ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-769

An Assessment Study of Features and Meta-level Features in Twitter Sentiment Analysis

Jonnathan Carvalho[1,2] and **Alexandre Plastino**[2]

Abstract. Sentiment analysis is the task of determining the opinion expressed on subjective data, which may include microblog messages, such as tweets. This type of message has been considered the target of sentiment analysis in many recent studies, since they represent a rich source of opinionated texts. Thus, in order to determine the opinion expressed in tweets, different studies have employed distinct strategies, which mainly include supervised machine learning methods. For this purpose, different kinds of features have been evaluated. Despite that, none of the state-of-the-art studies has evaluated distinct categories of features, regarding their similar characteristics. In this context, this work presents a literature review of the most common feature representation in Twitter sentiment analysis. We propose to group features sharing similar aspects into specific categories. We also evaluate the relevance of these categories of features, including meta-level features, using a significant number of Twitter datasets. Furthermore, we apply important and well-known feature selection strategies in order to identify relevant subsets of features for each dataset. We show in the experimental evaluation that the results achieved in this study, using feature selection strategies, outperform the results reported in previous works for the most of the assessed datasets.

1 Introduction

Sentiment analysis has been extensively used to determine the overall opinion expressed about different targets in many types of user-generated documents, generally on the Web, such as user reviews, blog comments, etc. Many companies have taken advantage of the area of sentiment analysis by automatically extracting the opinions expressed by consumers about their products and services, eliminating the need of extensive and expensive researches.

With the advent of social media and microblog platforms, such as Twitter[3], not only the opinion about products and services can be tracked, but also the sentiment expressed about entities in real time events, such as politics debates, world disasters, etc. Twitter is a microblog platform, in which users can send short messages about any topic, limited to 140 characters. In order to determine the sentiment expressed in this type of message, the so-called tweets, different approaches have been proposed in the literature of Twitter sentiment analysis. These approaches mainly include supervised machine learning strategies and they usually focus on the polarity classification of tweets, that is, whether the sentiment expressed on them carries a positive or negative connotation.

[1] Instituto Federal Fluminense, Itaperuna, Brazil, email: joncarv@iff.edu.br
[2] Universidade Federal Fluminense, Institute of Computing, Niterói, Brazil, emails: joncarv@ic.uff.br, plastino@ic.uff.br
[3] `http://twitter.com`

Supervised machine learning strategies aim at determining the sentiment expressed in tweets by exploring their contents in order to learn characteristics, commonly referred to as features, that can distinguish the positive tweets from the negative ones. The most usual feature representation in the task of sentiment classification of tweets is the bag-of-words model, first employed by Go et al. [17], in which each token of a tweet is taken as a feature. Regarding the challenging nature of tweets, such as misspelling words and the 140-character limit of each message, different studies have proposed other types of features, trying to improve the performance of the sentiment classification. These features include, in their vast majority, n-grams [2, 6, 7, 12, 13, 17, 19, 22, 24, 25, 27, 29, 30, 38, 40, 44], part-of-speech tags [2, 5, 8, 17, 24, 27], punctuation [2, 5, 13, 19, 22, 27], specific characteristics of Twitter and microblog messages [2, 5, 19, 22, 24, 27, 44], and lexicon-based features [2, 8, 12, 19, 22, 23, 24, 27].

Although these features have been successfully employed in the sentiment classification of tweets, none of the state-of-the-art studies has organized into categories the large set of features used in the sentiment classification of tweets from distinct domains, including the most common features and meta-level features that have already been proposed in the literature. In this context, considering that many of these features share similar characteristics, we propose to group them into different categories, i.e., features that are similar in structural aspects make up the same category. Then, we investigate whether the classification of tweets from different domains can benefit from distinct categories of features and meta-level features. In addition, we explore the application of feature selection strategies in order to identify relevant subsets of features for each evaluated dataset.

The main contributions of this study are the following.

1. We present an extensive literature review of the most common feature representation of tweets for supervised sentiment classification, including meta-level features, which have been proposed and employed in a relevant set of well-referenced works.
2. We propose to group the features and meta-level features sharing similar aspects into specific categories, in order to investigate the importance of each of these categories in the sentiment classification of tweets from distinct domains.
3. We use a collection of sixteen Twitter datasets in the experimental evaluation of the categories of features and meta-level features. To the best of our knowledge, this is the first study that evaluates distinct categories of features and meta-level features for a significant number of Twitter datasets.
4. In order to identify relevant subsets of features for each dataset, we apply important feature selection strategies on the full set of features, including measures such as Information Gain, Chi-Squared, and Relief-F.

5. We show that the results achieved in this study, using feature selection strategies, outperform previous results reported in the literature of Twitter sentiment analysis for the most of the assessed datasets.

The remainder of this work is organized as follows. Section 2 presents the related work on sentiment classification, focusing on the different types of features explored in the literature. In Section 3, we present the most common features and meta-level features identified in a set of well-referenced works. Besides, we propose the categorization of the features, based on their similar characteristics. The experimental evaluation of the different categories of features are reported in Section 4, as well as the results of the application of the feature selection strategies. Finally, in Section 5, we present the conclusions of this study and some future work.

2 Related Work

Sentiment analysis has been widely employed to determine the polarity of subjective data, that is, whether the sentiment expressed in opinionated text (movie reviews, blogs, microblogs, etc.) has a positive or negative connotation. For this purpose, different approaches have already been proposed, which mainly include supervised machine learning methods and lexicon-based strategies.

Regarding supervised machine learning methods, which are the main focus of this study, the precursor work by Pang et al. [31] applies different algorithms in the classification of movie reviews. In supervised methods, to classify a review as being positive or negative, a training dataset is used, which consists of pieces of texts, generally represented as a feature vector, whose polarities are already known. Such feature vector may contain relevant characteristics of each piece of text in the training dataset, the so-called features. In [31], in addition to the use of different machine learning algorithms (Naive Bayes, SVM, and Maximum Entropy), they also evaluate distinct sets of features, such as unigrams (bag-of-words), bigrams, and part-of-speech (POS) tags. Their experimental results show that better performance is achieved using only unigrams as features, and they conclude it is worthwhile to explore the data, in order to select good indicator features for sentiment classification.

More recent works explore the applicability of feature selection methods, attempting to improve the sentiment classification of movie reviews [3, 37]. Indeed, Sharma and Dey [37] show that feature selection may improve the performance of sentiment classification, although the improvement is dependent on the feature selection method employed and the number of features selected. Similarly, Agarwal and Mittal [3] study the effect of different feature selection methods and various sets of features from datasets of reviews in two distinct domains, such as movies and products. They show that features created from unigrams and bigrams achieves better results when compared to the use of unigrams or bigrams individually. Differently from [3] and [37], the strategy described in [1] does not evaluate distinct feature selection methods. In [1], Abbasi et al. propose a genetic algorithm that incorporates the Information Gain measure for feature selection on a corpus of movie reviews, using stylistic and syntactic features, such as word length distributions, and special character frequencies.

Beyond the classification of opinions expressed in movie and product reviews, the sentiment expressed in other types of documents has been evaluated, such as in tweets. The method presented by Go et al. [17] classifies the sentiment of tweets using a distant supervision approach, which relies on emoticons as noisy labels in a training dataset of 1,600,000 tweets. They also evaluate the performance of different sets of features, such as unigrams, bigrams, both unigrams and bigrams, and part-of-speech tags. The experimental evaluation, in [17], shows that unigrams and bigrams together represent a good set of features for the sentiment classification of tweets.

Inspired by Go et al. [17], many other approaches explore the use of n-grams in Twitter sentiment classification [2, 6, 7, 12, 13, 17, 19, 22, 24, 25, 27, 29, 30, 38, 40, 44]. For example, Davidov et al. [13] use higher order n-grams, such as bigrams, trigrams, 4-grams, and 5-grams as features in the classification process. To avoid data sparsity, due to the use of different range of n-grams, they only extract n-grams features from words which have a training set frequency over 0.5%.

Although n-grams have been a kind of feature largely evaluated in the literature, some works incorporate different types of features in the sentiment classification task, such as punctuation, microblog and Twitter-specific features, and lexicon-based features [2, 5, 8, 12, 13, 19, 22, 23, 24, 27, 44]. They claim that using only n-grams may not be sufficient for this task, since tweets are very noisy and short messages. For example, Kouloumpis et al. [24] take into account the presence of positive and negative emoticons, as well as abbreviations and intensifiers (all-caps and character repetitions). They point out that the best performance on their experimental results comes from using n-grams together to the lexicon-based and the microblog features. In addition to the use of different sets of features, in [24], a feature selection method is applied in order to select only the 1,000 most representative n-grams for the classification of tweets.

Feature selection methods have also been tried in Twitter sentiment classification [8, 33], as mentioned before, concerning the vast amount of redundancy features and to reduce the feature space. For example, Bravo-Marquez et al. [8] first compute the information gain of the proposed set of lexicon-based meta-level features. Then, they test combinations of feature subsets, selecting arbitrary sets of features with a best-first strategy, based on the information gain of the primarily evaluated features, on three datasets of tweets. Prusa et al. [33] evaluate the impact of feature selection techniques on a corpus of 3,000 tweets, using only unigrams as features. In their experimental evaluation, they applied ten different filter-based feature selection techniques and ten different sizes of the feature subsets, varying from 5 to 200, and they show that feature selection methods can be effective in the sentiment classification of tweets.

Despite the previous use of feature selection methods in Twitter sentiment classification, to the best of our knowledge, this work is the first study that evaluates the importance of distinct categories of features and meta-level features for a significant number of Twitter datasets. These features and meta-level features, which include n-gram-based features, microblog and Twitter-specific features, part-of-speech tags, punctuation features, and lexicon-based features, were identified in a set of well-referenced works in the literature of supervised sentiment classification of tweets, after an extensive literature review. We also study the effect of the application of different feature selection strategies on the full set of features, in order to identify relevant subsets of features for each evaluated dataset.

3 Features and Meta-level Features

Different types of features have been engineered and used in Twitter sentiment analysis, from the most common representation, such as n-gram-based features, to meta-level features. Meta-level features are usually extracted from other features, and can capture insightful new information about the data [9]. In this study, we consider merely as features the information that can be extracted primarily from tweets, such as the presence or absence of some particular characteristic in a tweet. In the other hand, we consider as meta-level features those

referred to counts and summations, which are, in general, secondary information extracted from tweets. For readability reasons, meta-level features are referred to hereafter as meta-features.

In this section, we describe the features and meta-features we have examined in a set of well-referenced works in supervised sentiment classification of tweets. These works were identified after an extensive literature review, from which we have detected the most common types of features and meta-features used to determine the sentiment expressed in tweets. In order to describe and evaluate these features and meta-features, as shown in Table 1, we have grouped them into five categories, namely N-grams, Twitter and Microblog, Part-of-Speech, Punctuation, and Polarity, so that features that share structural aspects fall into the same category.

Table 1. Categories of features employed in the literature of supervised sentiment classification of tweets.

Reference	$f1$	$f2$	$f3$	$f4$	$f5$
Agarwal et al. [2]	✓	✓	✓	✓	✓
Barbosa and Feng [5]		✓	✓	✓	
Bermingham and Smeaton [6]	✓				
Bifet et al. [7]	✓				
Bravo-Marquez et al. [8]			✓		✓
da Silva et al. [12]	✓				✓
Davidov et al. [13]	✓			✓	
Go et al. [17]	✓		✓		
Hagen et al. [19]	✓	✓		✓	✓
Jiang et al. [22]	✓	✓		✓	✓
Khuc et al. [23]					✓
Kouloumpis et al. [24]	✓	✓	✓		✓
Lin et al. [25]	✓				
Mohammad et al. [27]	✓	✓	✓	✓	✓
Narr et al. [29]	✓				
Pak and Paroubek [30]	✓				
Speriosu et al. [38]	✓				
Wang et al. [40]	✓				
Zhang et al. [44]	✓	✓			

$f1$ N-grams $f2$ Twitter and Microblog $f3$ POS $f4$ Punctuation $f5$ Polarity

3.1 N-grams Features

N-grams features are contiguous sequences of n tokens from a text. The n-gram-based features were first employed in sentiment classification of tweets by Go et al. [17]. Since then, this category of features has been the one most used by supervised sentiment learning strategies [2, 6, 7, 12, 13, 17, 19, 22, 24, 25, 27, 29, 30, 38, 40, 44].

In the bag-of-words model, that is, the unigram category (n = 1), each word or token is used as a feature. It is the basic representation of a tweet for the classification process, and it is adopted by many strategies [2, 6, 7, 12, 13, 17, 19, 22, 24, 27, 29, 30, 38, 40, 44]. In an attempt to capture more sentiment expressions, some studies have varied the value of n. For example, Davidov et al. [13] vary the value of n from 1 to 5, which means that each consecutive word sequence containing one to five words is taken as a feature. Table 2 presents an overview of the n-grams features used in the literature of Twitter sentiment classification.

3.2 Twitter and Microblog Features

The Twitter and Microblog category refers to those features related to the syntax and vocabulary used in tweets and microblog messages, as used in [2, 5, 19, 22, 24, 27, 44]. More specifically, some characteristics of how microblog posts are written may be good indicators of sentiment, such as emoticons and internet slang present in the vocabulary of this type of text. Furthermore, Twitter-specific tokens, such as

Table 2. Overview of the n-grams features used in Twitter sentiment classification.

Reference	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
Agarwal et al. [2]	✓				
Bermingham and Smeaton [6]	✓	✓	✓		
Bifet et al. [7]	✓				
da Silva et al. [12]	✓				
Davidov et al. [13]	✓	✓	✓	✓	✓
Go et al. [17]	✓	✓			
Hagen et al. [19]	✓	✓	✓	✓	
Jiang et al. [22]	✓	✓			
Kouloumpis et al. [24]	✓	✓			
Lin et al. [25]				✓	
Mohammad et al. [27]	✓	✓	✓	✓	
Narr et al. [29]	✓	✓			
Pak and Paroubek [30]	✓	✓	✓		
Speriosu et al. [38]	✓	✓			
Wang et al. [40]	✓				
Zhang et al. [44]	✓				

user mentions (followed by the special character @), retweets (indicated by RT), URLs, and hashtags (followed by the special character #) have also been explored in the literature.

Twitter hashtags, which are often used as keywords for tweets, are a very informative mechanism. Thus, they may be a good evidence of positive or negative sentiment, as employed in [2, 5, 19, 22, 27, 44]. Similarly, others Twitter-specific tokens are taken as features in the literature, such as the presence of user mentions and retweets [5].

Regarding the 140-character limit of tweets, a very common trick established among Twitter users is the use of word shortcuts and internet slang (for example, "love" becomes "luv"). Another interesting aspect of tweets is the use of repeated letters as intensifiers (for example, in "looooove"). Thus, some works have defined these characteristics as meta-features as well [2, 19, 24, 27].

In this context, for this category, we identified the following set of nine features and meta-features, employed in the literature [2, 5, 19, 22, 24, 27, 44].

• *Whether the tweet has*: retweet, hashtag, user mentions, URL, emoticon, internet slang, repeated letters.

• *Number of*: internet slang, repeated letters.

3.3 Part-of-Speech Features

Although some studies have already acknowledged that part-of-speech (POS) features are not useful for sentiment classification [17, 31], this category of features is still used to determine the sentiment of tweets, in combination with other features [2, 5, 8, 17, 24, 27]. For example, assuming that some adjectives and verbs are good indicators of positive and negative sentiment, Barbosa and Feng [5] map each word in a tweet to its POS, using a POS database, which can identify nouns, verbs, adjectives, adverbs, interjections, and others. Similarly, Agarwal et al. [2] consider the number of adjectives, adverbs, verbs, and nouns as features. In order to capture the informal aspects of tweets, some works [8, 27] use a POS tagset, presented in [16], to identify some special characteristics of short and noisy texts, such as misspelling words.

In this context, for this category, we identified the following set of twenty-five features in the literature [2, 5, 8, 17, 24, 27].

• *Number of*: common noun, proper noun, personal pronoun, common noun + possessive, common noun + verb, proper noun + possessive, proper noun + verb, verb, adjective, adverb, interjection, determiner, pre or post-position, conjunction, verb particle, predeterminer, predeterminer + verb, hashtag, user mention, discourse marker ("RT" and ":" in retweet), URL or email address, emoticon, numeral, punctuation, abbreviation or symbol.

3.4 Punctuation Features

Punctuation may also play an important role in sentiment detection of microblog messages. For example, Bermingham and Smeaton [6] observed that the exclamation mark is the most discriminative unigram according to the Information Gain measure, in a corpus of 1,000 tweets labeled as being positive and negative. They also point out that the question mark and sequences of exclamation marks (for example, as "!!!") are in the top 10 most relevant features.

In this context, punctuation features have also been explored in the literature [2, 5, 13, 19, 22, 27]. The most usual meta-features in this category are the number of exclamation and question marks, as appearing in [2, 5, 13, 19, 22]. The total count of quotes in tweets has also been used [13]. Some works have already proposed more sophisticated meta-features, such as the number of contiguous sequences of exclamation and question marks [19, 27], regarding their use in microblog messages to convey intonation. Therefore, to make out this category of features, we identified the following set of ten features and meta-features [2, 5, 13, 19, 22, 27].

- *Whether the tweet has*: question mark, exclamation mark.
- *Whether last token contains*: question mark, exclamation mark.
- *Number of*: question mark, exclamation mark, sequence of question marks, sequence of exclamation marks, sequence of both question and exclamation marks, quotes.

3.5 Polarity Features

A different manner of exploring the content of tweets, in order to determine the sentiment expressed in them, is from using existing sentiment lexical resources or dictionaries in the literature. These lexicons consist of lists of words with positive and negative terms, such as Bing Liu's opinion lexicon [26], NRC-emotion [28], and OpinionFinder lexicon [43], as well as lexical resources containing words and phrases that are scored on a range of real values, such as SentiWordNet (SWN) [4], NRC-hashtag [27], and Sentiment140 lexicon (Sent140) [27]. Meta-features of this category have been widely explored in sentiment classification of tweets [2, 8, 12, 19, 22, 23, 24, 27], especially the total count of positive and negative words.

The polarity of emoticons may also be another relevant characteristic for Twitter sentiment analysis. Since emoticons are used by microblog users to summarize the sentiment they intend to communicate, some works have also extracted meta-features from emoticons, such as the number of positive and negative emoticons in a tweet, as employed in [2, 12, 19, 27].

Regarding negation, it has already been acknowledged it can affect the polarity of an expression [42]. Indeed, the expression "not good" is the opposite of "good". In this context, an interesting meta-feature proposed in the literature to handle negation is the number of negated contexts [27]. Mohammad et al. [27] have defined a negated context as a segment of a tweet that starts with a negation word, such as "shouldn't", and ends on the first punctuation mark after the negation word. Moreover, in [27], negated contexts change the n-gram-based features, that is, they add the tag _NEG on each token into a negated context. More specifically, in a negated context, Mohammad et al. concatenate the tag _NEG to every token between the negation word and the first punctuation mark after it. For example, in the sentence "He isn't a great book writer, but I read his books.", the unigrams "great", "book", and "writer" become "great_NEG", "book_NEG", and "writer_NEG", respectively.

Considering the polarity features identified in the literature [2, 8, 12, 19, 22, 23, 24, 27], the following features and meta-features compose this category.

- *Whether the tweet has*: positive emoticon, negative emoticon.
- *Whether the last token is*: positive emoticon, negative emoticon.
- *Number of*: positive emoticon, negative emoticon, extremely positive emoticon, extremely negative emoticon, positive adjective, negative adjective, positive noun, negative noun, positive adverb, negative adverb, positive verb, negative verb, negated contexts.
- *Sum of the scores of the adjectives, adverbs, verbs, and nouns*.
- For each of the six aforementioned sentiment lexicons:
  - *Number of*: positive words, negative words.
  - *Total score of*: positive words, negative words.
  - *Score of*: last token.
  - *Maximal score of*: positive words, negative words.

3.6 Miscellaneous

Some other features reported in the literature [2, 5, 13, 19, 24, 27] do not fit in any of the aforementioned categories. Thus, we have created this category to put these features together. They include the presence of abbreviations, the number of capitalized text, and the number of words in a tweet, as follows:

- *Whether the tweet has*: abbreviation.
- *Number of*: words, abbreviations, capitalized words, capital letters, words with all letters capitalized (all caps).

4 Experimental Evaluation

This section describes the computational experiments conducted to evaluate the different categories of features and meta-features introduced in the previous section, as well as the results of the application of the feature selection strategies on the full set of features. We first present the datasets of tweets used and then we describe the settings adopted in the computational experiments. Finally, we present the results and the discussions.

4.1 Twitter Datasets

We used a set of sixteen datasets in the computational experiments reported in this section. These datasets have been extensively used in the literature of Twitter sentiment analysis. To the best of our knowledge, this is the first study using a significant number of Twitter datasets in the evaluation of different types of features and meta-features that have already been employed in the literature. These datasets are: Irony [18], Sarcasm [18], Aisopos[4], SemEval-Fig[5], Sentiment140 [17], Person [11], Movie [11], Sanders[6], Narr [29], Obama-McCain Debate (OMD) [14], Health Care Reform (HCR) [38], STS-Gold [34], SentiStrength [39], Target-dependent [15], Vader [21], and SemEval13[7]. Some characteristics of these datasets are presented in Table 3, namely their total number of tweets, positive tweets and negative tweets.

4.2 Experimental Setting

In order to evaluate the different categories of features and meta-features described in Section 3, we applied the state-of-the-art machine learning algorithm Support Vector Machines (SVM), which

[4] http://grid.ece.ntua.gr
[5] http://alt.qcri.org/semeval2015/task11
[6] http://www.sananalytics.com/lab/twitter-sentiment
[7] https://www.cs.york.ac.uk/semeval-2013/task2.html

Table 3. Characteristics of the Twitter datasets.

Dataset	#tweets	#positive	#negative
Irony	65	22	43
Sarcasm	71	33	38
Aisopos	278	159	119
SemEval-Fig	321	47	274
Sentiment140	359	182	177
Person	439	312	127
Movie	561	460	101
Sanders	1,224	570	654
Narr	1,227	739	488
OMD	1,906	710	1,196
HCR	1,908	539	1,369
STS-Gold	2,034	632	1,402
SentiStrength	2,289	1,340	949
Target-dependent	3,467	1,734	1,733
Vader	4,196	2,897	1,299
SemEval13	4,378	3,183	1,195

has proven its robustness on large feature spaces [27]. In our experiments, we adopted the LIBSVM[8] [10] implementation of SVM for Weka [20]. The regularization parameter of LIBSVM was set to its default value (C = 1.0) and we adopted the linear kernel.

As a preprocessing step, we used the same strategy as done in [27]. First, for each tweet in a given dataset, we replaced URLs by the token "http://someurl" and user mentions by the token "@someuser". Then each tweet was tokenized and classified according to their part-of-speech tag, using the Twitter-specific part-of-speech tagset tool[9] [16]. This tagset consists of twenty-five POS tags, specifically designed for tweets, that takes into account the different aspects that tweets have as compared to regular text, such as the lack of conventional orthography and the 140-character limit of each message [16]. Regarding stopwords removal, we discarded stopwords only as unigram features, since it has been acknowledged that stopwords can affect the polarity of some expressions in higher order n-grams [38].

The features used in the computational experiments are exactly those already proposed in the literature, as introduced in Section 3. In this context, for the category N-grams, we used as features: unigrams, bigrams, trigrams, 4-grams, and 5-grams. Considering that negation words ("shouldn't", for example) can affect the n-gram-based features, we handle negation by employing the same approach as used by Mohammad et al. [27], as described in Subsection 3.5. We used the SentiWordNet lexicon to extract the features of the Polarity category related to the number of positive and negative adjectives, nouns, adverbs, and verbs. Regarding the features related to internet slang and emoticons, we used the internet slang dictionary and the emoticon dictionary introduced and used in [2]. Similarly, we used the Internet Lingo Dictionary [41] for abbreviations, as done in [24].

In the experimental evaluation, the predictive performance of the sentiment classification is measured in terms of classification accuracy. For each evaluated dataset, the accuracy of the classification was computed as the ratio between the number of correctly classified tweets and the total number of tweets, after a 10-fold cross validation.

4.3 Results and Discussion

In this section, we present the computational results obtained in the set of experiments performed in this study. The conducted experiments aimed to answer two main questions:

1. How effective are the different categories of features and meta-features identified in the literature in the task of sentiment classification of tweets?
2. Can the sentiment classification of tweets benefit from the application of feature selection methods on the full set of features?

[8] Available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
[9] `http://www.ark.cs.cmu.edu/TweetNLP`

4.3.1 Analysis of the Categories of Features and Meta-features

In order to answer the first question, we evaluated the performance of each individual category by using its features and meta-features to train an SVM classifier for each dataset. Table 4 shows the results of this evaluation, as well as the number of features of each category (presented in the *Number of features* row, except for the category N-grams, which are presented in the *#features* column). The bold-faced values indicate the best accuracies. As can be observed, the best accuracies were achieved by the categories Polarity ($f5$ column) and N-grams ($f1$ column). The category Polarity achieved better results in ten out of the 16 datasets, while the category N-gram performed better in six out of the 16 datasets. None of the other categories, namely Twitter and Microblog ($f2$ column), POS ($f3$ column), and Punctuation ($f4$ column), achieved meaningful results.

We can also notice from Table 4 that the worst accuracies achieved with the features of the category N-grams are referred to the datasets Irony, Sarcasm, and Aisopos. For the datasets Irony and Sarcasm, it may be due to the few number of tweets they contain, that is, 65 and 71, respectively. It seems that the n-gram-based features are not representative enough when employed individually in the sentiment classification of the tweets from these datasets, since the classification is performed based on the vocabulary extracted from the training set, that is, the n-grams themselves. Regarding the dataset Aisopos, although the n-gram-based features achieved better results as compared to the categories Twitter and Microblog, POS, and Punctuation, there is a great difference between the performances of the categories N-gram and Polarity. The accuracy achieved with the polarity-based features surpassed in more than 20% the accuracy achieved with the n-grams features in this dataset. It may be due to the great number of emoticons that the tweets of this dataset contain. Since the polarities of emoticons are taken into account in the features of the category Polarity, this information may have improved the classification when using the features of this category.

Table 4. Accuracies (in %) achieved by evaluating each category of features and meta-features.

Dataset	$f1$		$f2$	$f3$	$f4$	$f5$
	Acc.	#features	Acc.	Acc.	Acc.	Acc.
Irony	66.2	3,457	66.2	66.2	64.6	**84.6**
Sarcasm	47.9	3,540	54.9	54.9	57.8	**70.4**
Aisopos	67.3	12,657	57.9	65.1	54.0	**90.3**
SemEval-Fig	**88.2**	17,592	86.9	84.7	85.4	84.1
Sentiment140	**80.2**	16,003	50.1	55.7	57.9	78.6
Person	78.1	22,624	71.1	70.8	70.8	**79.3**
Movie	83.1	24,952	81.8	82.0	82.0	**83.8**
Sanders	**78.4**	52,091	57.6	64.5	57.5	75.9
Narr	80.0	50,181	60.2	65.4	61.5	**87.9**
OMD	**80.2**	73,249	62.8	62.3	62.8	75.6
HCR	**79.5**	98,199	72.0	72.0	71.6	71.9
STS-Gold	82.5	83,882	68.7	68.8	69.3	**89.4**
SentiStrength	70.2	110,212	60.8	61.9	58.8	**80.2**
Target-dependent	**81.5**	163,101	50.3	57.9	54.6	79.9
Vader	81.6	155,429	69.0	69.0	69.3	**87.9**
SemEval13	78.4	248,807	72.7	72.7	72.7	**84.5**
Number of features		–	9	25	10	54

$f1$ N-grams $f2$ Twitter and Microblog $f3$ POS $f4$ Punctuation $f5$ Polarity

Although the n-gram-based features were not effective in the sentiment classification of the tweets in some datasets, other datasets are benefited from the use of n-grams features, such as SemEval-Fig, OMD, and HCR. The tweets of these three datasets are considered as

challenging tweets in sentiment classification because of their nature, that are: figurative language (irony, sarcasm, and metaphor), politics, and both politics and health, respectively. Since n-gram-based features are able to capture more context than the other features, such as expressions of irony and sarcasm, and specific vocabulary used in politics and health domains, these features seem to be more appropriated to be used in the sentiment classification of this type of tweets.

Another point we can observe is that the worst accuracy achieved for the dataset SemEval-Fig is the one obtained using the features of the category Polarity. Since most of the tweets of this dataset are related to irony and sarcasm, it is possible that polarity features were not helpful in the classification as in this type of tweets the polarity is usually reversed.

Besides the evaluation of the importance of each individual category, we also inverted this evaluation. More specifically, we investigate how each category contributes to the set of all features, by removing one of the categories at a time from the full set of features and meta-features. The results of this evaluation are presented in Table 5. The information in parentheses represents the lost or gain in accuracy when one category is removed, when compared to the accuracy achieved with the set of all features (All column). As we can see, in general, by removing the categories N-gram ($All-f1$ column) and Polarity ($All-f5$ column) the accuracies dropped considerably. In fact, these are the only categories in which their removal from the full set caused losses in accuracies for all datasets. It is in accordance with the previous results, in which the evaluation of these categories in isolation also performed better than the other categories.

Table 5. Accuracies achieved by evaluating different sets of features.

Dataset	All	All−$f1$	All−$f2$	All−$f3$	All−$f4$	All−$f5$
Irony	80.0	78.5 (-1.5)	80.0 (0.0)	**81.5** (+1.5)	80.0 (0.0)	61.5 (-18.5)
Sarcasm	67.6	64.8 (-2.8)	**71.8** (+4.2)	**71.8** (+4.2)	69.0 (+1.4)	60.6 (-7.0)
Aisopos	92.8	87.4 (-5.4)	**93.5** (+0.7)	92.4 (-0.4)	93.2 (+0.4)	71.9 (-20.9)
Sem-Fig	**90.7**	85.7 (-5.0)	90.3 (-0.4)	**90.7** (0.0)	90.3 (-0.4)	88.5 (-2.2)
Sent140	81.9	79.1 (-2.8)	**82.7** (+0.8)	81.3 (-0.6)	**82.7** (+0.8)	73.5 (-8.4)
Person	84.1	78.8 (-5.3)	**85.0** (+0.9)	82.7 (-1.4)	84.1 (0.0)	75.4 (-8.7)
Movie	85.6	83.2 (-2.4)	**85.9** (+0.3)	85.0 (-0.6)	85.6 (0.0)	84.1 (-1.5)
Sanders	83.8	77.0 (-6.8)	84.2 (+0.4)	**84.6** (+0.8)	83.5 (-0.3)	79.6 (-4.2)
Narr	88.0	86.8 (-1.2)	**88.5** (+0.5)	87.8 (-0.2)	87.2 (-0.8)	77.6 (-10.4)
OMD	83.7	78.5 (-5.2)	**83.9** (+0.2)	**83.9** (+0.2)	83.4 (-0.3)	81.7 (-2.0)
HCR	**80.5**	74.2 (-6.3)	79.9 (-0.6)	80.1 (-0.4)	79.8 (-0.7)	79.6 (-0.9)
STS-Gold	**90.7**	88.8 (-1.9)	**90.7** (0.0)	90.2 (-0.5)	90.2 (-0.5)	84.4 (-6.3)
SentiStr.	81.0	80.6 (-0.4)	**81.1** (+0.1)	80.9 (-0.1)	80.6 (-0.4)	72.1 (-8.9)
Target-dep.	83.6	79.8 (-3.8)	**83.7** (+0.1)	**83.7** (+0.1)	**83.7** (+0.1)	81.3 (-2.3)
Vader	88.6	87.6 (-1.0)	88.7 (+0.1)	**89.0** (+0.4)	88.7 (+0.1)	82.3 (-6.3)
SemEval13	86.6	85.6 (-1.0)	**86.7** (+0.1)	86.4 (-0.2)	85.8 (-0.8)	80.3 (-6.3)

$f1$ N-grams $f2$ Twitter and Microblog $f3$ POS $f4$ Punctuation $f5$ Polarity

The removal of the category POS ($All-f3$ column) causes loss in accuracy in nine datasets, it was indifferent for only one, and it led to a better performance in six datasets. Similarly, by removing the category Punctuation ($All-f4$ column), the accuracy dropped in eight datasets, it led to slightly higher accuracies in five datasets, and it was indifferent for three datasets.

The category Twitter and Microblog ($All-f2$ column) seems to be the less important one. Removing its features and meta-features from the full set caused loss in accuracy only for the datasets SemEval-Fig and HCR. Considering that such datasets are more challenging than the others, as mentioned before, it is possible that the presence of the features of this category improves the overall classification accuracy. Differently, the absence of the category Twitter and Microblog in the classification was indifferent for two datasets (Irony and STS-Gold) and led to a better classification performance in twelve out of the 16 datasets. This is probably because the information in the features of this category is also captured by the features and meta-features of other categories, that is, this category may add redundancy or inconsistency to the classification for the most of the datasets.

4.3.2 Application of the Feature Selection Methods

As mentioned before, it is possible that some redundant or inconsistent features are inserted into the classification when working with the full set of features. To further investigate this issue, and in order to answer the second research question, the next series of experiments aims at minimizing the redundancy and noise that distinct features may insert into the classification, by selecting the most relevant features for each dataset. To this purpose, we applied three commonly used feature selection measures, namely Information Gain (IG), Chi-Squared (CHI), and Relief-F. The adopted feature selection strategy ranks the features based on these measures and select the top n most relevant features for the classification, according to a predefined threshold (n). In this study, we have varied the threshold values from 75% of the full set of features to the top 5 features. More precisely, for each feature selection measure, we have used the following threshold values: 75%, 50%, 25%, 10%, 1000, 500, 100, 50, 25, 10, and 5. The accuracies achieved from using the measures Information Gain, Chi-Squared, and Relief-F are presented in Table 6, Table 7, and Table 8, respectively. For space reasons, we only reported the results for the threshold values from the top 5 to the top 1000 features. Moreover, the application of the omitted thresholds did not achieve meaningful results.

In some cases, when the results obtained by the application of two consecutive threshold values are very approximate and among the best, we varied the threshold between these two values. For example, regarding the dataset Vader, since the accuracies achieved by the IG measure using the threshold values 500 and 1000 were the best ones and very approximate (89.9% and 89.6%, respectively), we applied an extra variance in which the best accuracy was achieved using the top 700 features (90.0%). Similarly, regarding the Chi-Squared measure, the best results achieved for datasets Aisopos (93.9%) and Narr (88.6%) are from using the extra variance with the top 15 and 80 features, respectively. For space reasons, we only report the results of this extra variance when it led to a better result.

Table 6. Accuracies achieved by using Information Gain measure in the classification.

Dataset	#features							
	1000	700	500	100	50	25	10	5
Irony	64.6	-	76.9	80.0	78.5	**83.1**	72.3	73.9
Sarcasm	69.0	-	70.4	69.0	71.8	73.2	**74.7**	**74.7**
Aisopos	88.9	-	88.5	87.4	90.3	**92.8**	**92.8**	91.0
Sem-Fig	88.8	-	88.8	88.5	89.4	90.3	**91.6**	**91.6**
Sent140	84.4	**85.0**	84.7	84.7	82.5	80.2	81.1	79.1
Person	81.6	-	81.1	82.0	**82.9**	82.2	78.4	74.5
Movie	86.1	-	86.5	**86.8**	84.7	84.7	86.1	85.6
Sanders	**83.1**	-	82.9	80.0	79.9	79.2	75.3	73.9
Narr	87.3	-	88.0	**88.4**	87.9	86.9	86.6	84.9
OMD	83.1	-	**84.3**	82.8	81.4	80.9	78.8	69.4
HCR	77.6	-	**78.3**	77.7	75.4	74.0	71.8	71.8
STS-Gold	89.4	-	90.0	**90.4**	89.7	90.0	89.8	90.2
SentiStr.	79.1	-	79.5	**81.4**	81.0	80.2	77.6	76.8
Target-dep.	82.5	-	**83.1**	**83.1**	81.5	80.2	78.3	76.6
Vader	89.6	**90.0**	89.9	89.3	87.9	87.4	84.0	82.1
SemEval13	86.0	-	**86.5**	86.1	85.0	84.4	83.0	81.1

In general, regarding the three feature selection methods, the best results were achieved by using from the top 50 to the top 1000 features. However, we can observe that the datasets Irony, Sarcasm, Aisopos, and SemEval-Fig benefited from using a more compact set of features, ranging from 5 to 50 features, in general.

We can also note that the results achieved with feature selection for the dataset HCR did not surpass the accuracy from using the full set

of features (80.5%), shown in the previous experiment (Table 5). This may be an indication that using various and different kinds of features and meta-features is beneficial for this dataset, given its challenging and diverse domain (both politics and health).

Table 7. Accuracies achieved by using Chi-Squared measure in the classification.

Dataset	#features								
	1000	500	100	80	50	25	15	10	5
Irony	64.6	72.3	76.9	-	76.9	**81.5**	-	75.4	78.5
Sarcasm	67.6	71.8	73.2	-	71.8	**74.7**	-	73.2	**74.7**
Aisopos	89.2	88.5	88.5	-	89.6	93.2	**93.9**	92.8	91.0
Sem-Fig	88.8	88.8	88.2	-	91.0	**91.9**	-	91.6	91.6
Sent140	85.0	**86.9**	84.7	-	81.9	80.5	-	81.1	80.2
Person	80.6	81.1	**83.6**	-	**83.6**	82.7	-	77.9	75.2
Movie	85.4	84.9	85.0	-	85.0	84.7	-	**85.7**	85.6
Sanders	83.0	**83.5**	79.5	-	80.2	79.2	-	75.7	74.4
Narr	86.8	87.4	88.2	**88.6**	88.0	86.2	-	86.0	84.8
OMD	83.0	**83.5**	83.2	-	81.2	80.7	-	77.5	71.0
HCR	**77.7**	77.5	77.3	-	75.4	74.3	-	72.2	71.8
STS-Gold	88.7	90.1	**90.4**	-	89.8	90.0	-	89.7	90.2
SentiStr.	78.9	79.0	**81.0**	-	80.9	80.4	-	77.5	76.5
Target-dep.	82.5	**83.0**	82.8	-	81.5	80.2	-	78.2	76.8
Vader	**89.8**	89.7	89.4	-	88.3	87.2	-	83.8	82.1
SemEval13	85.7	**86.8**	86.1	-	84.9	84.5	-	82.3	81.1

Table 8. Accuracies achieved by using Relief-F measure in the classification.

Dataset	#features						
	1000	500	100	50	25	10	5
Irony	**75.4**	70.8	63.1	64.6	64.6	70.8	73.9
Sarcasm	**71.8**	67.6	67.6	**71.8**	70.4	67.6	**71.8**
Aisopos	87.4	88.5	92.1	**92.8**	90.3	91.0	91.4
SemEval-Fig	86.0	88.8	**91.6**	91.3	91.3	91.3	**91.6**
Sentiment140	79.4	78.6	**84.7**	82.2	81.6	79.1	79.4
Person	80.0	**80.6**	78.6	79.5	78.1	76.5	70.2
Movie	81.6	83.8	**84.3**	83.4	82.7	82.4	82.4
Sanders	78.8	78.9	**80.7**	78.4	75.2	71.2	61.0
Narr	87.1	87.8	87.0	87.8	**88.2**	80.0	75.4
OMD	81.0	**81.4**	80.3	77.3	74.7	69.8	68.5
HCR	76.9	**77.2**	74.7	74.2	74.1	72.9	71.8
STS-Gold	**89.9**	89.6	89.6	87.9	87.1	86.7	85.8
SentiStrength	80.0	**81.4**	79.4	76.5	74.6	74.6	65.9
Target-dep.	**82.2**	82.0	81.6	80.8	80.6	77.5	74.9
Vader	**89.1**	89.0	87.9	86.6	85.2	80.3	78.8
SemEval13	**85.2**	84.2	72.9	72.9	72.8	72.8	72.9

The best results achieved by each feature selection method are summarized in Table 9, as well as the results of the best category (Table 4) and the best set (Table 5), presented in the previous experiments. As we can see, the application of feature selection methods, in an attempting to minimize the redundancy and inconsistency that distinct features and meta-features may insert into the classification, led to better classification accuracies in ten of out the 16 datasets used in this evaluation. It means that while some datasets benefit with the presence of all features and meta-features of some categories, due to the nature of the tweets they contain, for other datasets this may cause the addition of noise and redundant features in the classification, which is not beneficial.

Besides improving the classification performance for the most of the datasets, it is important to highlight that the feature selection methods can also significantly reduce the feature space, that is, the number of features used in the classification. For example, for the dataset Narr, regarding the best result (88.6%), the application of the Chi-Squared measure have reduced the feature space of this dataset to only 80 features, as can be seen in Table 7, in contrast with more than 50,000 features from the full set of features.

Aiming at reporting the most relevant features selected for some datasets, Table 10 and Table 11 present the top features for datasets Aisopos and SemEval-Fig, respectively. For space reasons, we are

Table 9. Comparison among the best results achieved by each feature selection method, by the best category, and by the best set, respectively.

Dataset	IG	CHI	Relief-F	Best category	Best set
Irony	83.1	81.5	73.9	**84.6**	81.5
Sarcasm	**74.7**	**74.7**	71.8	70.4	71.8
Aisopos	92.8	**93.9**	92.8	90.3	93.5
SemEval-Fig	91.6	**91.9**	91.6	88.2	90.7
Sentiment140	85.0	**86.9**	84.7	80.2	82.7
Person	82.9	83.6	80.6	79.3	**85.0**
Movie	**86.8**	85.7	84.3	83.8	85.9
Sanders	83.1	83.5	80.7	78.4	**84.6**
Narr	88.4	**88.6**	88.2	87.9	88.5
OMD	**84.3**	83.5	81.4	80.2	83.9
HCR	78.3	77.7	77.2	79.5	**80.5**
STS-Gold	90.4	90.4	89.9	89.4	**90.7**
SentiStrength	**81.4**	81.0	**81.4**	80.2	81.1
Target-dep.	83.1	83.0	82.2	81.5	**83.7**
Vader	**90.0**	89.8	89.1	87.9	89.0
SemEval13	86.5	**86.8**	85.2	84.5	86.7

presenting the selected features for these two datasets only. For both datasets, we show the top features selected by the Chi-Squared measure, which achieved the best results for these datasets (Table 9). For presentation purposes, the features of each category are presented in the format <*category*>_<*featureName*>, wherein *category* can be NGRAM, TWITTER, POS, PUNC, and POL, representing the categories N-grams, Twitter and Microblog, Part-of-Speech, Punctuation, and Polarity, respectively.

Analyzing the selected features for the dataset Aisopos (Table 10), we can see that the top 4 features are related to emoticons. As mentioned before, this dataset contains a great number of emoticons. Among the 119 negative tweets of this dataset, 97 tweets contain at least one negative emoticon. Differently, none of the 159 positive tweets contain any negative emoticons. Moreover, among those 97 negative tweets that contain negative emoticons, in 67 of them the emoticons appear as the last token. For this reason, the most discriminative negative features for this dataset are *POL_hasNegativeEmoticon*, *POL_numberOfNegativeEmoticons*, and *POL_isLastTokenNegativeEmoticon*.

Table 10. Top 15 features selected for dataset Aisopos.

Ranking	CHI	Feature
1	191.95	POL_hasNegativeEmoticon
2	191.95	POL_numberOfNegativeEmoticons
3	117.95	POL_isLastTokenNegativeEmoticon
4	115.58	POL_hasPositiveEmoticon
5	94.10	POL_totalScoreOfNegativeWordsInSent140
6	89.86	POL_maximalScoreOfNegativeWordsInSent140
7	84.70	POL_numberOfPositiveEmoticons
8	68.21	POL_totalScoreOfPositiveWordsInSent140
9	66.24	POL_isLastTokenPositiveEmoticon
10	62.01	POL_maximalScoreOfPositiveWordsInSent140
11	43.37	POL_sumOfScoresOfAdjAdvVerbNounFromSWN
12	40.97	POL_numberOfNegativeWordsInSent140
13	31.54	POL_totalScoreOfNegativeWordsInSWN
14	30.11	POL_numberOfPositiveWordsInSent140
15	29.40	POL_maximalScoreOfNegativeWordsInSWN

Regarding the dataset SemEval-Fig (Table 11), we can notice that 18 out of the 25 ranked features are from the N-grams category. It is consistent with the previous results reported for this dataset (from Table 4), in which the best accuracy was achieved when using the n-gram-based features in the classification. Among the n-grams, we can see the unigrams "#not" and "#sarcasm", which are hashtags commonly used in tweets to express irony and sarcasm, respectively. We can also notice the trigram "pretty little liars", which may be used as an expression of sarcasm. For this dataset, the most discriminative positive feature is the unigram "literally", since this unigram has 21 occurrences among the 47 positive tweets of this dataset, and

occurs only eight times among the 274 negative tweets. The features *POS_numberOfHashtags* and *TWITTER_hasHashtag* are the most discriminative negative features for this dataset. This is probably because among the 274 tweets that contain hashtags, about 250 are related to negative tweets.

Table 11. Top 25 features selected for dataset SemEval-Fig.

Ranking	CHI	Feature
1	131.2	NGRAM_literally
2	72.9	POS_numberOfHashtags
3	72.9	TWITTER_hasHashtag
4	52.1	NGRAM_good
5	29.5	NGRAM_pretty
6	28.9	NGRAM_have a
7	28.8	POL_sumOfScoresOfAdjAdvVerbNounFromSWN
8	23.1	NGRAM_pretty little
9	22.9	POL_maximalScoreOfPositiveWordsInSWN
10	21.9	NGRAM_a good
11	21.4	NGRAM_#not
12	20.1	NGRAM_little
13	19.1	POL_numberOfPositiveWordsInSent140
14	19.0	NGRAM_one
15	18.0	NGRAM_#sarcasm
16	17.7	NGRAM_happy birthday
17	17.7	NGRAM_literally just
18	17.7	NGRAM_such a good
19	17.4	NGRAM_liars
20	17.4	NGRAM_pretty little liars
21	17.4	NGRAM_hilarious
22	17.4	NGRAM_little liars
23	15.9	POL_numberOfPositiveWordsInOpinionFinder
24	15.9	POL_scoreOfLastTokenInNRCHashtag
25	13.2	NGRAM_so good

4.3.3 Comparison with Results Reported in the Literature

In order to investigate the competitiveness of the computational results achieved in this work, we compared them with the results achieved in recent works in the literature. Specifically, we aim at comparing the classification performance using the most relevant features and meta-features we have identified using feature selection with the best results we found in recent works in the literature for this task. For space reasons, although we have found many works to compare with, we report only their best results. Moreover, we could not find any work to compare for some datasets because they have been used in the three-class classification problem, that is, the sentiment classification problem regarding the positive, negative, and neutral classes. Since we focus on the polarity classification (positive and negative classes, only), the results reported in such works are not comparable. The comparison is presented in Table 12.

Table 12. Comparison between the best results achieved in this work and the best results reported in the literature, in terms of accuracy and microF$_1$.

Dataset	Accuracy		MicroF$_1$	
	Our results	Best in literature	Our results	Canuto et al. [9]
Aisopos	–	–	**94.8**	89.2
Sentiment140	**86.9**	86.3 [35]	**87.1**	86.9
Sanders	84.6	**98.1** [8]	83.4	**86.5**
Narr	**88.6**	81.3 [29]	**90.5**	88.8
OMD	**84.3**	82.9 [36]	76.8	**80.0**
HCR	**80.5**	78.7 [34]	–	–
STS-Gold	**90.7**	85.7 [34]	–	–
SentiStrength	**81.4**	73.4 [34]	**84.3**	82.6
Vader	–	–	92.9	**97.2**
SemEval13	–	–	**91.1**	85.8
Win count	6	1	5	3

The results obtained in this work and the best results found in the literature are presented in terms of classification accuracy, except for the results achieved by Canuto et al. [9], which are presented in terms of microF$_1$. For this reason, we split the table in two parts. The first part shows the comparison among the results in terms of accuracy. In the second part, we present the comparison between the results in terms of microF$_1$. We also show, in the last row (*Win count* row), the number of times that each compared work achieved the best results.

As we can see, the computational results achieved in this work, using different kinds of features and meta-features, are the best in six out of the seven datasets, regarding the results reported in terms of classification accuracy. However, for dataset Sanders, the best result was achieved by Bravo-Marquez et al. [8]. Regarding the comparison with the results reported in terms of microF$_1$, the results achieved in this work were among the best in five out of the eight compared datasets. Similarly to the results presented in terms of accuracy, we did not achieve the best result for dataset Sanders. These results confirm the importance of selecting the appropriate set of features in the context of Twitter sentiment analysis.

5 Conclusions and Future Work

In this work, we presented a literature review of the most common feature representation in the sentiment classification of tweets, including meta-features. We proposed to group these features and meta-features in specific categories, in order to evaluate the importance of each category in the polarity classification of tweets from distinct domains. These categories include N-grams, Twitter and Microblog, Part-of-Speech, Punctuation, and Polarity. We used sixteen datasets of tweets in the series of experiments reported in this study. To the best of our knowledge, this is the first work that evaluates distinct categories of features for a significant number of Twitter datasets.

Our experiments showed that the categories Polarity and N-grams are the most important ones, achieving the best results. Indeed, when considering the full set of features, the removal of the features from these two categories made the performance drop considerably for all datasets. We could also notice that some datasets, such as HCR, benefited from the presence of the full set of features in the classification. This may be an indication that tweets from challenging domains need to be represented by all types of features.

We also applied feature selection strategies in order to select the most relevant features for the classification. This set of experiments showed that tweets from distinct domains can benefit from using different subsets of features in the classification. Finally, we compared the results achieved in this work with the best results previously reported in the literature for some datasets. This comparison confirmed we have achieved meaningful results by evaluating different categories of features and also using feature selection strategies.

For future work, we intend to examine and incorporate new features from more recent studies, such as the meta-features recently proposed by Canuto et al. [9]. Another idea of future work is the application of lazy feature selection methods, based on the hypothesis that knowing the values of the features of a particular tweet at classification time may contribute to identify the best features for the correct classification of that specific tweet [32].

ACKNOWLEDGEMENTS

This work was supported by CAPES and CNPq research grants.

REFERENCES

[1] A. Abbasi, H. Chen, and A. Salem, 'Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums', *ACM Transactions on Information Systems*, **26**(3), 12:1–12:34, (2008).

[2] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, 'Sentiment analysis of Twitter data', in *Proceedings of the Workshop on Languages in Social Media*, pp. 30–38. ACL, (2011).

[3] B. Agarwal and N. Mittal, 'Optimal feature selection for sentiment analysis', in *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 13–24. Springer, (2013).

[4] S. Baccianella, A. Esuli, and F. Sebastiani, 'SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining', in *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pp. 2200–2204, (2010).

[5] L. Barbosa and J. Feng, 'Robust sentiment detection on Twitter from biased and noisy data', in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 36–44. ACL, (2010).

[6] A. Bermingham and A.F. Smeaton, 'Classifying sentiment in microblogs: is brevity an advantage?', in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pp. 1833–1836. ACM, (2010).

[7] A. Bifet and E. Frank, 'Sentiment knowledge discovery in Twitter streaming data', in *Proceedings of the 13th International Conference on Discovery Science*, pp. 1–15. Springer-Verlag, (2010).

[8] F. Bravo-Marquez, M. Mendoza, and B. Poblete, 'Meta-level sentiment models for big social data analysis', *Knowledge-Based Systems*, **69**, 86–99, (2014).

[9] S. Canuto, M.A. Gonçalves, and F. Benevenuto, 'Exploiting new sentiment-based meta-level features for effective sentiment analysis', in *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, pp. 53–62. ACM, (2016).

[10] C. Chang and C. Lin, 'LIBSVM: A library for support vector machines', *ACM Transactions on Intelligent Systems and Technology*, **2**, 1–27, (2011).

[11] L. Chen, W. Wang, M. Nagarajan, S. Wang, and A.P. Sheth, 'Extracting diverse sentiment expressions with target-dependent polarity from Twitter', in *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, pp. 50–57, (2012).

[12] N.F.F. da Silva, E.R. Hruschka, and E.R. Hruschka Jr., 'Tweet sentiment analysis with classifier ensembles', *Decision Support Systems*, **66**, 170–179, (2014).

[13] D. Davidov, O. Tsur, and A. Rappoport, 'Enhanced sentiment learning using Twitter hashtags and smileys', in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 241–249. ACL, (2010).

[14] N.A. Diakopoulos and D.A. Shamma, 'Characterizing debate performance via aggregated Twitter sentiment', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1195–1198. ACM, (2010).

[15] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, 'Adaptive recursive neural network for target-dependent Twitter sentiment classification', in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: short papers*, pp. 49–54. ACL, (2014).

[16] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N.A. Smith, 'Part-of-speech tagging for Twitter: annotation, features, and experiments', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, pp. 42–47. ACL, (2011).

[17] A. Go, R. Bhayani, and L. Huang, 'Twitter sentiment classification using distant supervision', Technical Report CS224N, Stanford, (2009).

[18] P. Gonçalves, D. Dalip, J. Reis, J. Messias, F. Ribeiro, P. Melo, M. Gonçalves, and F. Benevenuto, 'Caracterizando e detectando sarcasmo e ironia no Twitter', in *Proceedings of the Brazilian Workshop on Social Network Analysis and Mining*, (2015).

[19] M. Hagen, M. Potthast, M. Büchner, and B. Stein, 'Twitter sentiment detection via ensemble classification using averaged confidence scores', in *Proceedings of the 37th European Conference on IR Research*, pp. 741–754. Springer, (2015).

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, 'The weka data mining software: an update', *SIGKDD Explorations Newsletter*, **11**(1), 10–18, (2009).

[21] C.J. Hutto and E. Gilbert, 'Vader: A parsimonious rule-based model for sentiment analysis of social media text', in *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, (2014).

[22] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, 'Target-dependent Twitter sentiment classification', in *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*, pp. 151–160. ACL, (2011).

[23] V.N. Khuc, C. Shivade, R. Ramnath, and J. Ramanathan, 'Towards building large-scale distributed systems for Twitter sentiment analysis', in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pp. 459–464. ACM, (2012).

[24] E. Kouloumpis, T. Wilson, and J. Moore, 'Twitter sentiment analysis: The good the bad and the omg!', in *Proceedings of the 5th International AAAI Conference on Web and Social Media*, pp. 538–541, (2011).

[25] J. Lin and A. Kolcz, 'Large-scale machine learning at Twitter', in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pp. 793–804. ACM, (2012).

[26] B. Liu, 'Sentiment analysis and opinion mining', *Synthesis Lectures on Human Language Technologies*, **5**(1), 1–167, (2012).

[27] S. Mohammad, S. Kiritchenko, and X. Zhu, 'Nrc-canada: building the state-of-the-art in sentiment analysis of tweets', in *Proceedings of the 7th International Workshop on Semantic Evaluation Exercises*, Atlanta, Georgia, USA, (2013).

[28] S.M. Mohammad and P.D. Turney, 'Crowdsourcing a word–emotion association lexicon', *Computational Intelligence*, **29**(3), 436–465, (2013).

[29] S. Narr, M. Hulfenhaus, and S. Albayrak, 'Language-independent Twitter sentiment analysis', in *Proceedings of the Workshop on Knowledge Discovery, Data Mining and Machine Learning*, (2012).

[30] A. Pak and P. Paroubek, 'Twitter as a corpus for sentiment analysis and opinion mining', in *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pp. 1320–1326, (2010).

[31] B. Pang, L. Lee, and S. Vaithyanathan, 'Thumbs up? sentiment classification using machine learning techniques', in *Proceedings of the ACL-02 conference on Empirical Methods in Natural Language Processing*, pp. 79–86. ACL, (2002).

[32] R.B. Pereira, A. Plastino, B. Zadrozny, L.H.C. Merschmann, and A.A. Freitas, 'Lazy attribute selection: Choosing attributes at classification time', *Intelligent Data Analysis*, **15**(5), 715–732, (2011).

[33] J.D. Prusa, T.M. Khoshgoftaar, and D.J. Dittman, 'Impact of feature selection techniques for tweet sentiment classification', in *Proceedings of the 28th International FLAIRS Conference*, (2015).

[34] H. Saif, M. Fernandez, Y. He, and H. Alani, 'Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold', in *Proceedings of the 1st Workshop on Emotion and Sentiment in Social and Expressive Media*, (2013).

[35] H. Saif, Y. He, and H. Alani, 'Alleviating data sparsity for Twitter sentiment analysis', in *Proceedings of the 2nd Workshop on Making Sense of Microposts*, pp. 2–9. CEUR-WS, (2012).

[36] H. Saif, Y. He, M. Fernandez, and H. Alani, 'Semantic patterns for sentiment analysis of Twitter', in *Proceedings of the 13th International Semantic Web Conference*, pp. 324–340. Springer, (2014).

[37] A. Sharma and S. Dey, 'A comparative study of feature selection and machine learning techniques for sentiment analysis', in *Proceedings of the 2012 ACM Research in Applied Computation Symposium*, pp. 1–7. ACM, (2012).

[38] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge, 'Twitter polarity classification with label propagation over lexical links and the follower graph', in *Proceedings of the 1st Workshop on Unsupervised Learning in NLP*, pp. 53–63. ACL, (2011).

[39] M. Thelwall, K. Buckley, and G. Paltoglou, 'Sentiment strength detection for the social web', *Journal of the American Society for Information Science and Technology*, **63**(1), 163–173, (2012).

[40] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, 'A system for real-time Twitter sentiment analysis of 2012 US presidential election cycle', in *Proceedings of the ACL 2012 System Demonstrations*, pp. 115–120. ACL, (2012).

[41] L. Wasden, 'Internet lingo dictionary: A parents' guide to codes used in chat rooms, instant messaging, text messaging, and blogs', Technical report, Idaho Office of the Attorney General, (2010).

[42] M. Wiegand, A. Balahur, B. Roth, D. Klakow, and A. Montoyo, 'A survey on the role of negation in sentiment analysis', in *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pp. 60–68. ACL, (2010).

[43] T. Wilson, J. Wiebe, and P. Hoffmann, 'Recognizing contextual polarity in phrase-level sentiment analysis', in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 347–354. ACL, (2005).

[44] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, 'Combining lexicon-based and learning-based methods for Twitter sentiment analysis', Technical Report HPL-2011-89, HP Laboratories, (2011).