili (URV)

Abstract

This work is built on the observation that our society is heavily affected by gender stereotypes in the way they speak and in the messages they receive from others, and that our political representatives are not immune to these stereotypes but rather amplify them due to their visibility as public figures. The traits of such conversational biases are studied through a data-driven analysis based on thousands of tweets published by Spanish politicians, as well as the tweets from individuals in reply to politicians' messages. The gender of the author of each tweet is annotated when its identification is possible, and replies are also annotated by the gender of the receiver. Then, machine learning classifiers are trained that separate these data by gender, using linear models that allow for straightforward interpretation through the observation of its coefficients. From these models, the words more indicative of a certain gender are identified, extracting revealing knowledge about the differences in how men and women politicians speak and how they are addressed. For instance, the existence of men and women topics gets confirmed, with infrastructures and territorial politics being clearly associated with male speech, whereas gender and social affairs are women things. Women are also addressed in a more condescending tone than men, who are instead insulted more directly. This qualitative analysis is followed by quantitative measures of gender bias in the corpus, calculated from an embedded representation where a gender subspace is clearly recognized, and the words that were identified previously as indicative of gender appear strongly aligned with this subspace.

Having confirmed the prevalence of these stereotypes in political conversation, a tool for its automatic identification is proposed. This requires the manual annotation of a subset of tweets, in terms of specific categories whose detection is of interest. The result is an identifier of hostile replies to politicians obtained after extensive fine-tuning, including the choice of a certain text preprocessing strategy, a set of features and a learning algorithm. The best performing configuration is a multi-layer perceptron with 1 hidden layer, using a bag-of-character-trigrams approach to obtain the features. I hypothesize that the success of ngrams of characters versus ngrams of words is a consequence of the briefness and informal tone employed on Twitter, which leads to heavy use of abbreviations and typos that are easier to manage by sequences of characters. An F1 score of 73.98% is achieved on a test set, and the main sources of error are analyzed. The two main sources identified were: the use of irony to express a hostile message with fake positive words, which causes most of the false negatives; and the fact that a message can agree with the politician it replies to and, at the same time, be hostile against a third person, leading to a number of false positives. I conclude that the two principal directions for improvement are a more complex language model that corrects the aforementioned subtleties in the tweets, as well as a larger supervised corpus, because there are many different ways to be impolite against politicians and, apparently, thousands of tweets are not enough to learn all of them.

“One should use common words to say uncommon things”

— Arthur Schopenhauer

Table of contents

List of tables

List of figures

1 Introduction

This work explores how gender stereotypes affect communication in a particular scenario: the messages of Spanish politicians and the answers they receive in the social platform Twitter. It is focused on the use of language and how it reflects (and perpetuates) diverse patterns of bias by gender, and proposes a solution based on the automatic detection of such patterns.

1.1 Motivation

Gender affects plenty of the living circumstances of society, at several levels. From the communicative point of view, gender is directly present in many languages, including Spanish, making men and women express in a grammatically different way when talking about themselves or other people. Language could also reflect social biases and stereotypes that are due to gender. Examples of such biases include the association of a conversation about cars or sports with men and a conversation about family or cooking with women. These stereotypes can be seen in how men and women express themselves, but also in how they are addressed by others in a conversation. Besides, they can happen involuntarily, because they are deeply bound to our thinking; and voluntarily, when used with the purpose of attacking or harassing, because plenty of such biases have a negative connotation for one of the genders. For instance, it is often claimed that women have worse driving capabilities than men, or that blond women are less intelligent.

There is increasing concern that the emergence of social media has generated a surge of uncivil, hostile behavior. There is psychological evidence that it is harder to talk unkindly to someone when looking at their eyes [1], whereas in social media the aggressor cannot see their victim, hence it becomes easier to be unpleasant and even hostile. While this hostility can be aimed at anyone (no one is immune to online harassment), we can expect that politicians are a major target for opponents, trolls or any citizen; because they are public figures and deal with sensitive issues that directly affect the lives of people, such as Economy or civil rights [2]. Consequently, one should consider whether their communicative acts are also affected by gender biases and how they are related to the abusive speech they face.

Among the variety of social networks available nowadays, Twitter is especially directed to the last minute, breaking news. Besides, its mechanism of posts, mentions and replies allows for direct interaction between users. This explains the importance of Twitter for contemporary political debate: indeed, 85.14% of the deputies for the Spanish parliament are registered in the social network, as well as 79.33% of the autonomous parliamentarians. Twitter allows that their messages reach a broad audience immediately, and this audience can express their agreement or disagreement by directly replying to politicians, sometimes being unkind to them.

Millions of tweets are shared on Twitter daily, becoming a reliable source for extracting the state of public opinion, for instance using data analysis techniques on which this work is focused. Such information, properly interpreted, may reveal whether men and women politicians are treated differently in social media, for instance by the use of language against them. On the other hand, acquiring, managing and exploiting such volumes of data presents a technical challenge that requires Big Data-specific solutions to make a data-driven analysis feasible.

1.2 Objectives

The first goal of this work is to analyze the text messages that politicians emit and receive on Twitter from a gender perspective. This thesis answers the following key questions on gender issues: Do men and women politicians communicate differently with the public? Are their messages treated unequally depending on their gender? Are politicians discriminated by their gender? This discrimination could come in the form of more hostile replies to women's messages, in comparison to men's, or vice versa; or even in subtler ways, like women being texted in a more condescending tone than men.

Such questions are approached with the use of intelligent methods that extract useful information from the tweets, with a focus on interpretability of the results that allows for further analysis. Specifically, this thesis analyzes the use of language conditioned by gender, both for tweets written by politicians and for tweets directed to politicians. This is achieved by fitting linear models on vector spaces consisting of features observable by humans, such as the words themselves and their parts-of-speech. Then, when properly interpreted, the coefficients of the model are explanatory of the variables used. Therefore, linear models help characterize the language used when individuals refer to women politicians, or to men politicians, or when women or men politicians express themselves. I complement this statistical analysis with a non-linear approach based on [3]: from the corpus of all downloaded tweets, an embedding space is trained and, following the procedure presented on [3], I show that the same gender-related characteristics are present in the tweets, such as the identification of a gender direction and the calculation of direct bias, with some adjustments required that are related to how the English and Spanish languages treat gender.

Secondly, while the aforementioned study is valuable for understanding the problems of gender bias in social media and it may help research on that area, this knowledge can be applied in practice by creating a detector that automatically discerns biased tweets from those that aren't, starting from the descriptive models explained above and increasing their complexity. Among the different gender biases that appear in everyday language, the initial objective was to detect sexism

and hostility. Only the results for hostility are presented, because the number of sexist tweets in the corpus is too little (around 100 samples) compared to non-sexist tweets (1:99), and so detection of sexism was discarded until more tweets are labeled. Nevertheless, most of the knowledge presented is transferable to other labels when available.

1.3 Context of this thesis

This project originates from a collaboration between *Institut Barcelona d'Estudis Internacionals* (IBEI), an inter-university institute that promotes research in politics and international affairs; and *Universitat Politècnica de Catalunya* (UPC), where this work is presented as a Master's Thesis for the Master in Artificial Intelligence. It attempts to bring data-driven methods from data science closer to research in social sciences, and has led to the presentation of a research note in the European Political Science Association (EPSA) in June 2018.

1.4 Organization of this thesis

First, a brief survey of the related work is presented in Section 2, focusing on the need for explainability when doing research in computational social sciences, in addition to the main results in automatic detection of abuse and hostility in social media. Then, Section 3 describes the data used: how the tweets were obtained and the labeling procedure to achieve a ground truth. Data transformations that were required for these tweets are described in Section 4, namely text preprocessing, assignment of gender to the authors of tweets and the different feature extraction strategies considered. Next, Section 5 explains the linguistic analysis of tweets, including the linear regression models in the aim of achieving high explainability of the importance of the variables, as well as the training of word embeddings and posterior analysis of gender bias in the embedding space. Finally, Section 6 describes the procedure followed for training a classification algorithm of hostile messages to politicians, as well as its evaluation and an error analysis. General conclusions top off the thesis at Section 7.

2 Related work

The scientific context in which this work can be inscribed is two-fold: on the one hand, it continues the research in machine learning methods for text classification that have been largely applied to problems like sentiment analysis and opinion mining, generally tackled as classification using supervised learning since [4]. That is the first statistical approach to sentiment, and presents the classic decisions about feature extraction of text, such as the use of a bag-of-words model, the size of n-grams and the parts-of-speech. [5] deepens into feature extraction and how to apply vector-space models to textual data, introducing techniques widely used in natural language processing like the pointwise mutual information (PMI) or the tf-idf measure. It is worth noting that, while most of the literature uses English corpora, one can find works centered in Spanish sentiment analysis [6]. Similar experimental practices are followed by most research in text classification, with a remarkable exception being [7], in which sentiment analysis is solved in an unsupervised manner, by calculating the ratio between the PMI of a message and the word “excellent” and the PMI of the message and the word “poor”. Even though their results did not achieve those of supervised methods, they give an insightful hint: when working with text, hand-picking a subset of words directly related to the task may be beneficial.

On the other hand, my aim is to discover the shape and extent of gender stereotypes in the tweets of Spanish politicians and in the replies they receive. Much work in related tasks has been done in Social Science, both about verbal abuse and discrimination [8] and in aggressiveness with political topics [1] or academic debate [9]. The importance of emojis as features of sentiment is also studied in Psychology [10]. Explanation and interpretation are the usual objectives when working with social data, which certainly contrasts with the prediction-guided methodology in machine learning [11]. Fortunately, there exists previous work about the use of natural language processing for social sciences [12], which focuses both on the caveats that must be considered with social data, like sample biases (e.g. about the validity of Twitter as representative sample of society or not); as well as how to interpret results and reach significant conclusions. A popular application related to this work is the automatic detection of verbal abuse, for which many examples use Twitter corpora not necessarily related to Politics [13] [14], with one remarkable exception by Amnesty International [15] that performs statistical analysis of the tweets received by Members of the Parliament (MPs) of the United Kingdom, as well as Naive Bayes for classification of abusive tweets. My aim is slightly different, trying to classify tweets no matter if they are directed to women or men, for instance a message between two men can also be biased towards other women. I haven’t found other works with the very same objective, most of the related work being sentiment analysis in general [4] instead of gender-specific, and it is not centered in the case of politicians, with the exception of Amnesty [15]. Differently to this thesis, none of the aforementioned works use Spanish corpora.

Regarding embedding representations of tweets, two levels of embedded representation have been used for text classification: the word-level [16] and the comment-level [17], with the aim of improving the representation space and providing better prediction. In the area of gender bias, [3] is the first (and, from my knowledge, the only) that defines measures of gender bias for an embedding space of words, and additionally proposes strategies to debias the embedding. My objective is not to obtain a gender-unbiased space, because that would hide the presence of such stereotypes in the corpus, but rather to follow the quantitative analysis of bias they present.

3 Data sources

The data needed for this work consists of tweets originally posted by Spanish members of the National and Regional Parliaments, i.e. their user timeline, as well as tweets directed at Spanish politicians from other people. Such interaction may take two forms in Twitter: a reply, where a user's text is posted as an answer to a politician's message; and a mention, where a user's text includes a reference to a politician (like @BarackObama), no matter if such text is a reply to another user or an original tweet. A sufficiently large corpus with this data would suffice to perform language analysis of politicians and their respondents by gender, yet there is a specific type of interaction that provides more context: the pairs formed by a politician's original message and an individual's reply to it. They represent micro-dialogues that facilitate the understanding of the general tone of the interactions and of their two parts; otherwise, it may be hard to discern the mood of an answer without knowing what it's answering to.

I didn't find any existing corpus of Spanish tweets that already satisfies these needs. There exist datasets in English for topic identification, labeled by "political" or "non-political" [18]. Others are more specific and contain the sentiment of each tweet [19]. It is unfeasible to translate these corpora into Spanish because the loss of information would be substantial, given that much of the slang language typically used by individuals in a media like Twitter would have no direct translation; besides, most political topics and affairs -- and the terms used when speaking about them -- are country-specific. In any case, there is no trace of a corpus that annotates the content of the tweets in terms of several specific phenomena, sentiment analysis being the closest approach, but negative sentiment is a much wider concept than certain stereotypes or attitudes.

What most Twitter corpora used in the literature have in common is how they retrieve such tweets in the first place: Twitter Developer Tools provides developers with several APIs that can be accessed through web endpoints. This is how I created the corpus, by downloading the politicians' original tweets from the *user_timeline* endpoint, and downloading the mentions and replies to politicians from the Search API, which allows to filter by specific criteria like such. Search API comes in different tiers at a variety of prices and functionalities. I used the Standard Search API, which is free yet limited to only retrieve tweets posted in the last week. This was overcome by running a search weekly since the start of the project.

A corpus as such should be as representative and unbiased as possible. For instance, one downloaded exclusively in a week previous to elections could only represent topics that were popular in these elections; or a corpus based only on the tweets of the leading candidate of each party would underrepresent women [15]. My sample is built from a list containing all Spanish

parliamentarians of the present legislatures, both national and autonomic. Despite that, the temporal window in which all tweets have been downloaded is small, so the corpus should be considered as a snapshot of the political moment in Spain in the dates when this work was developed. Specifically, the oldest tweets were downloaded in the week of December 18th of 2017; and the newest in the week of June 4th of 2018, making a total of 825561 tweets coming from 209812 different users. More info about the specific numbers of each type of tweet and their users is present in Tables 1, 2 and 3.

3.1 A multilingual corpus

Even though the corpus is limited to tweets from Spanish politicians and replies to them, Spanish is not the only language spoken in Spain. Indeed, the presence of Catalan tweets in the corpus is very common, while the number of Galician and Basque tweets is small but not negligible. This raises the question whether to remove tweets written in languages other than Spanish or not, in order for the models and analysis to be more specific. Yet, because multilingual interactions are extremely common (e.g. a Catalan parliamentary speaking in Catalan and an individual answering in Spanish, or vice versa) and because Spanish, Catalan and Galician have plenty of commonalities, the decision was to keep all tweets and let the models discern what's relevant and what's noisy. This results in the fact that a small portion of the most important word features in some models correspond to Catalan words instead of Spanish, with Galician and Basque never appearing. When this happens in the results presented in this thesis, the abbreviation *(CAT)* follows the word in question. Besides, when the character level is used (see Section 4.2), the aforementioned commonalities may be beneficial to detect common stems that are indicators of the same word in different languages.

3.2 Labeling the corpus

Because most of the objectives of this work are achieved with supervised learning techniques, there is the need for a supervised dataset from which classification models can learn. Specifically, the language analysis task explained in Section 5.1 requires that the tweets are labeled by gender of the author and, when they exist, the receiver; and the automatic detectors of different types of negative speech, such as sexism and hostility, require its presence or absence in every tweet as labels. While the gender of the author can be automatically identified in many cases by looking at the name of the user, a reliable ground truth for sexism and hostility has to be constructed manually.

3.2.1 Automatic gender assignment

Knowing the gender of a user is fundamental for several aspects of my analysis, because it allows to extract relations specific to a certain gender, like the use of language characterizing men and women speech. While the gender of politician users is known in all cases, it is not for the individual users that answer to them. There is not a Gender field in Twitter profiles, so it has to be estimated from other fields like the complete name of the user. Fortunately, several tools exist for gender identification of names, with the R package "*gender*" being a common option and the one that I use in this work. In general, the nature of such predictors is probabilistic and so the confidence of the gender prediction is also returned.

While some names are clearly associated with a gender, others are common in both, and others are rare so the gender identifiers are more susceptible to fail with them. My approach attempts to enhance the probabilistic assignment of gender, by bootstrapping from information that I *do* have labeled by gender: the names of politicians. Individuals whose first name is the same as any Spanish politician of the corpus have their gender assigned deterministically, and the predictor only works on the rest of individuals. The confidence obtained by the predictor ranges between 0, where the predictor is certain that a name is masculine; and 1, where it is certain that it's feminine. I discard predictions with a confidence between 0.4 and 0.6, because both genders are almost equally likely and the assignment would be close to arbitrary. There are also cases where the predictor does not return an answer because it considers the word is not a name. I label all such cases as "unknown", with 45% of the users falling into this category, and hence their tweets are omitted from the analyses that require a gender variable. The total numbers of users and tweets are described in Tables 1, 2 and 3 organized into categories like gender, the type of tweet or the type of user.

Users	*Politicians*	*Individuals*	*Σ*
Male	687	78705	79392
Female	538	35598	36136
Unknown	0	94284	94284
Σ	1225	208587	209812

Table 1: Distribution of Twitter users present in the corpus of tweets downloaded, organized by their position and their gender

Tweets by author gender	*Originals by politicians*	*Mentions to politicians by individuals*	*Replies to politicians by individuals*	*Σ*
Male	109803	66526	160649	336978
Female	64942	30504	72301	167747
Unknown	0	125480	195356	320836
Σ	174745	222510	428306	825561

Table 2: Distribution of tweets downloaded, organized by their type and gender of the author

		Receiver			
	Replies by author and receiver gender	*Male*	*Female*	*Unknown*	*Σ*
Author	*Male*	83912	33205	43532	160649
	Female	35026	17727	19548	72301
	Unknown	103131	43488	48737	195356
	Σ	222069	94420	111817	428306

Table 3: Distribution of tweets downloaded that are replies, organized by gender of the author and gender of the receiver

3.2.2 Manual assignment

The initial aim of the automatic classifiers of Section 5.1 is to detect answers to politicians that are sexist and hostile. The corpus of tweets obtained is huge (825561), and its subset consisting of replies (428306) is also unmanageably large to be completely labeled by hand in the extent of this thesis. That's why only a small subset of replies to politicians have been assigned a sexism label and a hostility label. In order to achieve a reliable ground truth, every tweet considered was evaluated independently by two people, and only those where there is agreement in the judgment are kept for the supervised sample, making it smaller but more reliable. This also allows to calculate how much human agreement is obtained for the task, using Cohen's Kappa score, which is a handy measure of the difficulty of the problem as it returns how much agreement there is relative to agreement by coincidence. That is, Kappa=1 means total agreement, while Kappa=0

means agreement as good as random. The scores obtained for the judgments of sexism and hostility were:

- Sexist / Not sexist: **Kappa = 43.78 %**
- Hostile / Not hostile: **Kappa = 53.33 %**

Kappa scores over 40% are considered moderately good, and in both cases this threshold is achieved, although by a small margin in the case of sexism. This means that, while in both cases the difficulty for humans to agree on how to classify tweets in these categories is relatively hard, it is harder for the sexism classification than for the hostility. A cause for this could be that hostility is easier to understand because we all have a similar definition in mind of what a hostile message is, whereas for sexism it is usually a topic of debate whether a certain behavior, attitude or speech is sexist or not. In any case, these percentages are not problematic for the prediction tasks because I only keep as supervised data those tweets on which there is agreement. Specific numbers can be found in Tables 4 and 5, separated by gender of the author.

	Sexist?	*Yes*	*No*	*Σ*
Author	*Male*	76	6250	6326
	Female	39	3042	3081
	Σ	115	9292	9407

Table 4: Distribution of tweets labeled by sexism, organized by their label and gender of the author

	Hostile?	*Yes*	*No*	*Σ*
Author	*Male*	1471	4383	5854
	Female	490	2423	2913
	Σ	1961	6806	8767

Table 5: Distribution of tweets labeled by hostility, organized by their label and gender of the author

Unfortunately, the labels of sexism are extremely skewed, with the class *sexist* containing only 115 of the tweets, versus 9292 non-sexist. While these odds of 1:99 surely pose a challenge for machine learning performance in small datasets, that is not the focus of this thesis and it would tangle up the completion of the objectives posed. For this reason, automatic detection is only

performed with the dataset of tweets tagged by hostility, which are also skewed but in a manageable scale (1:2).

3.3 Infrastructure

The corpus of tweets downloaded is too large to be loaded on commodity hardware in memory, which is required for tasks of this work such as fitting some of the classification models or training embeddings. In order to overcome this bottleneck, Amazon's cloud computing solution Elastic Compute Cloud (EC2) is used. All the experiments presented are run on an Ubuntu 16.04 installed on an *r4.xlarge* machine, which belongs in the *memory optimized* category of machines available at EC2. Data storage is handled with a MongoDB database that consists of two collections, one for the tweets and another for the users who appear as authors of these tweets. Such a simple schema is the main reason to work in a NoSQL paradigm, for instance duplicating some of the user's information in the corresponding tweets entries, which allows to use queries that are simpler and faster to process.

At the application level, the experiments are run using a common suite of Python 3 scientific libraries. Pandas is used for managing and preprocessing the datasets, FreeLing for text preprocessing and scikit-learn for classification algorithms and model selection.

4 Data transformations

The need to appropriately transform the data into a format that is valid for the methods used becomes especially relevant when working with text classification, because supervised methods generally require numeric inputs and because such transformations may notably change the performance of the algorithms. This section describes how the text of the corpus is preprocessed, considering the peculiarities of tweets, and how it is transformed into features for machine learning classification.

4.1 Preprocessing

Textual data is highly unstructured for several reasons. The set of lexical vocabulary that may appear is huge, if we consider its variations and the fact that languages are in permanent evolution. This can be aggravated in a context like tweets, where the language register used by most individuals is generally colloquial, makes heavy use of slang and jargon related to the topic being discussed, as well as abbreviations and other forms typical of the Internet (e.g., imho for *in my honest opinion* in English), often due to the limitation of 280 characters per message. Furthermore, Twitter has its own set of language forms that are not present in any other written form: mentions (a username preceded by the symbol @, e.g. @BarackObama) and hashtags (a short text without spaces, preceded by the symbol #, e.g. #Elections). This brings additional difficulties for an appropriate preprocessing, because these forms are not to be tokenized like the rest of the text. For instance, special symbols such as punctuation should be separated from the word they are adjacent to, but if @ and # are also separated then we miss what a mention and a hashtag is.

I use the text processing tool FreeLing to perform this preprocessing, because the range of languages it supports includes Spanish, Catalan and Galician; providing tokenization, sentence splitting, lemmatization and morphological analysis, and it can handle Twitter text. Each tweet is input as raw text and the output produced is a list of tokens (grouped by sentences), one for each atomic unit in the text (generally single words, but also entities formed by several words). A token contains its original form in the raw text as well as the lemmatized version. A parts-of-speech (POS) tag is also attached to each token, as a result of the process of morphological analysis. A FreeLing POS tag is an alphanumeric code where each digit describes a morphological feature of the word in the text, e.g. gender, number, modality. How all this information can be used as features is studied below.

4.2 Feature extraction

Because a list of words cannot be directly input to machine learning algorithms, they must be transformed into a feature space. A key aspect in the design of this work is to determine the set of features that solve the tasks proposed better. Such tasks have fairly different objectives, ranging between interpretability and predictive capability, and certain features will be more suitable for certain objectives. What they have in common is their focus on analyzing the language used by politicians and citizens in political conversation on Twitter, hence the features of interest are extracted directly from the text of the tweets and consist of the lexical units, for instance the words. A well-established method to represent words as features is the bag-of-words representation, which consists of a vector of V components, each corresponding to one of the V different words that appear in the corpus, whose value may indicate the presence or absence of such word in a sample, or its number of appearances in that piece of text, or the tf-idf of such word in a document [5]. This work uses the presence/absence for language analysis in Section 5.1, for easier analysis; and the tf-idf for detection in Section 6, because it generally achieves the best performance. By keeping this parameter fixed, the comparisons between the three approaches are omitted from the experiments and the presentation of major results gets clearer.

The bag-of-words model has the major disadvantage that the order in which the words appear in the text is lost, so the model is not aware of which words appear at the beginning, the middle or the end, or which words appear in the context of others (in this scenario, context means the positional vicinity of a word). The problem can be partially solved by several solutions. One is to extend the bag-of-words to a bag-of-ngrams, where an ngram is a succession of *n* words in the corpus. Usual sizes of *n* range between 1 and 3, which are respectively called unigrams, bigrams and trigrams. The bag-of-ngrams provides a better representation because the immediate vicinity of the words is maintained (for instance, the bigram "*stop being"* could be a good predictor of the negative sentiment of a sentence). On the other hand, the dimensionality of the feature space grows dramatically as n increases. Another solution is to use the bag-of-ngrams in a level lower than words: the characters themselves. When applied to the sentence itself, without tokenizing it into words, then the ngrams with a whitespace between letters also keep track of the vicinity between adjacent words. Besides, because a single word unigram corresponds to several character ngrams, this model is more robust to noisy variations of words, which are often due to typos (e.g. *womna* for *woman*), colloquial speaking (*cauze* for *because*), abbreviations (*dems* for *democrats*), etc. We can expect that such variations are especially common in Twitter, where the tone is mainly casual and the users are encouraged to express themselves briefly due to the limitation of 280 characters. Bag-of-POS-tags is discarded after observing it couldn't learn the data properly, and thus omitted from this text.

A problem that is common through all the bag-of-ngram models described is the choice of the vocabulary. Usually the whole vocabulary present in the corpus means an unfeasibly large feature vector size. A practical solution is to keep track of the most common words only, and this number is fixed at 5000 in all the experiments of this thesis, after observing that fewer words yield worse performance and more words mean too high dimensionality relative to the size of some datasets used, such as the supervised tweets. The aforementioned strategy cuts out the tail of the words frequency distribution. Interestingly, cutting out the head may be beneficial too. The most common words usually correspond to stopwords, which give no semantic information about the text in which they are, because they are present in the vast majority of messages written in that language. Examples for Spanish are "*un, por, mientras, de"* (English for *"a, by, while, of"*). On the other hand, some stopwords may actually be beneficial, especially for ngrams with n>1. For instance, think of the bigram for "*you're a* (directed at a female)" in Spanish: "*eres una"*. It might result in an accurate indicator of the sexism of a message. In this work I choose whether to maintain the stopwords or not as part of the fine-tuning of the models. The feature vocabulary also depends on the preprocessing of the tweets because, as explained above, two forms of every tokenized word can be used: the real form, as appears in the text; or the lemma, in which several versions of the same word get reduced to a common lemma, such as different verbal conjugations.

Here follows a list with all the feature extraction strategies used in this work, and an abbreviation used as identification in the rest of the document. Depending on the ngrams used:

- Bag-of-word-unigrams: **W1**.
- Bag-of-word-bigrams: **W2**.
- Bag-of-word-trigrams: **W3**.
- Bag-of-character-trigrams: **C3**.
- Bag-of-character-4grams: **C4**.
- Bag-of-character-5grams: **C5**.

Depending on the word form used:

- Real forms from text: **FM**.
- Lemmatized forms: **LM**.

Depending on the stopword removal:

- Maintaining stopwords: **SW**.
- Removing stopwords: **NOSW**.

For instance, a feature extraction strategy that consists of bag-of-unigrams extracted from lemmas and removing stopwords will be mentioned as **W1-LM-NOSW**, and so on with the different combinations.

5 Language analysis

The first task is to study the use of language by different groups of Twitter users associated with the problem presented, namely politicians and those individuals that answer to them, and how such language changes as a function of their gender. This exploratory analysis is intended to guide understanding on how gender biases are present in everyday conversation about politics.

5.1 Linear models

This pursuit of interpretability leads to think of linear models as a tool for modeling the use of language. Because I'm interested in the differences between messages from (and to) men and women, the task is approached as a binary classification problem: a linear classifier is trained from the tweets downloaded, using gender as the binary target variable. Depending on the question being posed to the model, a certain subset of the tweets is used, with the following two being explored in this work:

- Tweets written by politicians, using their gender as target. This allows to find trends about the use of language by politicians depending on their gender.
- Tweets written by individuals as a reply to other tweets from politicians, using the gender of the politicians as target. This allows to find differences in how people write to politicians depending on their gender.

The features extracted of tweets to solve this task have been chosen prioritizing their human interpretability. That's why, out of the different sets of features defined in Section 4.2, the bag of lemma unigrams is used, which summarizes variations of the same word (like the gender or number) into a common lemma. This allows to check whether some words are indicative of gender, no matter if such words have gender variation (e.g. "*guapo/a"*) or not (e.g. "*feminismo"*). If the gender were not summarized and we could find "*guapo"* and "*guapa"* as different features, it would hold little interest to observe that they are indicative of the male and female gender, respectively. Certainly, some cases will be lost in which the male and female versions are actually used in different contexts of interest, but in return the model will be focused on clearer differences, where concepts themselves are strongly related to a specific gender.

Features are extracted in their binary form, that is, they only account for the presence/absence of words in the message, instead of keeping a count of their frequencies, in order to facilitate interpretation of the models used. The use of single words instead of deeper n-grams is also motivated by the search of interpretability. Character n-grams are discarded because they provide

little interpretation. POS tags were actually considered as features, but they couldn't produce a good model, with all coefficients ranging close to 0, which means that the morphological properties alone aren't indicative of gender at all. For instance, a hypothesis was that one gender might make further use of imperative verbs than the other, but this is not the case. Hence, results for POS tags are omitted and only those for bag-of-unigrams are described. Stopwords are removed because they provide no semantics (hence, the model is **W1-LM-NOSW**).

Among the set of linear binary classifiers available, my first choice is logistic regression because of its high interpretability relative to other methods. While its coefficients are proportional to how much a feature is associated with the outcome, logistic regression allows for an even more explainable value: the odds ratios, which can be translated directly from the coefficients and are linearly related to the outcome, that is, they explain how predicted probabilities change (multiplicatively) when a variable is modified. In the case of binary variables like our words, an odds ratio of *r* in word *w* can be explained as "the presence of word *w* makes a message *r* times more likely to be of a specific gender compared to the absence of *w*, *ceteris paribus*". The same could not be said about coefficients, due to the non-linearity of the logistic function they are affected by. For this reason, results for logistic regressions report not only the features and their coefficients but also their translation to odds.

Two different models are trained, depending on the regularization strategy: one is Ridge Regression (L2 regularization, I call it **L2** for short), the other is Lasso (**L1**, idem for short). [9] uses Lasso to detect gender stereotypes in academic discussion, and the model has a theoretical advantage for this task: its tendency to pull coefficients to zero in the optimization, hence providing implicit feature selection. This is useful for interpretability [20], as the set of features to be understood by humans becomes reduced to a relevant set, which is especially helpful given the large amount of variables that bag-of-words models contain. Also, Lasso doesn't shrink the magnitude of non-zero coefficients as L2 does. This shrinkage could make it harder to observe the relative differences in relevance between variables, as their values would be generally smaller and closer to each other. In contrast, L2 could result in a smoother model that generalizes better. The following equations express the cost function of logistic regressions with L2 and L1 regularization, respectively:

$$J_{L2}(w) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}x_i w - y_i\right)^2 + \frac{1}{C}\|w\|^2$$

Equation 1: Cost function from logistic regression using L2 regularization

$$J_{L1}(w) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}x_i w - y_i\right)^2 + \frac{1}{C}\|w\|^1$$

Equation 2: Cost function from logistic regression using L1 (Lasso) regularization

The first term of minimization is the mean squared error and the second is the regularization term, and they are weighted by the parameter C, which in this case is the inverse of the regularization strength. Besides logistic regression, another model that allows interpretation of coefficients is **SVM** with L2 regularization and a linear kernel. Specifically, the linear kernel formulation can be rewritten as a linear combination of *data* times *coefficient*:

$$f(x) = \sum_{x_i \in S_v} \alpha_i x' x_i = \sum_{x_i \in S_v} \alpha_i \sum_{j=1}^{N} x^j x_i^j = \sum_{j=1}^{N} x_j \left(\sum_{x_i \in S_v} \alpha_i x_i^j \right) = \sum_{j=1}^{N} x_j \beta_j$$

Equation 3: Transformation of the linear kernel formulation of SVM into beta coefficients which are equivalent to those in logistic regression

The strength of regularization in the decision boundary can be adjusted and poses a trade-off against misclassified training samples. Again, this can be adjusted through the parameter C, as expressed in the following formulation of the SVM minimization (Equation 4), where the first term corresponds to the weights regularization, as in the L2 formulation above, and the second is the hinge loss, which accounts for samples falling on the wrong side of the margin, and C leverages its effect on the cost function.

$$J_{SVM}(w) = \frac{||w||^2}{2} + C \sum_{i=1}^{N} max(0, 1 - y_i (w x_i + b))$$

Equation 4: Cost function from SVM using L2 regularization

Because the effect of the C parameter is comparable in all three cases, as it poses a trade-off between the cost function of each method and the regularization penalty, I follow the same procedure to fine-tune the models of all algorithms: using 10-fold cross-validation for every configuration, each consisting of a model (L1 regression, L2 regression or SVM) and a certain value for C, reporting the metrics for the training and validation sets, averaged through the cross-validation. The numbers tried for C are in a logarithmic scale between 0.0001 and 100. The evaluation of each configuration uses the Area Under the Curve (AUC) score of the ROC curve, which for the binary case coincides with the accuracy balanced for each class, i.e. male and female. This allows to check the standalone performance and fine-tuning of each algorithm, as well as a comparison between them.

5.1.1 Results from tweets authored by politicians

The first experiments take a sample of 100000 tweets written by politicians, whose gender is identified in all cases. 10000 of these are kept for testing, and the other 90000 are cross-validated in 10 folds, i.e. each time 9000 tweets are used as validation and the other 81000 as training. The hyper-parameter C is adjusted looking at value of the AUC score in the validation set. These scores for each C between 10^{-4} and 10^{2} are presented in Figures 1, 2 and 3 for the L1, L2 and SVM algorithms, respectively. Two trends are clear: in the logistic regressions, C is directly related to overfitting, with small values meaning a high regularization that simplifies excessively the model and thus makes it underfit; and high values meaning small regularization and a model too adjusted to training examples that generalizes worse to unseen data, with the best validation score in an intermediate value. Similarly, results for SVM express the trade-off that C poses between the decision surface, more complex when C is large, and the tolerance for misclassified samples, higher when C is small. Neither a very complex boundary nor a high amount of training misclassifications leads to the best out-of-sample error, which again lies between both. A summary of the scores of the best model for each algorithm is presented in Table 6.

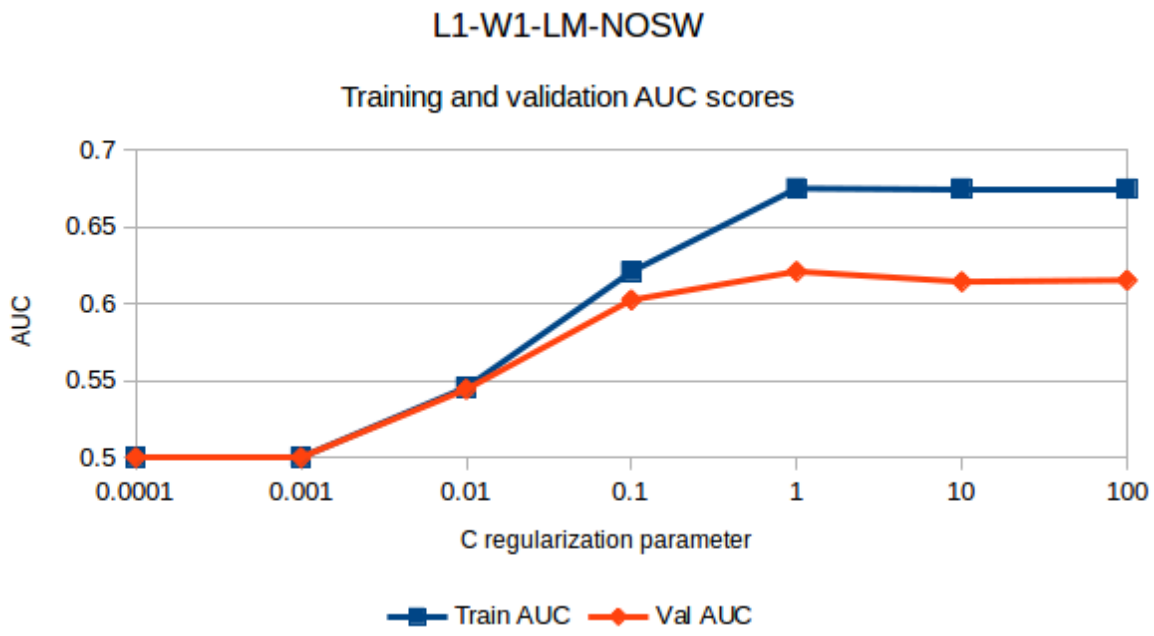

Figure 1: Overfitting curve of the models trained with L1 in the task of language analysis by author gender

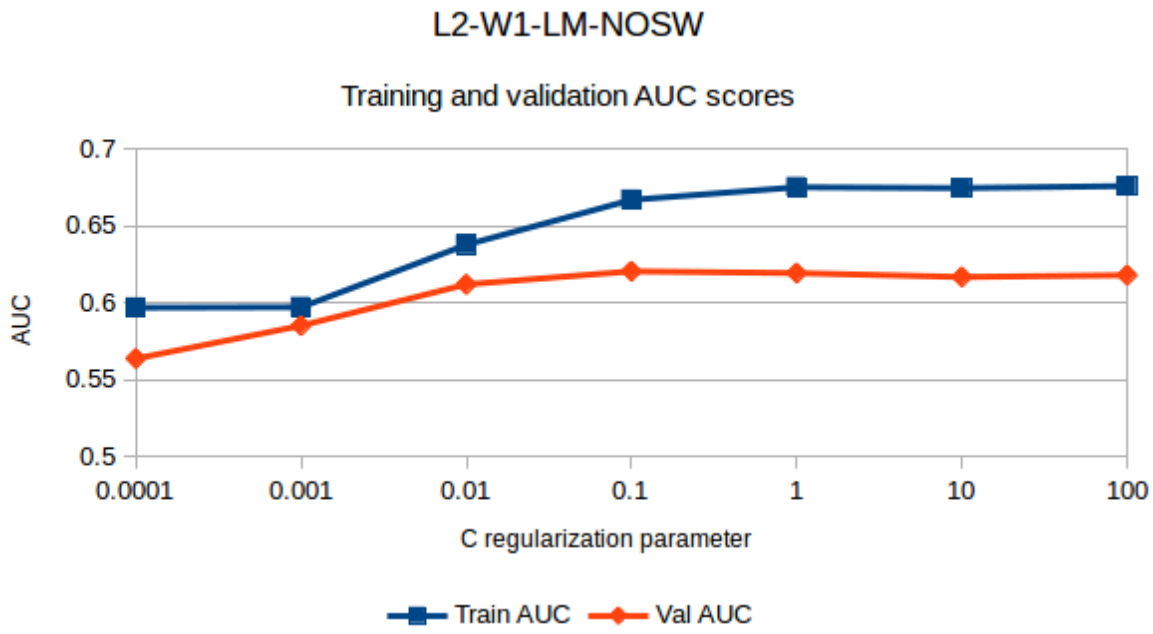

Figure 2: Overfitting curve of the models trained with L2 in the task of language analysis by author gender

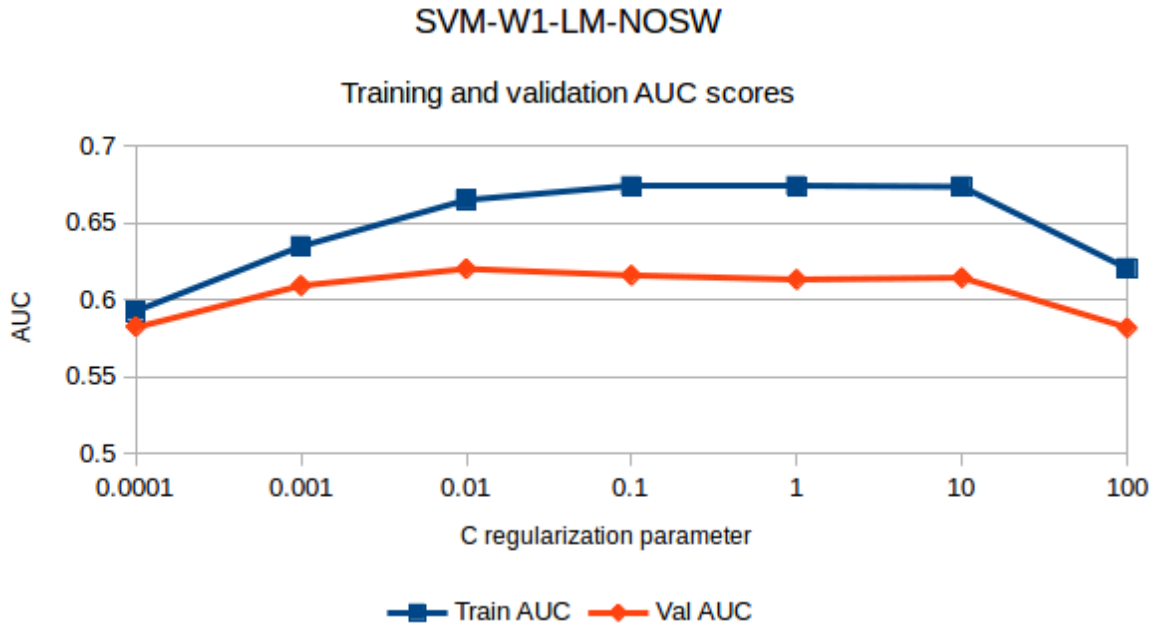

Figure 3: Overfitting curve of the models trained with SVM in the task of language analysis by author gender

Model	*C*	*Training AUC*	*Validation AUC*
***L1**-W1-LM-NOSW*	1	67.51%	**62.11%**
***L2**-W1-LM-NOSW*	0.1	66.70%	62.05%
***SVM**-W1-LM-NOSW*	0.01	66.50%	62.02%

Table 6: Training and Validation AUC scores for the best model obtained by each algorithm in the task of language analysis by author gender

Regarding the comparison between the three algorithms, all validation AUCs range around 60%, and the best value is obtained by the L1 logistic regression with C=1, by a small margin with the others. **The test AUC score achieved by this model is 60.93%**. It is rather low, given that a classifier assigning labels randomly would obtain an AUC of 50%. The interpretation of a model with limited predictive capabilities should be taken cautiously, so I follow these results with a qualitative analysis based on the features and their coefficients assigned by the best models. To do so, I check the top 100 features of highest positive coefficient (i.e. the most indicative of male gender) and the top 100 of highest negative coefficient (i.e. the most indicative of female gender), for the best model of each algorithm. If the logistic regressions with different regularizations and the SVM, which is trained with a completely different algorithm, share a fair amount of features, this could indicate that:

1. The features shared are meaningful for classification, especially those that are common to all three models.
2. Even though these models are not suited for correctly guessing the gender of *any* tweet, they still manage to arrive to a certain success, e.g. a certain amount of *easy* cases get correctly classified; and this allows for explainability: even though the model cannot completely explain what makes a tweet male or female, it finds several hints in this direction.

The coincidences in features between each model are shown in Table 7, both for the male and the female side of the model. There is notable agreement: 37 of the top 100 male features are shared by the three models, and this number rises to 52 for female features. Interestingly, L2 regression and SVM have more coincidences than they have with L1, probably due to the implicit feature selection performed by the latter, that sets to 0 those features that it finds irrelevant in the fitting, so a number of features are directly dropped to the end of both lists, male and female, making it harder to coincide with the L2 and SVM top features.

Intersection of models	*Male features*	*Female features*
L1 - L2	46%	60%
L1 - SVM	47%	59%
L2 - SVM	65%	78%
L1 - L2 - SVM	37%	52%

Table 7: Percentages of shared features by each combination of algorithms in the task of language analysis by author gender, using the results produced by the best models

What follows (Table 8) are lists of words that can be found as the most indicative male features in the models obtained, that is, words that are mostly associated with the category “the author of the tweet is a man”, sorted into categories and only if they are present in the lists of at least two algorithms. Words that were originally gendered in Spanish but are lemmatized as features are indicated with both genders.

Politics / Ideology	*Transport / Territory*	*Emotions*
Bipartidismo *(=Bipartidism)*	Aeropuerto *(=Airport)*	Joder *(~Fuck)*
Comunista *(=Communist)*	Avería *(=Fault)*	
Concejal/a *(=Councillor)*	Inauguración *(=Opening)*	
Consejería *(=Ministry)*	Millora *(CAT) (=Improvement)*	
Derecha *(=Right)*	Radial *(=Radial road)*	
Economía *(=Economy)*	TAV *(=High Speed Train)*	
Eleccions *(CAT) (=Elections)*	Tramo *(=Stretch)*	
Inversió *(CAT) (=Investment)*	Trasvase *(=Diversion)*	
Tabarnia *(Entity)*		
Unanimidad *(=Unanimity)*		

Table 8: Examples of top male features by author gender, organized into topics

Regarding the most indicative female words, these are some categories found (Table 9):

Gender	*Social affairs*	*Emotions*
Estereotipo *(=Stereotype)*	Asistencial *(=Assistive)*	Bravo *(Interjection)*
Feminismo *(=Feminism)*	Pensionista *(=Pensioner)*	Contentar *(=To please)*
Feminista *(=Feminist)*	Prevención *(=Prevention)*	D.E.P. *(=R.I.P.)*
Machismo *(=Machismo)*	Prisión *(=Prison)*	Doloroso/a *(=Hurtful)*
Machista *(=Machista)*		Harto/a *(=Jaded)*
Maltratador/a *(=Abusive)*		Inhumano/a *(=Inhuman)*
Masculino/a *(=Male)*		Precioso/a *(=Precious)*
Prostitución *(=Prostitution)*		
Sexual *(=Sexual)*		

Table 9: Examples of top female features by author gender, organized into topics

Discussion The tone of the messages changes visibly depending on the gender of the author: male politicians speak more often in terms of classic politics, such as ideological categories (e.g. *communism*, *right-wing*), institutional positions like *councillor*. Interestingly, there is a political topic clearly associated to men: territory, infrastructures and transportation, with words like *airport*, *high-speed train*, *water diversion*, *fault*. This confirms a common stereotype: engineering-related issues being a men thing, even within our representatives. On the other hand, other issues are clearly associated with women: one is Gender, with words like *feminism*, *feminist*, "*machista"*, "*machismo"*; as well as *stereotype*, *prostitution*, *sexual*. What's most surprising is that no words of this kind appear in the top male-features, which could mean that men aren't participating in the debate about feminism that is so common nowadays, or that women have been extremely mobilized about it in comparison to men. A cause of the latter could be the relevance that the Women's Day (March 8th) has had in Spain in 2018.

Social affairs are another political topic greatly associated to women politicians, which materializes in words like *prevention*, *prison*, *pensionist*, *assistive*. Finally, plenty of words in the top female features are emotionally charged, such as *hurtful*, *precious*, *inhuman*, *R.I.P.*, "*bravo"*. The only word with a comparable emotional charge in the top male features is the interjection "*joder"*, which is the only offensive word present in the lists of words by politicians. There is another source of emotions in tweets (and online messaging in general) that is emojis. Given the informal tone that is generally used in Twitter, the use of emojis is common even among our institutional representatives, and some can be found in the lists of top features obtained. Again, what's interesting are the differences between the tone in male and female emojis (also presented in

Table 10): men's are generally factual, such as a film projector, a data chart, a camera, a printed page; the only faced emojis are an elder and the grimacing face. On the other hand, women's emojis show more faces, like the screaming face, a smiling face, a face with heart eyes, a face throwing a kiss or an angry face; several hearts also appear, and among the factuals we find a phone, a rose, a green tick and a speaker. In summary, the messages of female politicians look more expressive thanks to their use of faces and hearts in comparison to males.

Emojis by male politicians	
Emojis by female politicians	

Table 10: Examples of emoji features of men and women by author gender

These results corroborate that the gender stereotypes present in our society are also reproduced by politicians in social media, which are communicative channels where biases and stereotypes are already present at many spheres, so our representatives could be affected and amplify the reach of these social phenomena. Questions remain about the distribution of these linguistic characteristics among social groups, such as age (are older politicians more prone to gender clichés?), geography (do all Spanish regions behave equally with respect to these results?) or political affiliation (are some parties more reluctant to follow gender stereotypes than others? Is there a correlation with the usual left-right political axis?). Finally, the complete lists of top 100 features for both genders using the three algorithms can be found in Appendix I. Some words in the list appear gendered despite the lemmatization strategy. This can be due to difficulties in the identification of the word, which is often aggravated by the existence of tweets that use more than one language. Besides, lemmatized words appear in a neutral form or, if it not exists, in masculine.

5.1.2 Results from tweets replying to politicians

This experiment takes 100000 tweets that are answers to politicians, and learns to predict the gender of the politician receiver. The separation of the sample into training, validation and test is the same as in the previous experiments as well as the cross-validation, model selection and evaluation measures. Figures 4, 5 and 6 show the training and validation AUC for the different

values of the hyper-parameter C with L1, L2 and SVM respectively; and the best scores achieved by each algorithm are summarized in Table 11.

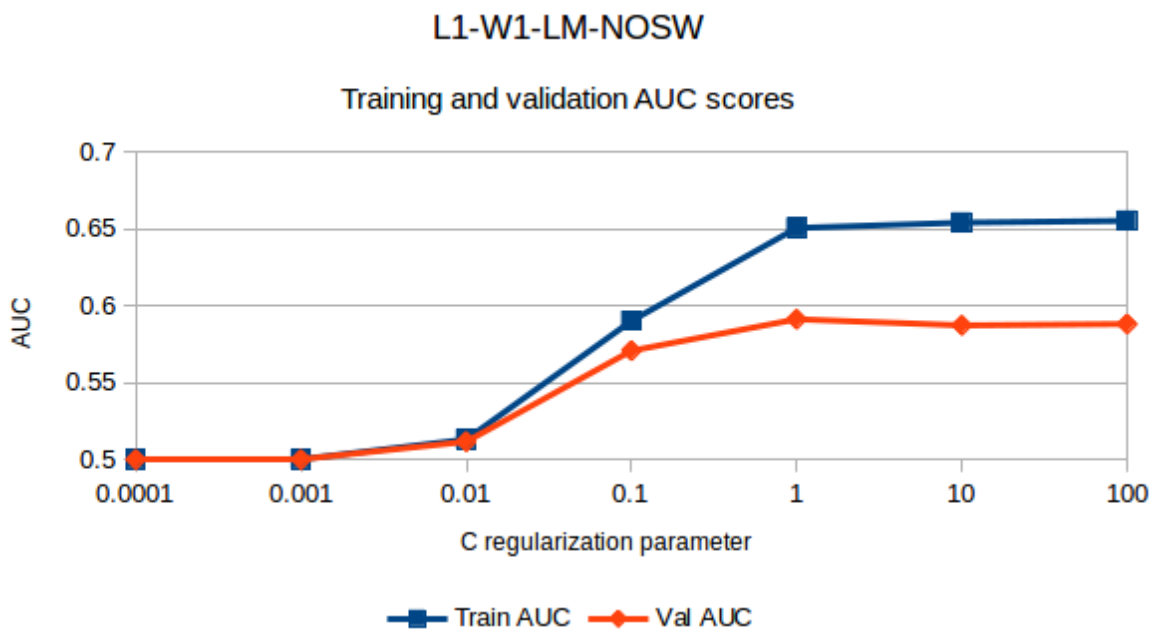

Figure 4: Overfitting curve of the models trained with L1 in the task of language analysis by receiver gender

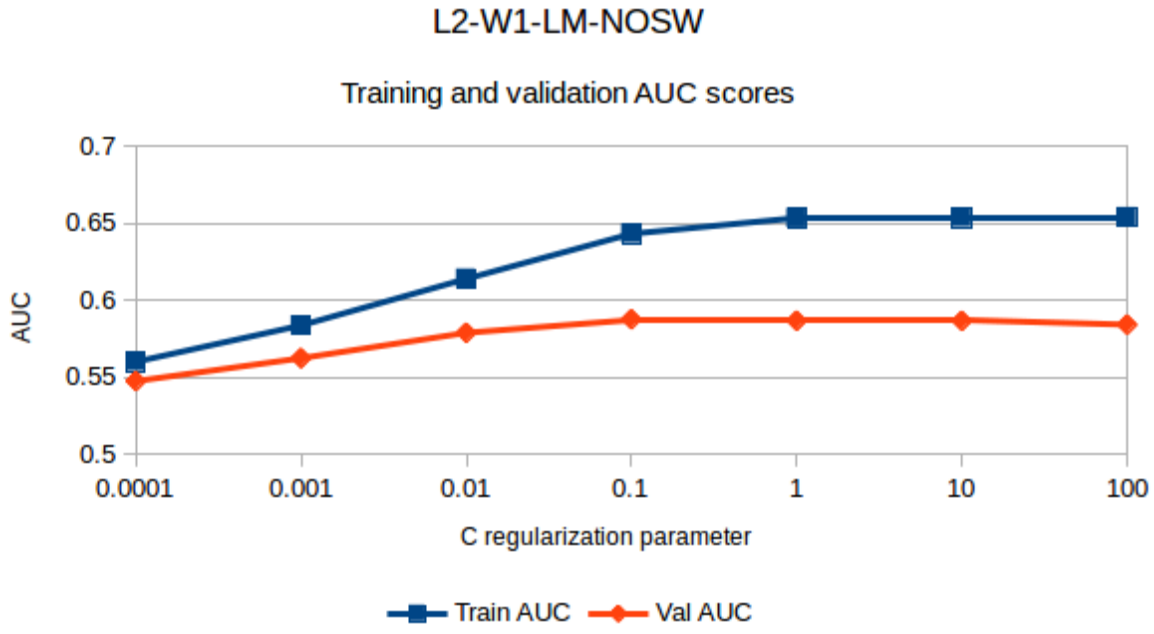

Figure 5: Overfitting curve of the models trained with L2 in the task of language analysis by receiver gender

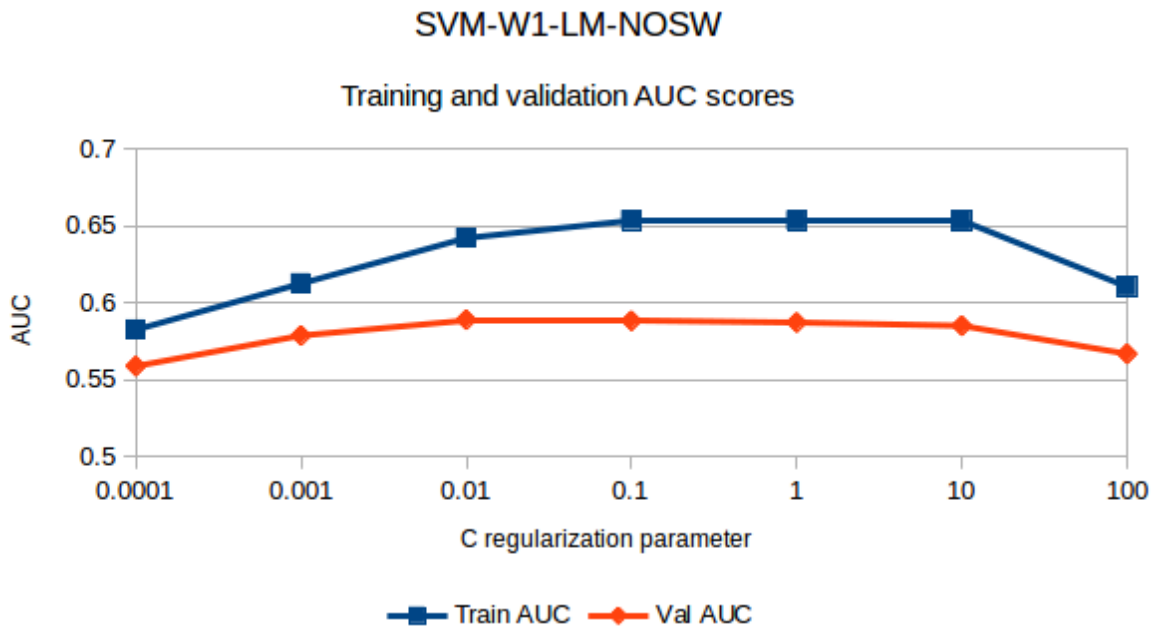

Figure 6: Overfitting curve of the models trained with SVM in the task of language analysis by receiver gender

Model	*C*	*Training AUC*	*Validation AUC*
***L1**-W1-LM-NOSW*	1	65.07%	**59.13%**
***L2**-W1-LM-NOSW*	0.1	64.33%	58.77%
***SVM**-W1-LM-NOSW*	0.01	64.22%	58.90%

Table 11: Training and Validation AUC scores for the best model obtained by each algorithm in the task of language analysis by receiver gender

The AUC scores for validation sets range between 50% and 60%. Again, the best model is L1 with C=1, **getting 58% AUC on the test set**. All numbers are slightly lower than in the previous experiments, from which I hypothesize that this problem could be of more difficult nature, that is, the differences between how men and women politicians express themselves are clearer than the differences between how men and women politicians are replied. Surprisingly, the agreement between different algorithms is higher as it can be seen in Table 12. The coincidences between L2 and SVM are even greater in this case, with 80 out of the 100 top features shared. This is may be due to a smaller set of word features that greatly affect the outcome, and a larger set of *difficult* cases for all the algorithms, so they end up finding the same words, which are clearly not enough for reliable prediction. This corroborates the inherent difficulty of the problem.

Intersection of models	*Male features*	*Female features*
L1 - L2	49%	67%
L1 - SVM	48%	68%
L2 - SVM	80%	80%
L1 - L2 - SVM	45%	59%

Table 12: Percentages of shared features by each combination of algorithms in the task of language analysis by receiver gender, using the results produced by the best models

What follows (Table 13) are lists of words that can be found as most indicative male features in the models obtained, that is, words that are mostly associated with the category "the receiver of the tweet is a male politician" sorted into categories and only if they are present in the lists of at least two algorithms. Again, words whose gender is lost after lemmatizing appear with both possibilities.

Proper nouns	*Political positions*	*Mentions to people*	*Offensive words*
Eduardo	Alcalde/sa *(=Mayor)*	Empresario/a *(=Businessman)*	Facha *(~Fascist)*
Juan	Ministro/a *(=Minister)*	Señor/a *(=Sir/Madam)*	Gilipollas *(Insult)*
Lluís *(CAT)*			Payaso *(Insult)*
Rafa			
Sports	*Transport / Territory*	*Emotions*	
Árbitro *(=Referee)*	Barco *(=Ship)*	Amenazar *(=To threaten)*	
Penalti *(=Penalty)*	Infraestructura *(=Infrastructure)*	Crack *(~Great guy)*	
	VPO *(=Social housing)*	Felicitats *(CAT) (=Congratulations)*	
		Frustración *(=Frustration)*	
		Violento/a *(=Violent)*	

Table 13: Examples of top male features by receiver gender, organized into topics

Regarding the most indicative female words, these are some categories found (Table 14):

Proper nouns	*Political positions*	*Mentions to people*	*Gender*
Anna *(CAT)*	Conseller/a *(CAT) (=Councillor)*	Guapo/a *(=Beautiful)*	Feminismo *(=Feminism)*
Adriana	Portavoza	Guapi *(=Beautiful)*	Feminista *(=Feminist)*
Eva		Rey/Reina *(=King/Queen)*	Género *(=Gender)*
Marta		Tío/a *(=Dude)*	Igualdad *(=Equality)*
Teresa		Tranquilo/a *(=Calm)*	Machista *(=Machista)*
Zaida			Mujer *(=Woman)*
			Patriarcado *(=Patriarchy)*
			Patriarcal *(=Patriarchal)*
			Violar *(=To rape)*

Table 14: Examples of top female features by receiver gender, organized into topics

Discussion As with the tweets authored by politicians, there are significant differences in the tone of the replies that politicians receive depending on the gender they have. For instance, men are directly attacked through offensive, explicit words more often, whereas none of such offensive terms appear in the top female features. Of course this doesn't mean that women are immune to online harassment, but it may take forms that are different from being directly insulted. For instance, if we take a look at the way men and women are addressed by qualificatives, some terms are ambiguous in Spanish because they're sometimes used with hostility in the examples of the corpus (e.g. *sir / madam*, *boy / girl*, *dude*), but some general trends are observed: there is much more variety of terms used when addressing to women, from "*reina"* to *girl*, and none refer to job positions as happens with the male feature *businessman*. Besides, mentions about the attractive of the receiver only appear in women, namely *pretty*. Women are apparently receiving more messages with a condescending tone, being infantilized or objectified for their physical aspect.

Beyond references to people, other relevant word features are also conditioned by gender: because messages about transport, infrastructures and territory are indicators of male tweets, men also receive messages about those topics and so features like *ship* or *infrastructure* appear in the list. On the other hand, because gender issues are highly indicative of female tweets, they also

receive messages about gender and some relevant features are *woman*, *rape*, *feminism*, *equality*. In summary, the way citizens address to their representatives on Twitter depends on both the topics that they use and the way the society treats men and women differently, which corroborates that gender stereotypes are strongly reflected in how we communicate. For more information, the complete lists of top 100 features for both genders using the three algorithms can be found in Appendix II. Again, some gendered words may appear in the lists and others are lemmatized but presented in the masculine form.

5.2 Gender bias in the embedding space

The previous results are achieved with models that are supervised and linear with respect to how the variables are combined to produce an output. While the choice for linear classifiers was purposely made due to their simplicity for explanation, they suffer from the limitation that they could fail to represent complex, non-linear decision boundaries between the categories, and let's not forget that the quality of interpretations will be proportional to the fitting of the model to the real distribution. Besides, such interpretations are limited to the coefficients of the model (or equivalently the odds), precisely how much a variable affects the outcome. Consequently, in order to obtain a more comprehensive analysis of the use of language by gender, I'm also considering other approaches, based on *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings* [3]. First, it introduces several ways to measure gender bias in a corpus through the application of distance metrics in an embedding space trained from such corpus, and then it proposes strategies to debias the embedding. I'm interested in the former, which allows to measure quantitatively the degree of gender bias in our political tweets. This is the only work I've found that reaches sociolinguistic conclusions out of word embeddings, but it has key implications for the suitability of data-driven methods and algorithms in a world that is increasingly concerned with the dangers of automatization for our society, such as the amplification of gender bias in intelligent systems.

A word embedding is a transformation of the representation of the words present in a corpus: from a one-hot encoded representation, where each component of the vector corresponds to a word in the corpus and it is set to 1 to represent such word, keeping the rest at 0 so the vectors are sparse; to an embedded representation which is lower-dimensional, dense and takes continuous values. In practice, well-trained embeddings keep useful semantic properties between words, like the renowned "*man is to king as woman is to queen" [21]*. Such characteristics of the vector space are analyzed in [3] with the purpose of finding gender traits in the embeddings, namely the direction that specifies gender (i.e. a direction in which words are arranged from male to female). Unfortunately, embedding vectors become *black-boxy*, and one cannot see the traits they're

interested in, such as the gender, as observable features. [3] resolves the calculation of the gender direction in the following way:

- Choosing a small set (10) of pairs of words that are gendered by definition and mean the same with respect to each gender, like *she-he* or *man-woman*.
- The difference between the embedding vectors of each pair is calculated, and will be called difference vector.
- Principal Components Analysis (PCA) is performed on the dataset of difference vectors, and a space is produced where the first principal component explains the majority of the variance in the gender-differences vectors. This first component is called the gender direction or gender subspace.

From the gender subspace, a metric for gender bias is created that studies the degree to which words that should be gender-neutral are aligned with the gender direction, measured with the cosine similarity between both. The average of this metric along a set of gender-neutral words in the corpus is called direct bias, and its value is proportional to the gender bias present in the text. The underlying intuition is that, if definitionally ungendered word vectors are aligned with the *she-he/woman-man/etc* direction, then they become gendered in practice by their use in the corpus and the contexts in which they appear.

5.2.1 Application to this work

It must be noted that the work presented in [3] uses, as most of the literature in embeddings and natural language processing in general, an English corpus. In this case, the differences between English and Spanish language affect the application of the gender bias measures proposed. In Spanish, many more nouns have gendered versions than in English. Hence it becomes difficult to obtain a significant set of gender-neutral words to be compared with the gender direction. For instance, [3] uses occupational words like *architect*, *nurse*, *homemaker*, *philosopher*, and finds out they are notably aligned with gender even though they are all applicable to men and women. None of such words are gender-neutral in Spanish, becoming “*arquitecto/a”*, “*enfermero/a”*, “*amo/a de casa”*, “*filósofo/a”*, which are inherently gendered and then they are unsurprisingly aligned with the gender direction. Besides, the corpus with which I’m working belongs to a very specific domain: political debate in Twitter between politicians and citizens. Consequently, the most common occupational words become *"presidente/a”*, “*diputado/a”*, “*concejal/a”*, “*ministro/a”*, etc. The majority of such occupational words are also indicative of gender, as it can be seen from the relevant words obtained in the previous section.

My solution is to use another set of words that are not occupations. Instead, I take words from the lists of top features from the previous section that are ungendered and can be applied to persons

or social groups. For instance: "*feminista", "coherente", "patriarcal"* (see Table 15 for more examples); but not "*violador/a", "guapo/a", "rey/reina"*, because they are all gendered.

5.2.2 Experiments and results

The embedding spaces used in my experiments are trained using the GloVe algorithm [22]. The size of the resulting vectors is a free parameter of the method, and it may drastically change the representation of words. For this reason, I adjust it by training several spaces of different dimensions and keeping the embedding that maximizes the gender component. This is the embedding that I use to return a measure of direct bias, as it is the one that best captures the gender component according to [3].

Gender subspace The relevance of the gender direction is not defined only by how much variance is explained by the first component, but how much is explained relative to the other components, in order to make sure that a single direction is enough to represent gender. In [3], using vectors of size 300, the first component explains 60% of the variance, and it is 6 times larger than the second component, which explains approximately 10%. The values that I obtain for each vector length are presented in Figure 7, showing the variance of the first and second components, and indicating the ratio between them. In my experiments **the best value is achieved by vectors of size 300, where a single direction explains 41% of the variance and it is 5.25 times larger than the second component**. The gender subspace appears to be more represented as dimensionality grows, but up to a certain point from which the ratio starts decreasing. A possible explanation for this is that small vectors have limited capability of representation, whereas large vectors are harder to train so the embedding space achieved could be worse. Consequently, subspace measures like such but extended to other domains are a reasonable evaluation for embedding spaces. A comparison between the percentages of variance explained by the 5 first dimensions of this gender subspace and a subspace created from random words (averaged through 100 runs) is presented in Figure 8, corroborating that there is a significant difference in the subspace obtained.

A question raises on why the numbers are smaller than in [3], and it could be due to their corpus (consisting of articles from Google News) being actually more biased by gender than mine, but it could also be due to the corpus size, which is much larger in their case so they might have achieved a better embedding representation. Therefore, a natural follow-up of my work is to increase the number of tweets, which only requires more weeks of Twitter API, and then check if these numbers change.

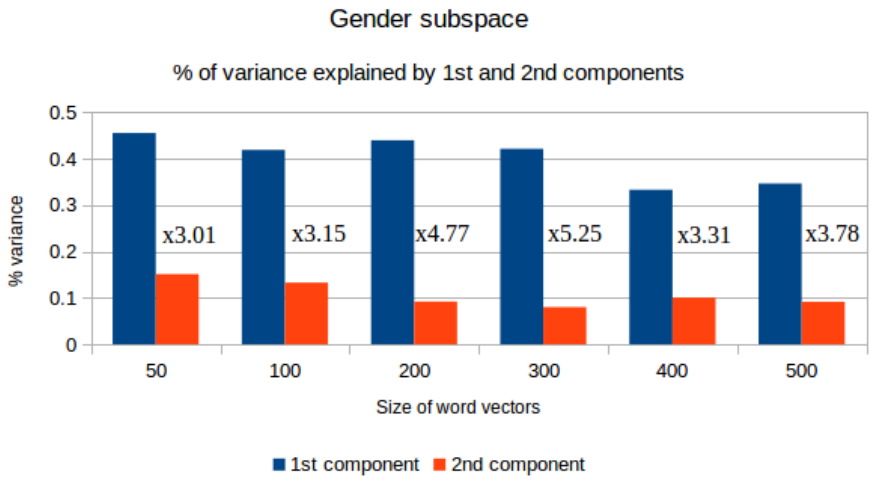

Figure 7: Percentage of variance explained by the first and second component of the gender subspace calculated for word embeddings of different vector lengths

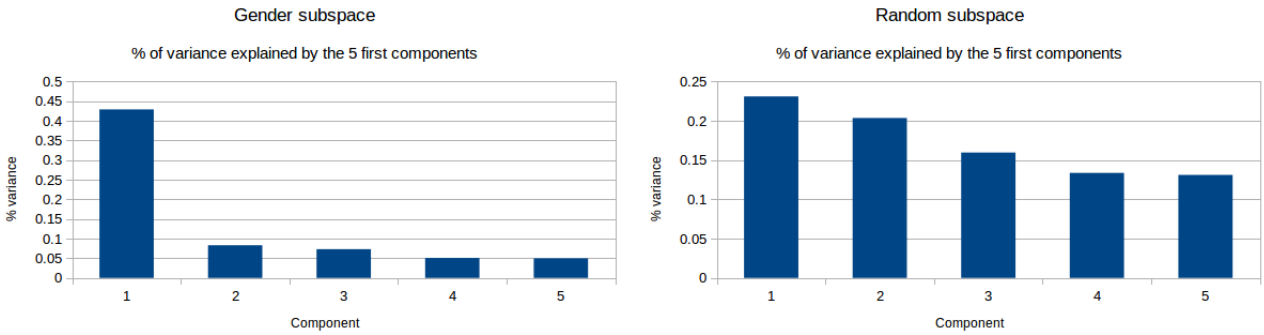

Figure 8: Comparison between the percentage of variance explained by the first 5 components of a subspace created with pairs of gendered words, and a subspace created with random words

Direct bias In [3], the direct bias obtained from occupational words is 8%, which they claim "confirms that many occupation words have substantial component along the gender direction". Because there are no more results available of this measure, it is hard to say which values can be interpreted as the presence of significant bias in the corpus, and I can only compare my result with theirs. Using ungendered words that appear in the lists of top features of both males and females in the classification models described above (shown in Table 15) **direct bias is 11%**. It must be noted how these words are expectedly biased as they appear as good indicators of a certain gender, which explains the *high* value obtained in comparison to [3]. Hence, this should not be taken as a standalone proof of the presence of bias in the corpus, but rather an additional corroboration of the meaningfulness of the results obtained in the language analysis performed.

Ungendered words that should not be associated with a specific gender	
Bolso *(=Handbag)*	Coherente *(=Coherent)*
Feminismo *(=Feminism)*	Feminista *(=Feminist)*
Guardería *(=kindergarden)*	Igualdad *(=Equality)*
Patriarcado *(=Patriarchy)*	Patriarcal *(=Patriarchal)*
Reproducción *(=Reproduction)*	Vulnerable *(=Vulnerable)*

Table 15: Examples of ungendered words that should not be associated with a specific gender, extracted from the results of language analysis by gender

6 Automatic detection of tweets

The language analysis presented above shows the existence of biases in the way politicians express themselves and how citizens refer to them. These biases reflect gender stereotypes present in our society and, specifically for the topic that affects us, in social networks like Twitter. This finding justifies the necessity for a tool that helps identify online messages that perpetuate such stereotypes. This task has usually been tackled as a classification problem of supervised machine learning [4], hence an annotated dataset is needed such as those that we manually labeled, as explained in Section 3. The corpus used consists of **replies from individuals to politicians**, and a subset of it was labeled: 9407 tweets as sexist or not sexist, and 8767 tweets as hostile or not hostile. Unfortunately, the sexist category is so small (only 115 tweets) that the detection of sexist messages is discarded from this work, as the learning task appears very difficult or unfeasible with such data. Hence, the experiments described in this section are for classification of tweets as hostile or not hostile.

As indicated in Table 5, 1961 tweets are hostile versus 6806 that aren't, that is, approximately 77% of the tweets are contained in the majority class (not hostile) making the problem rather unbalanced, which is a concern that must be faced when fitting the models. One possible solution could be to do the fine-tuning of parameters maximizing the AUC, which corresponds with the balanced accuracy, as in the language analysis tasks. That was an appropriate choice when both categories were equally important to assess, as happened with the male and female gender. Instead, this problem focuses on detecting the hostile cases, so the precision and recall of the positive class become useful measures for evaluating the quality of detection. For this reason, I optimize the F1 score in the fine-tuning process of the following experiments, which is the harmonic mean between the Precision score (i.e. how many of the tweets detected as hostile are actually hostile) and the Recall score (i.e. how many of the hostile tweets are actually detected).

The results obtained from previous section corroborate that word unigrams have certain predictive capabilities of the bias present in the corpus, leading to the observation that there are significant differences in how men and women speak and are addressed to. Other feature types were omitted from the language analysis task because of their scarce interpretability, such as ngrams of characters. Yet they could improve the predictive ability of a model, because they naturally handle the variability of colloquial language due to slang or typos, which are especially common in Twitter. Other parameters of importance are the preprocessing strategy used, for which I'm trying four combinations: with or without removing stopwords, and using lemmas or real forms. Naturally, several learning algorithms can be tried, each with their own set of parameters. In these experiments I use the following algorithms and parameters:

- Logistic regression with Lasso (L1) regularization: **L1**.
    - C: regularization parameter, in [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}].
- Logistic regression with Ridge (L2) regularization: **L2**.
    - C: regularization parameter, in [10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}].
- Support Vector Machines with RBF kernel: **SVM**.
    - C: regularization parameter, in [10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}, 10^{6}, 10^{7} 10^{8}].
- Multi Layer Perceptron neural network, with 1 hidden layer: **MLP**.
    - N: number of neurons in the hidden layer, in [100, 200, 500].
    - L: learning rate, controls the step-size in weights update, in [10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}].
    - C: L2 regularization parameter, in [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}].
- Random Forest with 50 trees: **RF**.
    - M: maximum number of features to consider when splitting a leaf, in [sqrt(n_features), 10%n_features, 20%n_features, 50%n_features]
    - L: minimum number of samples required for a node to be a leaf, in [1, 20, 50].

In order to simplify the fine-tuning process and presentation of results, an incremental procedure is performed as follows:

1. I start with the model used for the language analysis task (**W1-LM-NOSW**) with the algorithm that yielded the best result (**L1**): **L1-W1-LM-NOSW**, and try each of the following feature sets: **W1**, **W2**, **W3**, **C3**, **C4**, **C5**.
2. I stick to the best feature set, and with it check the four possible combinations of preprocessing: **LM-NOSW**, **LM-SW**, **FM-NOSW**, **FM-SW**.
3. I stick to the best preprocessing combination, and then check the following learning algorithms, doing grid-search of the parameters indicated above: **L1**, **L2**, **SVM**, **MLP**, **RF**.
4. The best performing algorithm, together with the best configuration obtained so far, is the selected model.

For this analysis, 20% of the tweets are kept for testing (1754 samples) and the rest for training (7013), out of which 10-fold cross validation is performed when fitting the model, and the average of F1 scores in the validation sets is used to determine the best model on each combination. Once the best overall configuration is found, it is evaluated on the test set.

6.1 Experiments and results

The first step of the training strategy consists of 6 models, one for each feature set, keeping fixed the preprocessing as lemmas (**LM**) and removing stopwords (**NOSW**), as well as the learning algorithm (**L1**), whose only parameter C controls the regularization strength, and thus overfitting can be prevented by properly adjusting it. This can be observed in the overfitting curves of Figures 9, 10, 11, 12, 13 and 14. The trend is always the same: a very small C means very high regularization, hence all coefficients tend to 0. As C grows, the model grows in complexity and is able to learn better decision boundaries, which is traduced in higher training and validation scores, up to the point where the model starts overfitting, when the validation score gets stuck and eventually starts decreasing. In my experiments, this point of inflexion always happens around C=10 or C=100, and that becomes the selected model in each case. Table 16 indicates such models for all feature sets and their F1 measures.

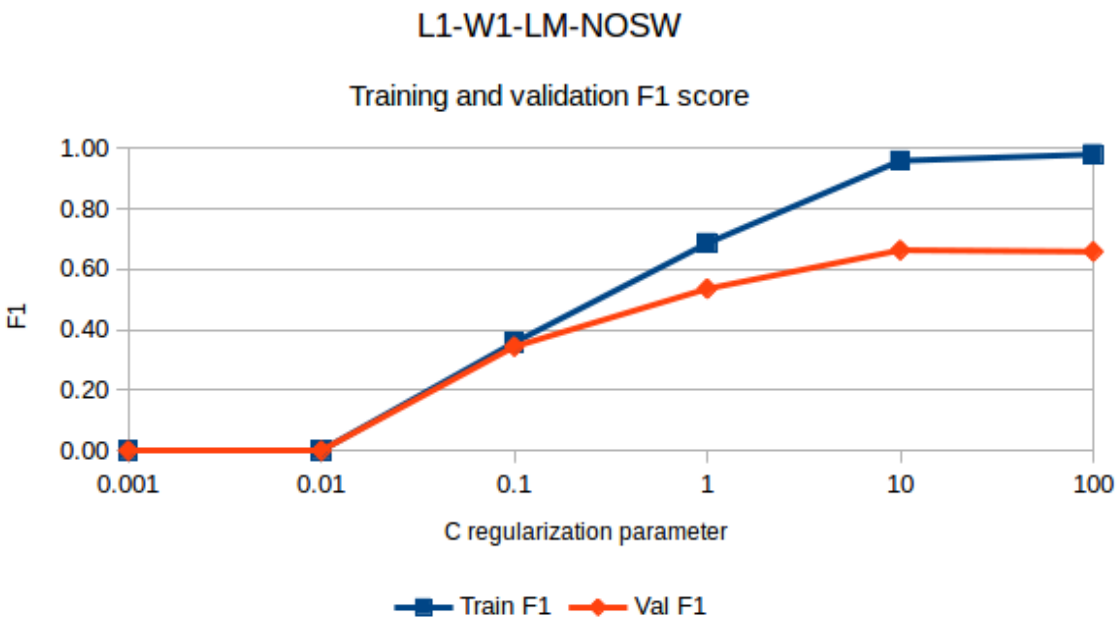

Figure 9: Overfitting curve of the models trained with feature set W1 in the task of automatic detection of hostility, holding fixed the training algorithm as L1 and the preprocessing strategy as LM-NOSW

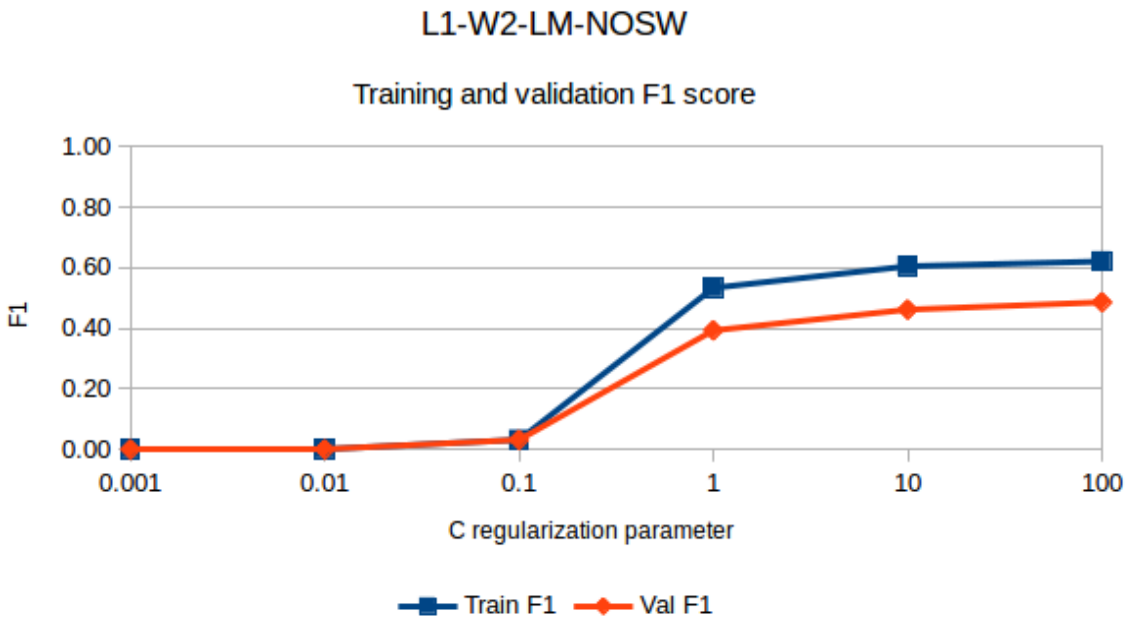

Figure 10: Overfitting curve of the models trained with feature set W2 in the task of automatic detection of hostility, holding fixed the training algorithm as L1 and the preprocessing strategy as LM-NOSW

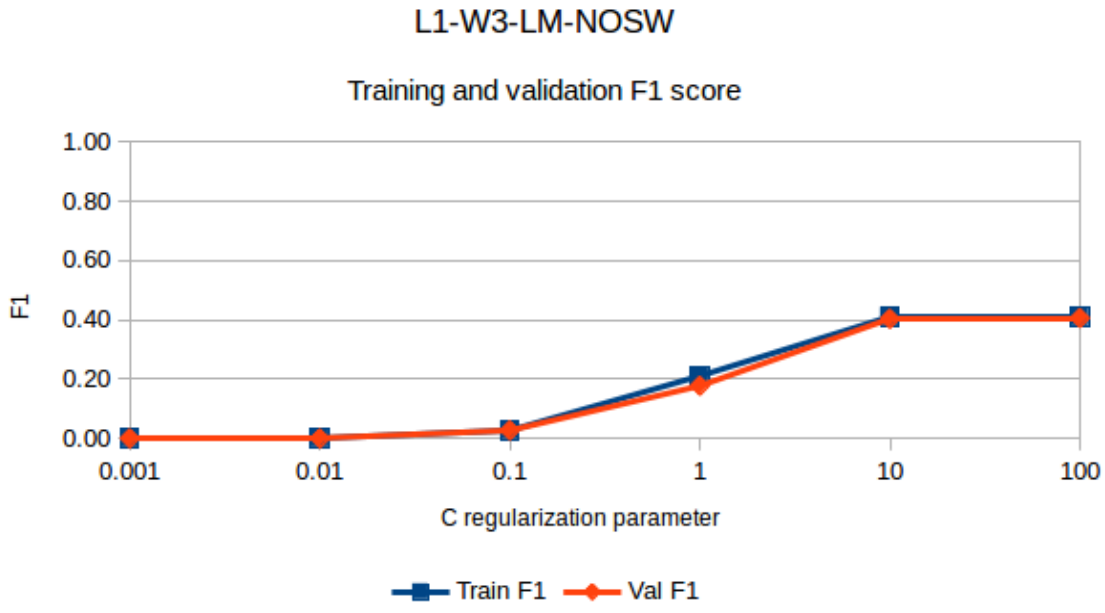

Figure 11: Overfitting curve of the models trained with feature set W3 in the task of automatic detection of hostility, holding fixed the training algorithm as L1 and the preprocessing strategy as LM-NOSW

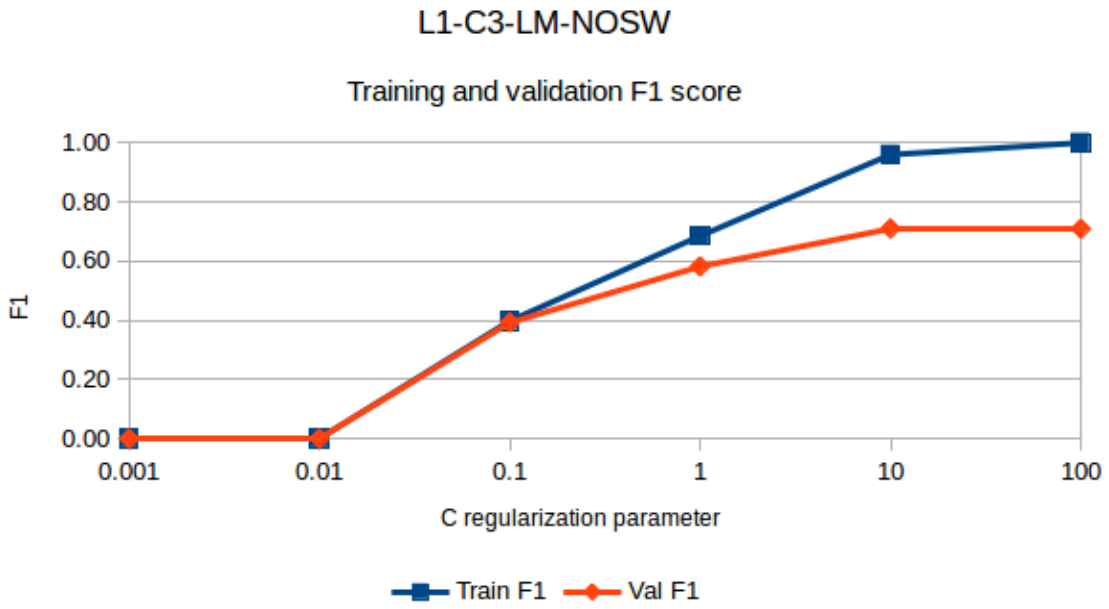

Figure 12: Overfitting curve of the models trained with feature set C3 in the task of automatic detection of hostility, holding fixed the training algorithm as L1 and the preprocessing strategy as LM-NOSW

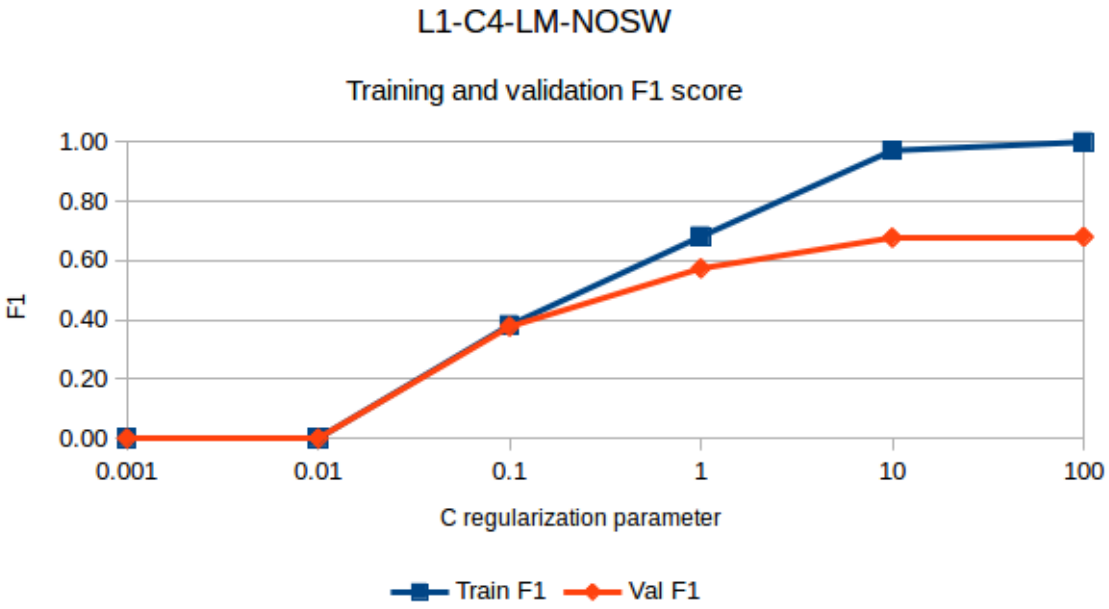

Figure 13: Overfitting curve of the models trained with feature set C4 in the task of automatic detection of hostility, holding fixed the training algorithm as L1 and the preprocessing strategy as LM-NOSW

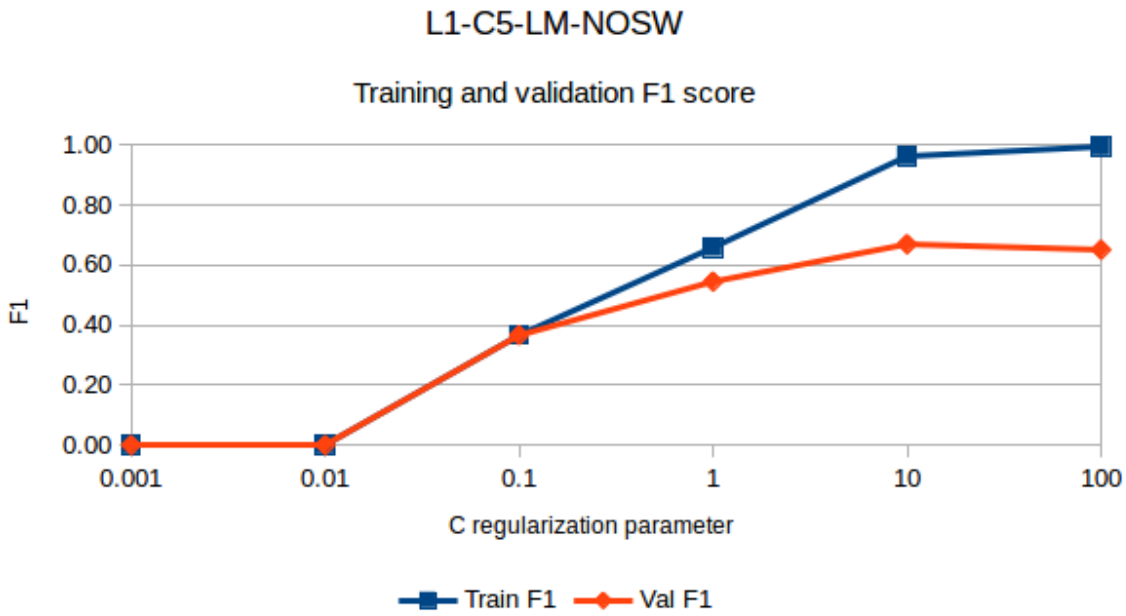

Figure 14: Overfitting curve of the models trained with feature set C5 in the task of automatic detection of hostility, holding fixed the training algorithm as L1 and the preprocessing strategy as LM-NOSW

Best model	*Best C*	*Training F1*	*Validation F1*
*L1-**W1**-LM-NOSW*	10	95.91%	66.31%
*L1-**W2**-LM-NOSW*	100	62.09%	48.62%
*L1-**W3**-LM-NOSW*	100	40.96%	40.55%
*L1-**C3**-LM-NOSW*	10	95.97%	**70.98%**
*L1-**C4**-LM-NOSW*	100	99.93%	68.00%
*L1-**C5**-LM-NOSW*	10	96.27%	66.95%

Table 16: Training and Validation F1 scores for the best model obtained by each feature set in the task of automatic detection of hostility, holding the algorithm fixed at L1 and the preprocessing fixed at LM-NOSW

Comparing the six feature sets available, one observation is clear: all bag-of-character-ngrams work at least as good as the best bag-of-word-ngrams, which is **W1**. Word bigrams and trigrams dramatically decrease the quality of prediction, and character trigrams are the only model reaching 70% F1 on the validation set, which is almost a 5% increase with respect to classic word unigrams. Hence, **L1-C3-LM-NOSW** is the best model so far, and the **C3** feature set is held fixed for all future experiments. I hypothesize that characters are more successful due to the enormous variability of

forms of the same concept that may appear in the corpus, due to the informal tone used and the multilingual corpus. That could be producing too many combinations of word ngrams with small frequencies, making it harder for word ngrams to keep counts of what's really happening in the data. Characters, on the other hand, can gather up these variations as long as there is a common part between them, such as the stem that many words in Spanish, Catalan and Galician may share.

Best model	*Best C*	*Training F1*	*Validation F1*
*L1-C3-**LM-NOSW***	10	95.97%	70.98%
*L1-C3-**LM-SW***	10	96.41%	71.60%
*L1-C3-**FM-NOSW***	100	99.96%	**72.42%**
*L1-C3-**FM-SW***	100	99.96%	71.32%

Table 17: Training and Validation F1 scores for the best model obtained by each preprocessing strategy in the task of automatic detection of hostility, holding the algorithm fixed at L1 and the feature set fixed at C3

Now the different preprocessing combinations are evaluated (Table 17). For space and redundancy reasons, now the overfitting curves aren't reported as their behavior is comparable, so the best configuration for each model is reported. Differences produced by preprocessing are smaller, with the best and worst model having a difference in F1 score of 1.5% approximately. The combination of real forms (**FM**) and removing stopwords (**NOSW**) yields the best mark, with **72.42% F1 on the validation set**, and again it is fixed for the following experiments. The only trend I can find in which attributes of preprocessing are more beneficial is the following: the simplest model (lemmatizing and removing stopwords) and the most complex (keeping all different real forms and keeping stopwords) perform worse than the other two, so the success of a certain preprocessing could depend, at least partially, on how many features it produces, and too many or too few are clearly undesired. Now the learning algorithm is changed, and the best configuration achieved by each algorithm is compared with the others.

L2 Results are expected to be similar to L1, as it is the same algorithm with a different regularization strategy. Indeed, the overfitting curve (Figure 15) behaves similarly to the ones presented above, but more smoothly due to the shrinkage of coefficients that could be making all possible models more similar. Likewise, the validation score achieved is pretty similar to the best results using L1 in the previous runs, but slightly lower, **with an F1 of 70.80%**. The implicit feature selection performed by L1 appears to be beneficial in vector spaces dimensionally large as those produced by ngram models.

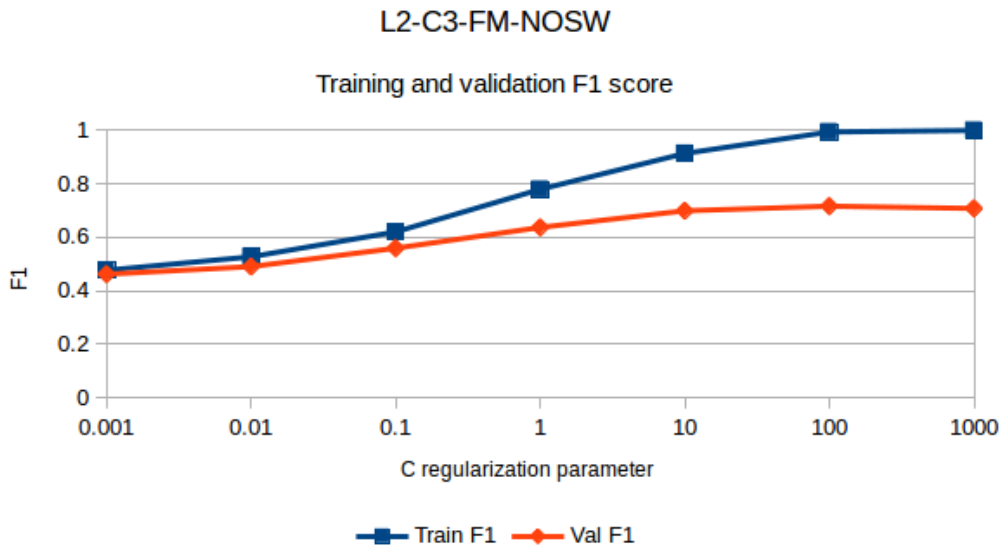

Figure 15: Overfitting curve of the models trained with algorithm L2 in the task of automatic detection of hostility, holding fixed the feature set as C3 and the preprocessing as FM-NOSW

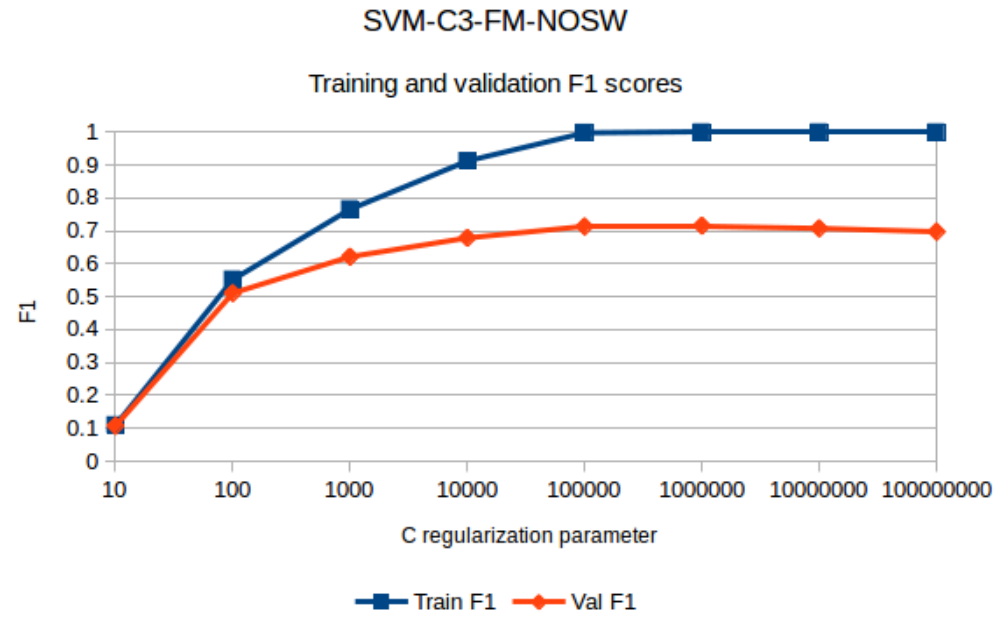

Figure 16: Overfitting curve of the models trained with algorithm SVM in the task of automatic detection of hostility, holding fixed the feature set as C3 and the preprocessing as FM-NOSW

SVM I'm using Support Vector Machines with a non-linear kernel of type RBF, and adjusting its regularization parameter that controls the complexity of the boundary against the amount of misclassified samples. As in the experiments with logistic regressions, there exists a balance between both sources of error that, in this case, results in the best model at C=1,000,000. **The F1 score in validation set is 71.62%**, slightly improving results of L2 but not of L1 by a margin of 1%.

MLP My experiments use a multi-layer perceptron architecture with ReLUs as activation functions and one hidden layer, for which different numbers of neurons are tried (100, 200, 500), as well as several learning rates for weights update in the backpropagation. The weights update is also affected by L2 regularization that controls overfitting, through a parameter alpha that is directly proportional to the amount of regularization (0.0001, 0.001, 0.01, 0.1, 1, 10). A grid search through these three parameters was performed and, in order to facilitate the presentation of results, they are organized by the different learning rates tried (0.001, 0.01, 0.1). More than that resulted in a worse performing model, and a smaller rate had no advantages against 0.001 at the expense of higher training time.

Figures 17, 18 and 19 show area charts of the validation F1 score for a learning rate of 0.001, 0.01 and 0.1 respectively, where the different numbers of neurons are represented each in one colored area, and the different regularizations are evolved in the X axis. They are zoomed in the part where the highest values are reached, so as to facilitate comparison. One can see how higher learning rates lead to more unstable learning, reaching lower scores and with more bounces between models. Besides, the number of neurons is affected by learning rates, with more neurons being better with a slower rate and vice versa. Hence the best results overall are yielded by a learning rate of 0.001 and using 500 neurons, where the area chart is consistently above. The best score is reached with an alpha of 0.001, getting a **validation F1 of 73.82%.**

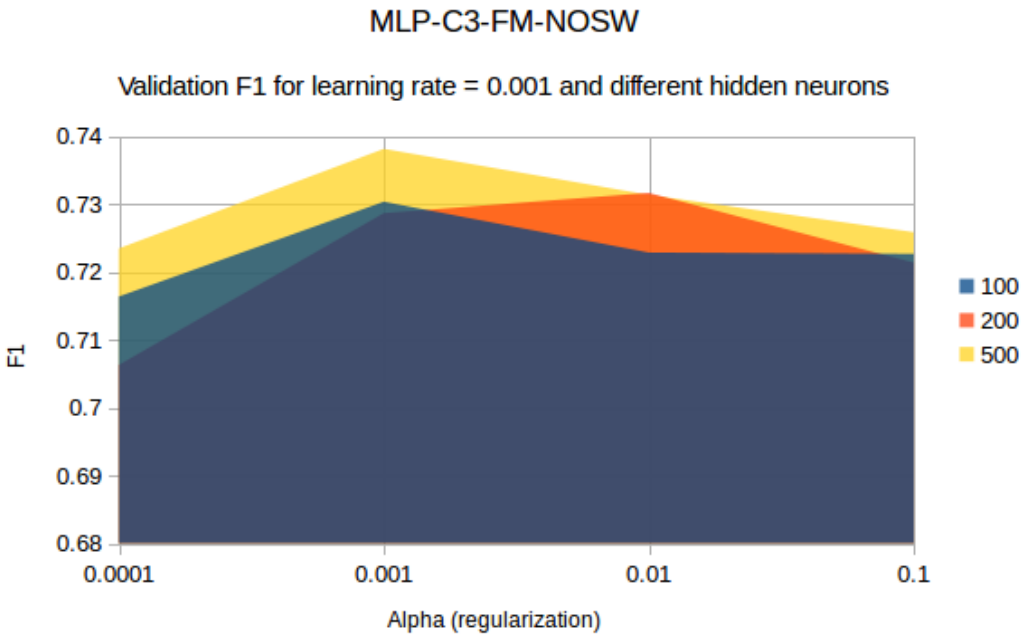

Figure 17: Evolution of the Validation F1 score for different regularization levels of the MLP algorithm, for configurations with different numbers of hidden neurons, keeping the learning rate at 0.001

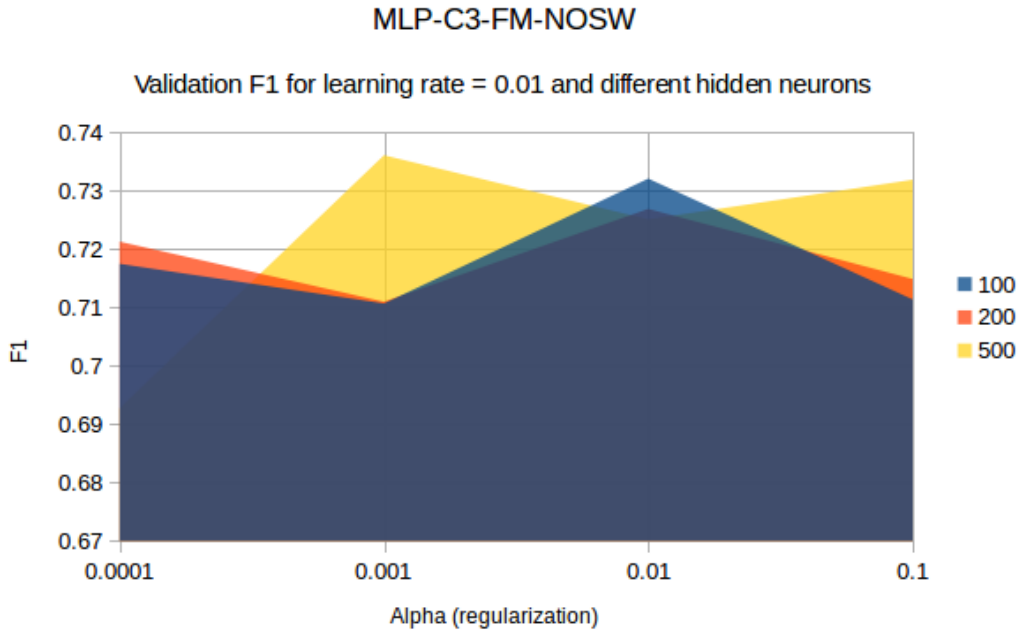

Figure 18: Evolution of the Validation F1 score for different regularization levels of the MLP algorithm, for configurations with different numbers of hidden neurons, keeping the learning rate at 0.01

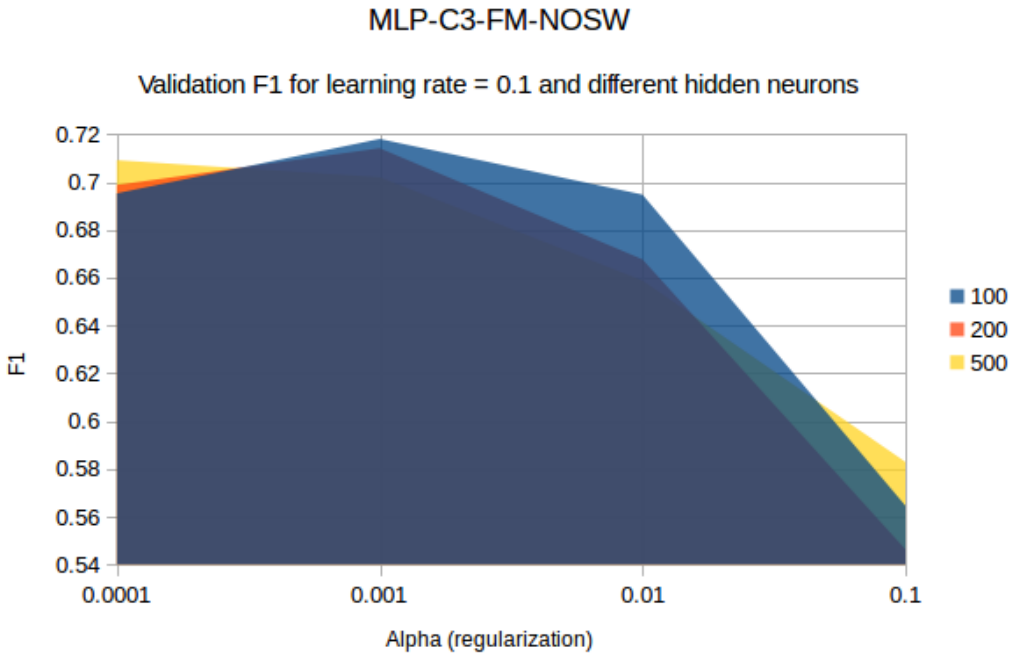

Figure 19: Evolution of the Validation F1 score for different regularization levels of the MLP algorithm, for configurations with different numbers of hidden neurons, keeping the learning rate at 0.1

In order to provide as well results of the different training and validation scores achieved at different levels of regularization, an overfitting curve of the configuration that yields the highest mark is presented in Figure 20. The behavior is as usual, small regularization producing slight overfitting and too much regularization underfitting the model.

RF My experiments use Random Forests of 50 trees with the Gini index as a measure of fitness of the splits. Two parameters are adjusted: the maximum number of features to consider when making a split (**MF**), which affects diversity of each individual tree (a small number will make different trees more diverse) and thus the variance of the model; and the minimum number of samples required for a node to be a leaf (**LEAF**), which affects how much noise is introduced into the model, consequently reducing overfitting. Results show the F1 scores for LEAF = 1, 2, 5 and 10 (more than that drastically reduces performance for this corpus) and MF = square root, 10%, 20%, 50%, 70%, 90% and 100% of the total features (5000), i.e. MF = 70, 500, 1000, 2500, 3500, 4500, 5000. Figure 21 shows an area chart with all F1 scores in the validation set, and Figure 22 is the same chart focused in the area where the maximum is achieved, in order to better observe the results.

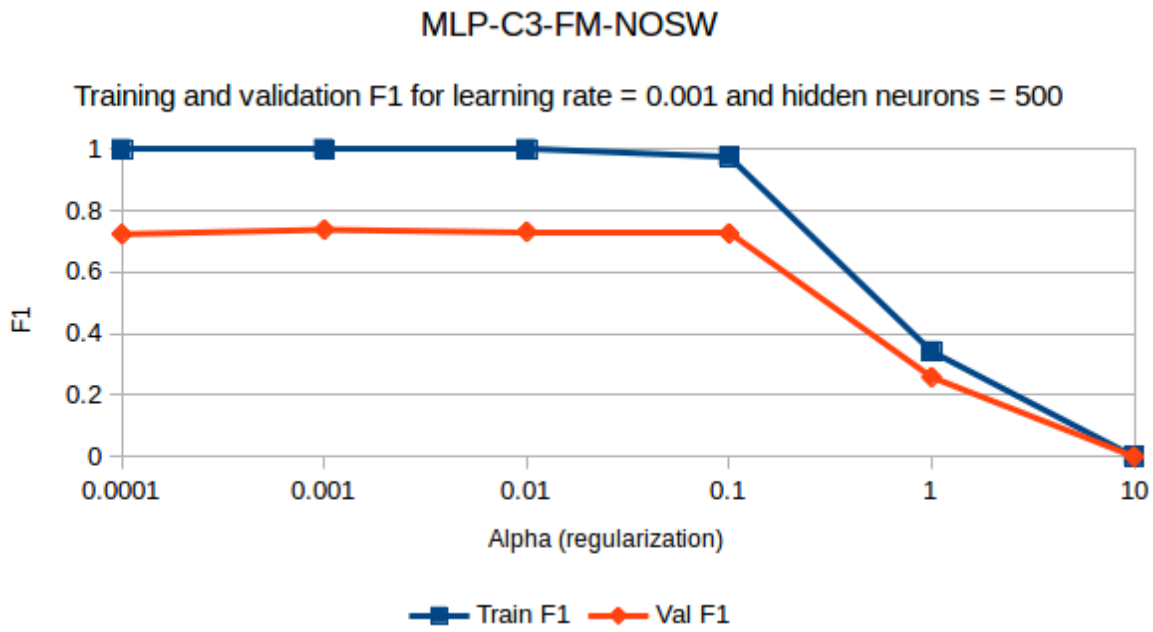

Figure 20: Overfitting curve for the winning configuration of MLP algorithm in terms of learning rate (0.001) and hidden neurons (500)

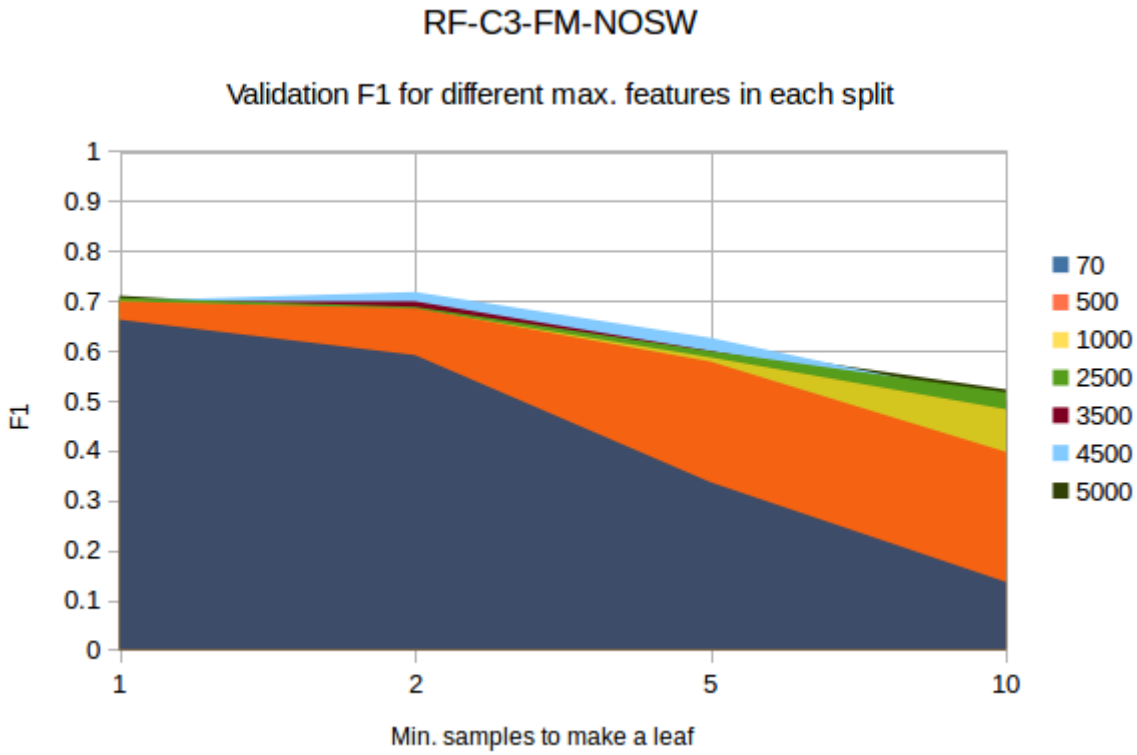

Figure 21: Validation F1 scores for different configurations of RF algorithm

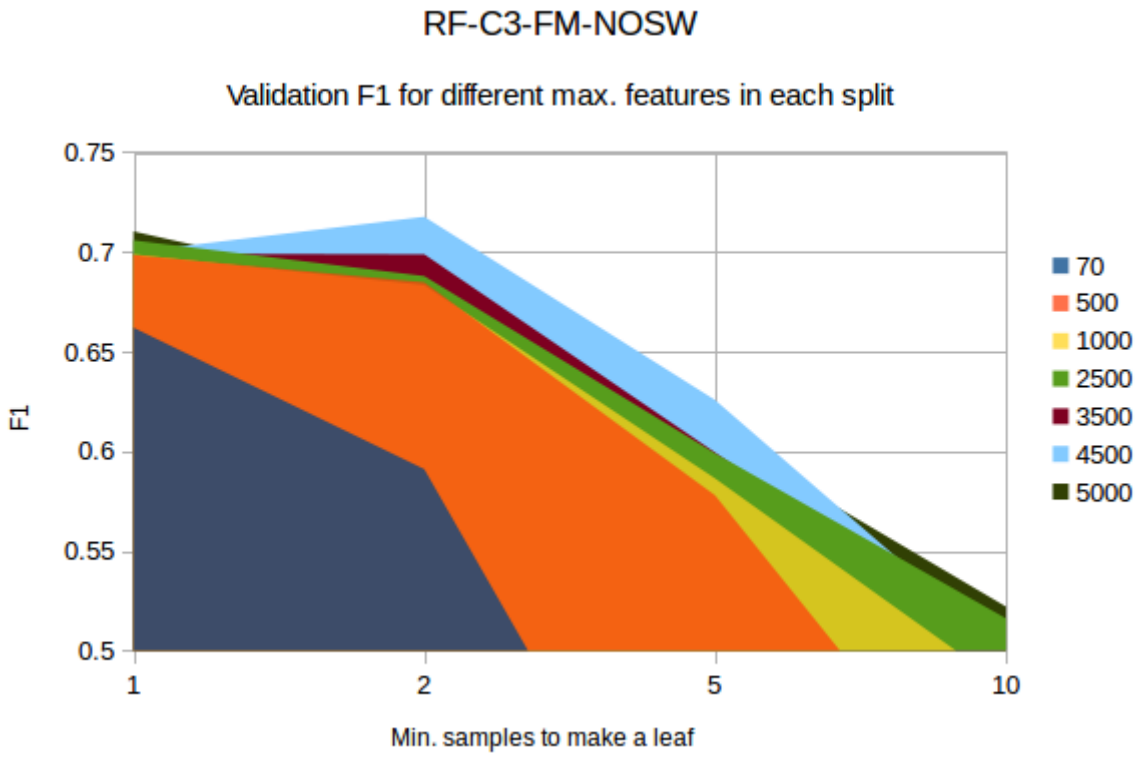

Figure 22: Validation F1 scores for different configurations of RF algorithms, focused in the part where the highest values are achieved

First, the highest results are always found with LEAF = 1 or 2. While this may seem too small, this parameter is certainly related to the size of the dataset, which in this case is rather small. Forbidding leafs with fewer than 10 samples makes the model too simple in terms of classification capability. Regarding MF, results increase as MF increases up to 90% of the features (i.e. 4500 features used), achieving the best result at LEAF=2 and MF=4500, with a **validation F1 of 71.76%,** with a training F1 of 98.97%.

Model	*F1*		*Precision*		*Recall*		*AUC*	
	Train	*Val*	*Train*	*Val*	*Train*	*Val*	*Train*	*Val*
***L1**-C3-FM-NOSW*	99.96%	72.42%	100%	70.83%	99.93%	**74.30%**	99.96%	**82.72%**
***L2**-C3-FM-NOSW*	99.96%	70.80%	98.75%	69.96%	100%	73.52%	99.82%	82.30%
***SVM**-C3-FM-NOSW*	100%	71.62%	100%	71.33%	100%	71.98%	100%	81.83%
***MLP**-C3-FM-NOSW*	99.98%	**73.82%**	99.98%	78.32%	99.98%	69.91%	99.99%	82.19%
***RF**-C3-FM-NOSW*	98.97%	71.76%	99.95%	**86.47%**	98.01%	61.39%	99.00%	79.32%

Table 18: Training and Validation F1, Precision, Recall and AUC scores for the best model achieved by each training algorithm, holding the feature set fixed at C3 and the preprocessing fixed at FM-NOSW

Discussion Table 18 summarizes the F1, Precision, Recall and AUC scores achieved by each algorithm. The F1 that I use to evaluate the best model varies only slightly, with the difference between the worst method in validation (L2 logistic regression) and the best (MLP) being 3%. Whilst all models achieve to (almost) perfectly learn the training set, it appears that their generalization capabilities are capped below 75% F1, even when some sort of regularization is applied. What's more, such regularization parameters appear not to have a major effect in reducing overfitting, but rather they end up underfitting the model even with values that are usually considered *default* (see how C evolves in SVM). My hypothesis is that **the dataset is especially noisy**, due to its small size and the enormous set of possibilities to construct a hostile tweet, and hence **the models suffer from a high variance** that cannot be significantly reduced by just regularizing the decision boundary; instead, the only apparent solution I can find is **to increase the sample size** by labeling more tweets, and then observe if the same models are able to generalize better. Only when such limitation is fixed may the choice of an algorithm yield more significant differences in out-of-sample performance.

In Table 19 I show the training, validation and test metrics of the winning configuration. **The F1 score in the test set is 73.98%,** which comes from a Precision of 77.74% and Recall of 70.57%. Both measures are connected by a trade-off that, in this case, is won by Precision, hence the model fails more due to false negatives than to false positives, i.e. it is conservative in affirming that a certain tweet is hostile in order to make more right guesses. Hypothesizing with a real-world

application of this detector, this behavior might be desired or the contrary. For instance, if all the detections were to be automatically deleted from the social network, then this focus on precision would be highly beneficial because fewer inoffensive tweets would be mistakenly banned. On the other hand, if the tweets detected were then analyzed by humans who judged their content, then false positives would be admissible if they allow for more hostile tweets to be retrieved. This can be decided by choosing another algorithm of Table 18, one with higher values for Precision at the expense of Recall, or vice versa; or modifying the evaluation measure: instead of using the F1 score, use an F-beta score that weighs Precision and Recall differently. Regarding the test AUC, the average accuracy between both classes, hostile and not, is 80%.

MLP-C3-FM-NOSW	*Training*	*Validation*	*Test*
F1	99.98%	73.82%	73.98%
Precision	99.98%	78.32%	77.74%
Recall	99.98%	69.91%	70.57%
AUC	99.99%	82.19%	82.29%

Table 19: Training, Validation and Test F1, Precision, Recall and AUC scores for the best model overall, which uses MLP as training strategy, C3 as feature set, and FM-NOSW as preprocessing

6.2 Error analysis

I performed a qualitative analysis of the errors that happen in the test set, both the false positives and false negatives, which also can be seen in the confusion matrix of Table 20. My analysis summarizes what traits in a tweet provoke that the model fails, supported with examples of the dataset. Regarding false positives, the main source of error comes from tweets that contain hostile speech that is not addressed at the politician who receives the message. The actual receiver of hostility usually is a third person that is mentioned in the original tweet. Because all features used are lexical, the model only sees words that are associated with hostility and cannot discern who they refer to. Some examples of the test set are:

- "*@amonterosoler @Irene_Montero_ Pedro Sánchez cada vez más patético",* where the word *pathetic* is used but addresses *Pedro Sánchez* and not *@amonterosoler*, who is the actual receiver.
- "*@ignacioaguado No tienen vergüenza , en todos los viejos partidos más de lo mismo . La ciudadanía espera mucho de vosotros , no hagáis la AVESTRUZ si veis indicios en compañeros . EDIFICANTE"*. Only the parts underlined use a tone that is clearly positive for the receiver. The rest is expressed in an angry tone that both criticizes other people and gives orders to the actual receiver, but still the tone is overall positive towards the receiver.

Regarding false negatives, one reason for errors is the same as for false positives, but the other way around: tweets containing words with a positive tone even though the general tone is hostile against the receiver. Such positive words may appear because they refer to a third person, but also because they are used with irony. This is actually the most common reason for the false negatives obtained. Some examples are:

- "*@nicadichiara ¡ Qué contento estoy ! somos los líderes en España en paro , que no decaiga ¡ no podemos ser los segundos ! así que sigamos con nuestras políticas reforcemos nuestra posición , tenemos que ser líderes ¡ Vamos , hombre ! que nos quieren quitar el primer puesto y no lo vamos a permitir ."*. The whole reply has an ironic tone, with sentences like *"I'm so happy we're leaders in unemployment in Spain"*. There is abundance of words that are positive when observed in isolation.
- "*@GFVara Oye , he visto en prensa en enchufas en cargos publicos a tus colegas . Tienes algo por ahi para mi ? Podriamos ser amigos , si fuese necesario"*. Again, irony is used but in a more subtle sense, by asking a politician for favorable job positions and asking for his friendship in case that's necessary.
- "*@ccifuentes A pesar de ti cumplen xq son profesionales responsables . Asi , pues no t apuntes el tanto . . ya t conocemos falsa arpia ."*. Some of the words appearing are clearly

associated with successful businesses, namely "*responsible professionals"*, yet the individual insists, even with offenses, that this happens *despite* the presence of the politician who receives the message.

In summary, **the two main causes for error are irony and unawareness of the addressee**. In order to solve this unawareness, first I suggest that the named entities of the text should be identified. Then, a method for co-reference resolution in dialogues should be used in the pairs of original message and reply, so the model knows about the existence of a third person mentioned by both. Besides, the scope of the text that refers to each entity could be estimated. If successfully identified, the model could account only for the features appearing in that scope, ignoring those words that lead it to error. Resolving irony appears as a harder problem, with a whole line of research in NLP related to this problem [22] [23]. If a detector of irony that works successfully with this corpus is achievable, then it could be used to invert the prediction of the model. In any case, none of these problems are easy because they are inherent to the communicative act, even with humans: if we maintain a conversation where several names appear, it may be hard to make out which stories correspond to whom; if someone starts talking with us for the first time, we might not be used to their use of sarcasm and not catch it at all.

Confusion matrix	*Real*		
Predicted	*Is hostile?*	*Yes*	*No*
	Yes	TP=1297	FP=74
	No	FN=93	TN=290

Table 20: Confusion matrix of the test set in the best model achieved, which is MLP-C3-FM-NOSW

7 Conclusion

The application of text classification techniques to social data was successful, achieving meaningful observations about the differences in how politicians speak, or how they are addressed, depending on their gender. Male and female politicians have preferences about the topics they deal with. Territorial politics and infrastructures are men topics, who also speak more often in terms of ideologies, such as communism, liberalism, left and right. On the other hand, gender issues are monopolized by women, as well as social affairs. Their language is more charged emotionally, both in words and emojis used. The existence of differences of such significance corroborates the prevalence of gender stereotypes, even among our political representatives. They are public figures with huge visibility in the media, so they could be reinforcing these stereotypes on a wide audience. Nevertheless, this is not entirely their fault: territory and gender are, respectively, two topics strongly associated with the language used by the people when addressing to men and women politicians. There are also differences in the qualificatives used to refer to each gender, with women being addressed in a more condescending tone and men being insulted more harshly. In summary, politicians are both victims and transmitters of gender biases in our society.

Having corroborated the presence of such clear stereotypes in political conversation in social media, there appears the need for an automatic solution that helps identify negative attitudes in communication. My approach to hostility achieves a satisfactory 73.98% F1 score on unseen data, and I discussed how the main source of improvement should be an enlargement of the supervised dataset, due to the enormous variety of cases that could make a reply in Twitter hostile. As such variety is underrepresented in my small supervised dataset, I expect it to be rather noisy, and the different algorithms tried appear to overfit to such noise even after thorough fine-tuning. While an increase of data will surely help learn new cases of hostility that are being overlooked by the model obtained, my analysis of errors shows some systematic mistakes: one is the unawareness of the model about the different named entities mentioned in a conversation, so it cannot discern whether the focus of hostility is the politician in question or anyone else; the other is its incapacity to detect irony, making it easy to be deceived by sarcastic words. The main lesson learned is that a more complex language model is needed, that works at more levels than lexical and offers a richer perspective of the text and its representation.

7.1 Future work

The interdisciplinary nature of this thesis leads to several directions of work. As a data-driven solution, it will benefit from enlarging the corpus by downloading more tweets, but the real progress will come from increasing the number of labeled data, improving accuracy of prediction of hostile

tweets and allowing for the desired prediction of sexist tweets. From the computational point-of-view, one could start considering deep learning algorithms for prediction when the supervised dataset is large, namely recursive or convolutional neural networks whose utility in text classification has been proved [24] [25]. Similarly, the embedding vectors calculated could be used as features, under the hypothesis that a vector space that scores a high direct bias should identify sexist tweets better. Improvements in prediction could also come from the use of non-textual features, for instance about characteristics of the Twitter users: is their account verified? How many followers do they have? These are indicators of the exposition of the users, which could be related to the roughness of their public statements.

From the socioscientific view, the generation of the annotated dataset allows to answer plenty of questions. One could analyze if sexism is spread uniformly across all regions in Spain, by looking at the messages from and to autonomic politicians conditioned on their autonomous community. Another concern is whether the effects of hostility on politicians change with their gender, because there's evidence that online harassment to women and minorities can reduce their participation [26]. The relation between anonimity and harassment could be analyzed too, due to the presence of Twitter users without a name to infer their gender. These users could be more conflictive because their real identity is harder to discern. It can be seen how many questions can be addressed as a follow-up to this work with real implications in our understanding of society and politics.

7.2 Final thoughts

The results of this thesis condensate negative attitudes against men and women that happen in the communicative act with politicians. In the era of information, when we're constantly engaged to the news through social media, we should start to consider the quality of the speech we're consuming. Otherwise we risk generalizing a vulgar conception of political debate, or even acquiring it ourselves. It is in our hands to be more demanding with our sources of information, also in everyday use of social networks like Twitter, because that's where politicians get closer to the common people today; and in the end it's us, the common people, who decide the future of our societies in the ballots every few years.

Tweets, despite being strictly limited to 280 characters, can hold impressive amounts of concepts and reveal intricate meanings in a few words. That's why this problem is challenging, as it attempts to convert the communicative act between the subject and the ruler into counts of words, when this act is endless. In a nutshell of the size of a tweet: Tweets will be as puzzling as the issues discussed by their writers, politics will be as puzzling as the people are, and the people will be as puzzling as the concepts they are able to express.

References

[1]: K. Munger. Don't @ Me: Experimentally Reducing Partisan Incivility on Twitter, 2017.

[2]: Y. Theocharis, W. Lowe. Does Facebook increase political participation? Evidence from a field experiment, 2015.

[3]: T. Bolukbasi, K. Chang, J. Zou, V. Saligrama, A. Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, 2016.

[4]: B. Pang, L. Lee, S. Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques, 2002.

[5]: D. Jurafski, J. Martin. Speech and Language Processing, 2017.

[6]: E. Tellez, S. Miranda-Jiménez, M. Graff, D. Moctezuma, O. Siordia, E. Villaseñor. A case study for Spanish text transformations for twitter sentiment analysis, 2017.

[7]: P. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, 2002.

[8]: J. Bohren, A. Imas, M. Rosenberg. The Dynamics of Discrimination: Theory and Evidence, 2017.

[9]: A. Wu. Gender Stereotyping in Academia: Evidence from Economics Job Market Rumors Forum, 2017.

[10]: L. Kaye. Why psychology needs to start taking emoji seriously, 2018.

[11]: H. Wallach. Computational Social Science ≠ Computer Science + Social Data, 2018.

[12]: B. O'Connor. Statistical Text Analysis for Social Science, 2014.

[13]: J. Chen, S. Yan, K. Wong. Verbal aggression detection on Twitter comments: convolutional neural network for short-text sentiment analysis, 2018.

[14]: E. Wulczyn, N. Thain, L. Dixon. Ex Machina: Personal Attacks Seen at Scale, 2016.

[15]: A. Dhrodia. Unsocial Media: Tracking Twitter Abuse against Women MPs, 2017.

[16]: D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, B. Qin. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification, 2014.

[17]: N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, N. Bhamidipati. Hate Speech Detection with Comment Embeddings, 2015.

[18]: M. Marchetti-Bowick, N. Chambers. Learning for Microblogs with Distant Supervision: Political Forecasting with Twitter, 2012.

[19]: C. Johnson, P. Shukla, S. Shukla. On Classifying the Political Sentiment of Tweets, 2018.

[20]: L. Melkumova, S. Shatskikh. Comparing Ridge and LASSO estimators for data analysis, 2017.

[21]: T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representation in Vector Space, 2013.

[22]: R. González-Ibáñez, S. Muresan, N. Wacholder. Identifying Sarcasm in Twitter: A Closer Look, 2011.

[23]: L. Weitzel, R. Prati, R. Aguiar. The Comprehension of Figurative Language: What is the influence of Irony and Sarcasm on NLP Techniques?, 2016.

[24]: B. Gambäck, U. Sikdar. Using Convolutional Neural Networks to Classify Hate-Speech, 2017.

[25]: P. Badjatiya, S. Gupta, M. Gupta, V. Varma. Deep Learning for Hate Speech Detection in Tweets, 2017.

[26]: B. Henson, B. Reyns, B. Fisher. Fear of Crime Online? Examining the Effect of Risk, Previous Victimization, and Exposure on Fear of Online Interpersonal Victimization, 2013.

Appendix I: top features indicative of author gender

L1 Male features				L1 Female features			
#	Word	Coef	Odds	#	Word	Coef	Odds
1	almansa	3.86	47.27	1	yours	3.68	39.68
2	bns	3.76	42.93	2	compromis	3	20.03
3	haiku	3.74	41.93	3	fuengirola	2.86	17.46
4	retweeted	3.61	36.85	4	📢	2.72	15.12
5	tabarnia	3.16	23.47	5	maltratador	2.55	12.79
6	urgell	3.14	23.17	6	masclisme	2.2	9.03
7	radial	3.13	22.87	7	estereotipo	2.18	8.81
8	barkos	3.11	22.34	8	❤	2.07	7.96
9	malpartida_de_cáceres	3.04	20.98	9	lle	2.06	7.85
10	auditorio	2.95	19.06	10	masclista	1.95	7.03
11	comunista	2.52	12.49	11	forta	1.91	6.75
12	haría	2.29	9.85	12	junta_de_castilla	1.73	5.64
13	tarragona	2.23	9.31	13	ceuta	1.63	5.09
14	rècord	2.21	9.08	14	♀	1.54	4.68
15	boadilla	2.17	8.78	15	doloroso	1.54	4.65
16	insular	2.16	8.65	16	javier_fernández	1.53	4.6
17	l'economia	2.04	7.73	17	principado	1.52	4.59
18	trasvase	2.03	7.64	18	compostela	1.52	4.55
19	tav	2.02	7.58	19	gitano	1.46	4.32
20	pablo	1.96	7.07	20	feminismo	1.44	4.21
21	variante	1.95	7.02	21	saneamiento	1.43	4.16
22	leonés	1.91	6.78	22	som-hi	1.42	4.15
23	constructivo	1.91	6.74	23	hallar	1.41	4.09
24	🦁	1.9	6.69	24	imos	1.39	4.03
25	l'independentisme	1.89	6.63	25	reproducción	1.39	4
26	♂	1.84	6.32	26	🍇	1.38	3.99
27	inversions	1.84	6.32	27	parlamento_de_canarias	1.38	3.98
28	certificar	1.75	5.74	28	ejército	1.38	3.96
29	cáceres	1.74	5.71	29	🌹	1.37	3.93
30	kilo	1.73	5.67	30	cuota	1.35	3.86
31	vots	1.72	5.59	31	😱	1.33	3.8
32	sedición	1.67	5.32	32	carroza	1.33	3.78
33	cat	1.66	5.28	33	▶	1.3	3.68
34	avería	1.63	5.11	34	😊	1.29	3.62
35	abonar	1.63	5.1	35	consejera	1.28	3.61
36	paisano	1.61	5.02	36	💗	1.28	3.61
37	gobierno_de_aragón	1.58	4.87	37	completamente	1.28	3.59
38	electrónico	1.55	4.73	38	retributivo	1.28	3.58
39	joder	1.54	4.67	39	asistencial	1.27	3.56
40	existente	1.54	4.66	40	😘	1.27	3.56
41	canal_de_isabel_ii	1.54	4.65	41	pensionistes	1.26	3.54
42	isabel	1.53	4.64	42	saúde	1.25	3.5
43	etarra	1.53	4.63	43	📲	1.24	3.47

44	renta_básica	1.49	4.45	44	🏙	1.24	3.45
45	fcse	1.49	4.44	45	bng	1.23	3.43
46	📃	1.48	4.4	46	acento	1.23	3.42
47	pescador	1.47	4.35	47	nunha	1.23	3.41
48	hores	1.47	4.35	48	page	1.22	3.39
49	📸	1.45	4.27	49	consumir	1.22	3.38
50	riojano	1.45	4.26	50	pasito	1.22	3.37
51	idioma	1.44	4.21	51	gobierno_de_navarra	1.2	3.31
52	clm	1.43	4.19	52	pacto_de_estado	1.19	3.3
53	cuenca	1.42	4.14	53	redar	1.19	3.28
54	latín	1.42	4.14	54	feminista	1.17	3.22
55	manipulable	1.42	4.13	55	✔	1.17	3.22
56	retornar	1.4	4.04	56	🎄	1.17	3.21
57	toni	1.4	4.04	57	fuerzas_armadas	1.16	3.2
58	pista	1.39	4.03	58	litoral	1.16	3.2
59	😁	1.39	4	59	posicionar	1.16	3.19
60	romano	1.38	3.98	60	decilitro	1.16	3.19
61	🗓	1.38	3.97	61	maior	1.16	3.18
62	dictar	1.32	3.75	62	malagueño	1.15	3.17
63	consumidor	1.32	3.75	63	masculino	1.14	3.13
64	región_de_murcia	1.32	3.73	64	treballem	1.14	3.13
65	🐿	1.32	3.73	65	sexual	1.14	3.12
66	obispo	1.31	3.71	66	gobierno_de_cifuentes	1.13	3.11
67	millora	1.3	3.68	67	herri	1.13	3.1
68	febrer	1.3	3.67	68	reserva	1.12	3.07
69	fotografía	1.3	3.66	69	ourense	1.11	3.05
70	de_sentido_común	1.3	3.65	70	tantes	1.11	3.02
71	medalla	1.29	3.65	71	prision	1.1	3.02
72	ribera	1.28	3.6	72	egiten	1.1	3
73	occidente	1.27	3.56	73	lingua	1.09	2.99
74	carril	1.25	3.48	74	segregar	1.09	2.98
75	complemento	1.25	3.48	75	trobem	1.09	2.96
76	exportar	1.24	3.47	76	banquillo	1.08	2.95
77	indústria	1.24	3.44	77	😍	1.08	2.94
78	qual	1.23	3.42	78	alde	1.08	2.94
79	negativa	1.23	3.41	79	illes_balears	1.06	2.89
80	aconseguit	1.22	3.39	80	lluitar	1.06	2.88
81	cospedal	1.21	3.36	81	criminalizar	1.05	2.86
82	bilateral	1.21	3.36	82	galiza	1.04	2.84
83	cartagena	1.2	3.34	83	haurien	1.04	2.83
84	post	1.2	3.33	84	persoal	1.04	2.83
85	patio	1.19	3.3	85	relat	1.04	2.82
86	burgos	1.19	3.29	86	degradar	1.03	2.81
87	aeropuerto	1.19	3.28	87	😣	1.02	2.77
88	lleida	1.18	3.24	88	zorionak	1.01	2.75
89	contrari	1.17	3.22	89	mujer	1.01	2.74
90	león	1.16	3.19	90	machismo	1.01	2.73
91	teoría	1.16	3.18	91	dijous	0.99	2.69

92	novela	1.15	3.15	92	patriarcado	0.99	2.68
93	eólico	1.14	3.13	93	testigo	0.98	2.66
94	san_esteban	1.14	3.12	94	silenciar	0.97	2.65
95	patrimonial	1.14	3.11	95	apoio	0.97	2.64
96	upn	1.13	3.1	96	arena	0.97	2.64
97	monumento	1.13	3.09	97	junta_de_andalucía	0.97	2.63
98	llegir	1.12	3.08	98	cohesió	0.97	2.63
99	separatismo	1.12	3.07	99	prostitución	0.96	2.61
100	exclusión	1.12	3.07	100	don	0.95	2.6

L2 Male features				L2 Female features			
#	Words	Coef	Odds	#	Words	Coef	Odds
1	♂	1.17	3.21	1	compromis	0.74	2.1
2	tarragona	1.1	3.02	2	📢	0.71	2.03
3	bns	1.06	2.89	3	😊	0.49	1.64
4	almansa	1.06	2.88	4	❤	0.48	1.61
5	retweeted	1	2.73	5	♀	0.45	1.57
6	leonés	0.94	2.56	6	😘	0.44	1.56
7	tabarnia	0.94	2.56	7	feminismo	0.41	1.51
8	comunista	0.9	2.46	8	🎹	0.41	1.51
9	cáceres	0.84	2.33	9	mujer	0.4	1.5
10	🐭	0.84	2.32	10	🌹	0.39	1.47
11	malpartida_de_cáceres	0.81	2.26	11	🤍	0.38	1.46
12	base	0.8	2.22	12	▶	0.36	1.43
13	and	0.76	2.14	13	fuengirola	0.35	1.42
14	cuenca	0.76	2.13	14	🎄	0.34	1.4
15	🦁	0.75	2.12	15	sexual	0.34	1.4
16	canario	0.75	2.12	16	inhumano	0.33	1.39
17	📻	0.74	2.11	17	feminista	0.33	1.39
18	hores	0.74	2.1	18	😍	0.33	1.39
19	riojano	0.74	2.09	19	ceuta	0.32	1.38
20	león	0.73	2.08	20	🏐	0.32	1.38
21	incapaz	0.73	2.06	21	machismo	0.31	1.37
22	haiku	0.71	2.03	22	zorionak	0.31	1.36
23	pista	0.7	2.01	23	maltratador	0.31	1.36
24	radial	0.7	2.01	24	presidenta	0.31	1.36
25	región_de_murcia	0.69	2	25	😣	0.3	1.36
26	editorial	0.69	1.99	26	😱	0.3	1.35
27	😁	0.67	1.96	27	principado	0.29	1.34
28	tramo	0.67	1.96	28	😚	0.29	1.34
29	upn	0.67	1.96	29	pacto_de_estado	0.28	1.32
30	the	0.67	1.95	30	galiza	0.28	1.32
31	nacionalista	0.67	1.95	31	completamente	0.27	1.31
32	concierto	0.66	1.93	32	🔝	0.27	1.31
33	clm	0.65	1.92	33	📲	0.26	1.3
34	dictar	0.64	1.9	34	masclisme	0.26	1.3
35	ses	0.64	1.9	35	junta_de_andalucía	0.26	1.3
36	urgell	0.64	1.9	36	acento	0.26	1.29

37	novela	0.64	1.9	37	❤	0.25	1.29
38	detener	0.62	1.87	38	machista	0.25	1.29
39	boadilla	0.62	1.86	39	don	0.25	1.29
40		0.62	1.86	40	estereotipo	0.25	1.29
41	exclusión	0.61	1.85	41	precioso	0.25	1.29
42	murcia	0.61	1.85	42		0.25	1.28
43	liberal	0.6	1.81	43	drets	0.25	1.28
44	derecha	0.59	1.81	44	junta_de_castilla	0.25	1.28
45	barkos	0.59	1.81	45	✔	0.25	1.28
46		0.59	1.81	46	parlamento_de_canarias	0.25	1.28
47	trasvase	0.58	1.79	47	saneamiento	0.24	1.27
48	teruel	0.58	1.78	48	avalar	0.24	1.27
49	idioma	0.58	1.78	49	harto	0.24	1.27
50	militante	0.57	1.77	50	ejército	0.24	1.27
51	romano	0.57	1.77	51	contentar	0.24	1.27
52		0.56	1.76	52	forta	0.24	1.27
53	inversions	0.56	1.76	53	illes_balears	0.24	1.27
54	cert	0.56	1.75	54	invisible	0.23	1.26
55	concejal	0.55	1.74	55	doncs	0.23	1.26
56	cospedal	0.55	1.74	56	masclista	0.23	1.26
57	cat	0.55	1.73	57	prevención	0.23	1.26
58	millora	0.54	1.72	58	vigo	0.23	1.26
59	ganador	0.54	1.72	59	pensionistes	0.23	1.26
60	inauguración	0.54	1.71	60	andalucía	0.23	1.25
61	referir	0.54	1.71	61	som-hi	0.22	1.25
62	unanimidad	0.54	1.71	62	gallego	0.22	1.25
63	vino	0.54	1.71	63	anular	0.22	1.25
64		0.54	1.71	64	vanguardia	0.22	1.25
65	humildad	0.53	1.7	65	imos	0.22	1.25
66	pnv	0.53	1.7	66	almería	0.22	1.25
67	dudar	0.53	1.7	67	bravo	0.22	1.24
68	bipartidismo	0.53	1.7	68		0.22	1.24
69	detalle	0.52	1.69	69	reproducción	0.22	1.24
70	estable	0.52	1.68	70	lgtbi	0.22	1.24
71	pib	0.52	1.68	71	madrileño	0.21	1.24
72	manifiesto	0.52	1.68	72	piel	0.21	1.24
73	homenajear	0.52	1.68	73	cohesió	0.21	1.24
74	lleida	0.51	1.67	74	comité	0.21	1.24
75	gobierno_de_canarias	0.51	1.67	75	apoio	0.21	1.24
76	eleccions	0.51	1.67	76	d.e._p	0.21	1.24
77	joder	0.51	1.66	77	compostela	0.21	1.24
78	consejería	0.51	1.66	78	lenguaje	0.21	1.24
79	constructivo	0.51	1.66	79	veu	0.21	1.23
80	receta	0.5	1.65	80	decilitro	0.21	1.23
81	murciano	0.5	1.65	81	segregar	0.21	1.23
82	barcelona	0.49	1.63	82	nunha	0.21	1.23
83	haría	0.49	1.63	83	luz	0.21	1.23
84	lío	0.49	1.63	84	doloroso	0.21	1.23

85	élite	0.49	1.62	85	alde	0.21	1.23
86	sentencia	0.48	1.62	86	page	0.21	1.23
87	burgos	0.48	1.62	87	ourense	0.21	1.23
88	fotografía	0.48	1.62	88	alquiler	0.21	1.23
89	badajoz	0.48	1.62	89	prision	0.21	1.23
90	climático	0.48	1.62	90	circular	0.2	1.23
91	tenerife	0.48	1.61	91	javier_fernández	0.2	1.23
92	saludo	0.48	1.61	92	dimitir	0.2	1.23
93	transversal	0.48	1.61	93	arena	0.2	1.22
94	lliures	0.48	1.61	94	ple	0.2	1.22
95	castilla-la_mancha	0.48	1.61	95	asistencial	0.2	1.22
96	parálisis	0.47	1.61	96	dui	0.2	1.22
97	rodear	0.47	1.6	97	😑	0.2	1.22
98	negativa	0.47	1.6	98	boa	0.2	1.22
99	ministro_de_fomento	0.47	1.6	99	felicitats	0.2	1.22
100	latín	0.47	1.6	100	alcalá_de_henares	0.2	1.22

SVM Male features			SVM Female features		
#	Words	Coef	#	Words	Coef
1	♂	0.49	1	compromis	0.74
2	bns	0.44	2	📢	0.71
3	tarragona	0.43	3	😊	0.49
4	almansa	0.41	4	❤	0.48
5	tabarnia	0.4	5	♀	0.45
6	retweeted	0.38	6	😘	0.44
7	comunista	0.36	7	feminismo	0.41
8	leonés	0.36	8	🎹	0.41
9	🐿	0.34	9	mujer	0.4
10	malpartida_de_cáceres	0.34	10	🌹	0.39
11	región_de_murcia	0.32	11	💗	0.38
12	cáceres	0.32	12	▶	0.36
13	león	0.32	13	fuengirola	0.35
14	and	0.31	14	🎄	0.34
15	canario	0.31	15	sexual	0.34
16	base	0.3	16	inhumano	0.33
17	haiku	0.3	17	feminista	0.33
18	👼	0.3	18	😍	0.33
19	concierto	0.29	19	ceuta	0.32
20	riojano	0.29	20	🏐	0.32
21	📸	0.28	21	machismo	0.31
22	hores	0.28	22	zorionak	0.31
23	novela	0.28	23	maltratador	0.31
24	inversions	0.27	24	presidenta	0.31
25	urgell	0.27	25	😣	0.3
26	barkos	0.27	26	😱	0.3
27	editorial	0.27	27	principado	0.29
28	😁	0.27	28	😉	0.29
29	clm	0.27	29	pacto_de_estado	0.28

30 	0.27	30 galiza	0.28
31 saludo	0.26	31 completamente	0.27
32 idioma	0.26	32	0.27
33 detener	0.26	33	0.26
34 upn	0.25	34 masclisme	0.26
35 romano	0.25	35 junta_de_andalucía	0.26
36 incapaz	0.25	36 acento	0.26
37 boadilla	0.25	37 ❤	0.25
38 tav	0.25	38 machista	0.25
39 radial	0.25	39 don	0.25
40 vuelo	0.25	40 estereotipo	0.25
41 	0.25	41 precioso	0.25
42 l'economia	0.25	42	0.25
43 derecha	0.24	43 drets	0.25
44 veto	0.24	44 junta_de_castilla	0.25
45 desfile	0.24	45 ✔	0.25
46 	0.24	46 parlamento_de_canarias	0.25
47 cospedal	0.24	47 saneamiento	0.24
48 bipartidismo	0.24	48 avalar	0.24
49 millora	0.23	49 harto	0.24
50 tramo	0.23	50 ejército	0.24
51 crec	0.23	51 contentar	0.24
52 huesca	0.23	52 forta	0.24
53 evolución	0.23	53 illes_balears	0.24
54 	0.23	54 invisible	0.23
55 cartagena	0.23	55 doncs	0.23
56 the	0.23	56 masclista	0.23
57 cole	0.23	57 prevención	0.23
58 pista	0.23	58 vigo	0.23
59 unanimidad	0.23	59 pensionistes	0.23
60 confirmar	0.22	60 andalucía	0.23
61 vino	0.22	61 som-hi	0.22
62 	0.22	62 gallego	0.22
63 junqueras	0.22	63 anular	0.22
64 liberal	0.22	64 vanguardia	0.22
65 preservar	0.22	65 imos	0.22
66 división	0.22	66 almería	0.22
67 cuenca	0.22	67 bravo	0.22
68 penal	0.22	68	0.22
69 superior	0.22	69 reproducción	0.22
70 castellano	0.22	70 lgtbi	0.22
71 teruel	0.22	71 madrileño	0.21
72 detalle	0.21	72 piel	0.21
73 eleccions	0.21	73 cohesió	0.21
74 aeropuerto	0.21	74 comité	0.21
75 murciano	0.21	75 apoio	0.21
76 ganador	0.21	76 d.e._p	0.21
77 consejería	0.21	77 compostela	0.21

78	extremeño	0.21	78	lenguaje	0.21
79	impresentable	0.21	79	veu	0.21
80	auditorio	0.21	80	decilitro	0.21
81	pablo	0.21	81	segregar	0.21
82	inauguración	0.21	82	nunha	0.21
83	lleida	0.21	83	luz	0.21
84	avería	0.21	84	doloroso	0.21
85	dictar	0.21	85	alde	0.21
86	viejo	0.21	86	page	0.21
87	referir	0.21	87	ourense	0.21
88	concejal	0.21	88	alquiler	0.21
89	mil	0.21	89	prision	0.21
90	constructivo	0.21	90	circular	0.2
91	murcia	0.21	91	javier_fernández	0.2
92	complemento	0.21	92	dimitir	0.2
93	lío	0.21	93	arena	0.2
94	paisano	0.2	94	ple	0.2
95	tranquilo	0.2	95	asistencial	0.2
96	secesionista	0.2	96	dui	0.2
97	cobarde	0.2	97	😑	0.2
98	🍥	0.2	98	boa	0.2
99	leo	0.2	99	felicitats	0.2
100	medalla	0.2	100	alcalá_de_henares	0.2

Appendix II: top features indicative of receiver gender

L1 Male features				L1 Female features		
#	Words	Coef	Odds	# Words	Coef	Odds
1	lluís	4.9	134.62	1 zaida	4.26	70.66
2	gaspar	4.29	72.9	2 carolina	4.18	65.67
3	borja	3.13	22.92	3 teresa	4.01	54.99
4	anchoa	3.13	22.77	4 miriam	3.99	53.93
5	carrizosa	3.08	21.79	5 marta	3.96	52.57
6	presi	2.99	19.85	6 mireia	3.91	49.93
7	sergi	2.97	19.44	7 eva	3.83	45.84
8	juan	2.96	19.38	8 dolors	3.72	41.19
9	héctor	2.89	18.06	9 ana	3.61	36.85
10	eduardo	2.88	17.81	10 tranquila	3.4	29.98
11	sergio	2.86	17.47	11 andrea	3.39	29.8
12	odón	2.79	16.24	12 patricia	3.31	27.31
13	santander	2.78	16.05	13 anna	3.29	26.94
14	neptuno	2.74	15.55	14 presidenta	3.29	26.87
15	coscu	2.71	15.05	15 lorena	3.27	26.42
16	harvard	2.67	14.42	16 mónica	3.16	23.6
17	gràcies_lluís	2.64	13.98	17 irene	3.15	23.37
18	juanma	2.57	13.09	18 hitleriano	3.08	21.73
19	profesorado	2.52	12.42	19 president_mas	2.89	18.05
20	raúl	2.52	12.4	20 psoe-cha	2.88	17.86
21	joan	2.44	11.45	21 adriana	2.76	15.74
22	miquel	2.44	11.43	22 carme	2.65	14.22
23	oriol	2.42	11.28	23 ines	2.5	12.24
24	idus	2.42	11.28	24 sra.	2.5	12.19
25	panem	2.38	10.82	25 patriarcado	2.45	11.56
26	ramon	2.33	10.24	26 señora	2.43	11.41
27	segueixen	2.3	9.99	27 katana	2.43	11.3
28	victòria	2.28	9.75	28 inés	2.37	10.7
29	📻	2.27	9.66	29 portavoza	2.33	10.23
30	rafa	2.26	9.63	30 gabriela	2.32	10.15
31	penalti	2.26	9.59	31 liberal_de_castilla	2.14	8.54
32	árbitro	2.25	9.47	32 galiza	2.12	8.34
33	toni	2.25	9.45	33 susana	2.1	8.16
34	catedral	2.23	9.28	34 anticapitalista	2.05	7.75
35	josep	2.22	9.22	35 consellera	2.05	7.74
36	xavier	2.22	9.2	36 retirada	1.98	7.21
37	iniesta	2.21	9.1	37 senyora	1.95	7.06
38	ramón	2.14	8.5	38 maca	1.93	6.92
39	comunero	2.13	8.42	39 dracs	1.91	6.73
40	transformador	2.11	8.21	40 bolso	1.89	6.59
41	jesús	2.08	8	41 suizo	1.88	6.58
42	lluis	2.07	7.92	42 suiza	1.85	6.36
43	quántum	2.04	7.71	43 susanita	1.81	6.12

44 vpo	2.03	7.58	44 ombligo	1.8	6.02
45 equiparacion	2.01	7.49	45 patriarcal	1.76	5.81
46 paco	1.98	7.24	46 techo	1.76	5.8
47 raül	1.98	7.23	47 sra	1.75	5.76
48 suma	1.97	7.19	48 zara	1.75	5.76
49 frustración	1.95	7.03	49 inscribir	1.74	5.69
50 conseller	1.91	6.75	50 brujo	1.71	5.55
51 traducir	1.9	6.71	51 marruecos	1.68	5.39
52 carlos	1.88	6.54	52 pescar	1.66	5.28
53 papi	1.86	6.41	53 cristina	1.66	5.26
54 miguel	1.84	6.28	54 balear	1.64	5.16
55 👀	1.83	6.21	55 carmen	1.63	5.1
56 tranquil	1.81	6.13	56 tania	1.58	4.86
57 caseta	1.81	6.12	57 rota	1.55	4.71
58 revilla	1.79	5.98	58 isabel	1.54	4.65
59 girauta	1.77	5.85	59 correo	1.52	4.59
60 infraestructura	1.76	5.82	60 molotov	1.51	4.54
61 noi	1.74	5.72	61 feminismo	1.51	4.53
62 aravaca	1.73	5.63	62 vulnerable	1.51	4.51
63 tamaño	1.72	5.58	63 eliminación	1.5	4.49
64 mio	1.7	5.49	64 asistente	1.5	4.47
65 consecuente	1.7	5.49	65 marxa	1.48	4.41
66 patología	1.7	5.45	66 pspv	1.44	4.24
67 postureig	1.69	5.42	67 bon_sant_jordi	1.42	4.14
68 singular	1.69	5.4	68 inés_arrimadas	1.42	4.12
69 black	1.69	5.4	69 oltra	1.4	4.05
70 serrat	1.67	5.32	70 argelino	1.39	4.02
71 luis	1.67	5.31	71 congelar	1.39	4
72 promoción	1.66	5.28	72 ipc	1.38	3.98
73 ciudadania	1.64	5.15	73 carencia	1.37	3.93
74 antonio	1.64	5.14	74 noia	1.36	3.91
75 motiu	1.64	5.14	75 igualment	1.36	3.9
76 abc	1.62	5.03	76 tia	1.36	3.89
77 estable	1.6	4.96	77 inmigrante	1.33	3.78
78 xustiza	1.6	4.96	78 requisar	1.32	3.76
79 d'europa	1.6	4.96	79 capacitat	1.31	3.72
80 monarquia	1.59	4.91	80 alfa	1.31	3.71
81 xavi	1.58	4.85	81 contentar	1.3	3.68
82 albano	1.58	4.84	82 guapi	1.3	3.67
83 caure	1.57	4.83	83 front	1.28	3.61
84 cartagena	1.57	4.81	84 vigo	1.27	3.58
85 superioridad	1.57	4.8	85 valenta	1.26	3.54
86 juan_carlos	1.56	4.74	86 el_liberal_de_castilla	1.25	3.48
87 rufián	1.55	4.7	87 marroquí	1.24	3.46
88 teoría	1.54	4.65	88 irán	1.24	3.45
89 aumento	1.54	4.64	89 brecha	1.23	3.43
90 literatura	1.53	4.64	90 tolerancia	1.22	3.39
91 íñigo	1.51	4.52	91 sant	1.21	3.35

92	rufi	1.51	4.5	92	guardería	1.2	3.33
93	hoja	1.5	4.5	93	euskera	1.19	3.3
94	ignacio	1.5	4.47	94	diablo	1.19	3.29
95	estimat	1.49	4.44	95	clara	1.19	3.28
96	gilipollez	1.49	4.42	96	peste	1.19	3.27
97	vicepresident	1.48	4.4	97	admiración	1.18	3.25
98	héroe	1.46	4.33	98	educació	1.18	3.25
99	puerto	1.45	4.26	99	feminista	1.14	3.11
100	payaso	1.44	4.2	100	efectiu	1.13	3.1

L2 Male features				L2 Female features			
#	Words	Coef	Odds	#	Words	Coef	Odds
1	oriol	1.28	3.61	1	marta	2.11	8.27
2	conseller	1.26	3.53	2	presidenta	1.58	4.86
3	gaspar	1.25	3.51	3	dolors	1.49	4.42
4	lluís	1.09	2.97	4	anna	1.47	4.34
5	joan	1.07	2.92	5	zaida	1.45	4.28
6	payaso	1.02	2.79	6	carme	1.41	4.08
7	toni	1.02	2.77	7	teresa	1.35	3.86
8	rafa	0.9	2.45	8	ana	1.34	3.83
9	black	0.85	2.35	9	carolina	1.2	3.33
10	miquel	0.84	2.32	10	suiza	1.2	3.33
11	revilla	0.83	2.29	11	irene	1.2	3.32
12	beca	0.8	2.22	12	feminismo	1.16	3.2
13	juan	0.78	2.18	13	susana	1.15	3.16
14	miguel	0.78	2.18	14	mireia	1.1	3.01
15	jordi	0.78	2.18	15	lorena	1.06	2.87
16	vpo	0.77	2.17	16	consellera	1.05	2.86
17	raül	0.77	2.17	17	inés	1.02	2.77
18	xavier	0.74	2.09	18	andrea	1	2.72
19	crack	0.74	2.09	19	mujer	0.98	2.67
20	carlos	0.73	2.08	20	portavoza	0.96	2.6
21	gilipollas	0.72	2.06	21	patricia	0.95	2.58
22	alcalde	0.72	2.05	22	galiza	0.93	2.55
23	coscu	0.71	2.03	23	patriarcado	0.93	2.54
24	sergio	0.7	2.02	24	eva	0.92	2.5
25	ministro	0.7	2.02	25	adriana	0.91	2.48
26	entregar	0.7	2	26	miriam	0.88	2.41
27	presi	0.69	2	27	cristina	0.84	2.31
28	ramón	0.69	1.99	28	sra.	0.83	2.28
29	albano	0.68	1.97	29	ines	0.82	2.28
30	iñigo	0.67	1.96	30	señora	0.82	2.28
31	📻	0.67	1.96	31	feminista	0.82	2.27
32	pablo	0.66	1.94	32	tranquila	0.77	2.15
33	violento	0.66	1.94	33	mónica	0.76	2.14
34	girauta	0.66	1.93	34	guapo	0.76	2.14

35 senyor	0.65	1.92	35 president_mas	0.76	2.14
36 figura	0.65	1.92	36 psoe-cha	0.75	2.12
37 josep	0.65	1.91	37 hitleriano	0.73	2.07
38 héctor	0.64	1.9	38 patriarcal	0.72	2.06
39 borja	0.64	1.9	39 sra	0.71	2.04
40 bar	0.63	1.89	40 😘	0.69	1.99
41 rufián	0.63	1.89	41 igualdad	0.69	1.99
42 comisaría	0.63	1.88	42 alimentar	0.65	1.91
43 albert	0.63	1.87	43 diada	0.64	1.9
44 harvard	0.63	1.87	44 tania	0.63	1.89
45 odón	0.62	1.87	45 senyora	0.63	1.88
46 anchoa	0.61	1.84	46 katana	0.62	1.86
47 mariano	0.61	1.84	47 sant	0.62	1.86
48 paco	0.6	1.83	48 molotov	0.62	1.85
49 noi	0.6	1.82	49 bon_sant_jordi	0.6	1.83
50 generar	0.6	1.82	50 argelino	0.6	1.83
51 carretera	0.59	1.81	51 carmen	0.6	1.82
52 frustración	0.58	1.79	52 marroquí	0.6	1.82
53 pelea	0.58	1.79	53 coherente	0.6	1.82
54 barco	0.58	1.79	54 anticapitalista	0.6	1.82
55 monarquia	0.58	1.79	55 euskera	0.6	1.81
56 penalti	0.58	1.78	56 rosa	0.59	1.81
57 alberto	0.58	1.78	57 zara	0.59	1.81
58 raúl	0.58	1.78	58 película	0.59	1.81
59 eduardo	0.57	1.78	59 tolerancia	0.59	1.8
60 juanma	0.57	1.77	60 andalucía	0.59	1.8
61 asistir	0.56	1.75	61 gabriela	0.59	1.8
62 reflexión	0.56	1.74	62 suizo	0.58	1.79
63 malversación	0.56	1.74	63 machista	0.58	1.79
64 garzón	0.56	1.74	64 acta	0.58	1.79
65 carrizosa	0.55	1.74	65 reina	0.58	1.78
66 victòria	0.55	1.73	66 género	0.57	1.77
67 mio	0.55	1.73	67 violar	0.57	1.77
68 david	0.55	1.73	68 vulnerable	0.56	1.76
69 felicitats	0.55	1.72	69 firme	0.56	1.76
70 sergi	0.54	1.72	70 clara_campoamor	0.56	1.76
71 ave	0.54	1.72	71 noia	0.56	1.75
72 árbitro	0.54	1.71	72 guapi	0.56	1.75
73 transición	0.53	1.71	73 pescar	0.54	1.71
74 politico	0.53	1.7	74 marruecos	0.53	1.7
75 acudir	0.53	1.7	75 ipc	0.52	1.68
76 caseta	0.53	1.7	76 beso	0.52	1.68
77 facha	0.53	1.69	77 violador	0.51	1.67
78 democràtic	0.52	1.69	78 vital	0.51	1.67
79 amenazar	0.52	1.69	79 atacar	0.51	1.66
80 aragón	0.52	1.68	80 bloque	0.51	1.66
81 futur	0.52	1.68	81 congelar	0.5	1.65
82 empresario	0.52	1.68	82 arrimadas	0.5	1.65

83	copa	0.51	1.67	83	polo	0.5	1.65
84	latín	0.51	1.67	84	hagués	0.5	1.65
85	jamar	0.51	1.67	85	peste	0.5	1.65
86	extremadura	0.51	1.67	86	cuidar	0.5	1.65
87	murcia	0.51	1.67	87	socialista	0.5	1.64
88	equiparación	0.51	1.67	88	guardería	0.5	1.64
89	pedro	0.51	1.66	89	correo	0.49	1.64
90	infraestructura	0.51	1.66	90	entrevista	0.49	1.63
91	iceta	0.5	1.66	91	sentencia	0.49	1.63
92	equiparacion	0.5	1.66	92	animar	0.49	1.63
93	internacional	0.5	1.65	93	inmigrante	0.49	1.63
94	ahir	0.5	1.65	94	isabel	0.48	1.62
95	financiación	0.5	1.65	95	princesa	0.48	1.62
96	jubilar	0.5	1.65	96	rescatar	0.48	1.62
97	patología	0.5	1.65	97	buenos	0.48	1.62
98	burgos	0.5	1.65	98	tia	0.48	1.61
99	decena	0.5	1.64	99	techo	0.48	1.61
100	carles	0.49	1.64	100	bolso	0.48	1.61

SVM Male features		SVM Female features	
# Words	Coef	# Words	Coef
1 oriol	0.51	1 marta	0.8
2 gaspar	0.5	2 presidenta	0.64
3 conseller	0.48	3 anna	0.62
4 lluís	0.42	4 dolors	0.59
5 joan	0.42	5 zaida	0.56
6 payaso	0.41	6 carme	0.56
7 toni	0.41	7 irene	0.53
8 beca	0.38	8 susana	0.51
9 rafa	0.37	9 ana	0.5
10 revilla	0.36	10 teresa	0.49
11 xavier	0.36	11 mireia	0.48
12 black	0.35	12 carolina	0.47
13 miquel	0.33	13 suiza	0.47
14 jordi	0.33	14 inés	0.47
15 josep	0.32	15 lorena	0.46
16 violento	0.32	16 miriam	0.46
17 ministro	0.31	17 andrea	0.44
18 raül	0.3	18 feminismo	0.43
19 vpo	0.3	19 mujer	0.39
20 juan	0.3	20 patricia	0.39
21 cantabria	0.3	21 patriarcado	0.38
22 upn	0.29	22 sra.	0.38
23 miguel	0.29	23 ines	0.38
24 ramón	0.29	24 portavoza	0.37
25 pablo	0.29	25 consellera	0.37
26 girauta	0.29	26 feminista	0.37
27 crack	0.29	27 eva	0.37

28 alcalde	0.28	28 galiza	0.37
29 coscu	0.28	29 cristina	0.36
30 gilipollas	0.27	30 adriana	0.35
31 sergio	0.27	31 señora	0.34
32 albano	0.27	32 andalucía	0.33
33 murcia	0.27	33 igualdad	0.31
34 carrizosa	0.27	34 president_mas	0.31
35 borja	0.27	35 patriarcal	0.31
36 antonio	0.27	36 hitleriano	0.3
37 felicitats	0.27	37 guapo	0.3
38 president	0.27	38 😘	0.3
39 carlos	0.26	39 machista	0.3
40 rufián	0.26	40 sra	0.29
41 malversación	0.26	41 carmen	0.29
42 albert	0.26	42 tranquila	0.27
43 humor	0.26	43 vigo	0.27
44 empresario	0.26	44 tania	0.27
45 senyor	0.26	45 molotov	0.27
46 extremadura	0.25	46 género	0.26
47 facha	0.25	47 gabriela	0.26
48 bar	0.25	48 psoe-cha	0.26
49 ave	0.25	49 mónica	0.26
50 presi	0.25	50 rosa	0.25
51 anchoa	0.25	51 marruecos	0.25
52 alemania	0.25	52 katana	0.25
53 pelea	0.25	53 marroquí	0.25
54 figura	0.25	54 suizo	0.24
55 paco	0.25	55 argelino	0.24
56 héctor	0.25	56 película	0.24
57 noi	0.24	57 guapi	0.24
58 reflexión	0.24	58 admiración	0.24
59 estupendo	0.24	59 non	0.23
60 monarquia	0.24	60 tolerancia	0.23
61 harvard	0.24	61 isabel	0.23
62 pedro	0.24	62 violar	0.23
63 david	0.24	63 anticapitalista	0.23
64 chaval	0.24	64 firme	0.23
65 internacional	0.23	65 sant	0.23
66 generar	0.23	66 beso	0.23
67 politico	0.23	67 entrevista	0.23
68 victòria	0.23	68 sentencia	0.23
69 infraestructura	0.23	69 ipc	0.22
70 equiparacion	0.23	70 animar	0.22
71 odón	0.23	71 🤍	0.22
72 juanma	0.23	72 zara	0.22
73 luis	0.23	73 noia	0.22
74 comisaría	0.23	74 senyora	0.22
75 mio	0.23	75 diada	0.22

76	mariano	0.22	76	constituir	0.22
77	caseta	0.22	77	efectiu	0.22
78	amenazar	0.22	78	dracs	0.22
79	burgos	0.22	79	reina	0.22
80	fiscal	0.22	80	nens	0.22
81	badalona	0.22	81	vulnerable	0.22
82	árbitro	0.22	82	correo	0.21
83	eduardo	0.22	83	bona	0.21
84	garzón	0.22	84	socialista	0.21
85	mañana	0.22	85	cuestión	0.21
86	listo	0.22	86	susanita	0.21
87	papi	0.22	87	clara_campoamor	0.21
88	trilero	0.21	88	inés_arrimadas	0.21
89	raúl	0.21	89	histórico	0.21
90	futur	0.21	90	alimentar	0.21
91	scc	0.21	91	bon_sant_jordi	0.21
92	😜	0.21	92	bat	0.21
93	financiación	0.21	93	tornis	0.21
94	📻	0.21	94	machismo	0.21
95	maltratar	0.21	95	alfa	0.21
96	barco	0.21	96	acta	0.21
97	vasco	0.21	97	revisable	0.21
98	🎗	0.21	98	inscribir	0.21
99	entregar	0.21	99	congelar	0.2
100	frustración	0.21	100	maca	0.2