On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages

David Vilares (corresponding author)

Research Group on Language and Information Society, Departamento de Computación, Universidade da Coruña, Campus de Elviña, 15071, A Coruña, Spain, E-mail: david.vilares@udc.es

tel. +34 981 167 000 ext.1302

Fax. +34 981 167 160

Miguel A. Alonso

Research Group on Language and Information Society, Departamento de Computación, Universidade da Coruña, Campus de Elviña, 15071, A Coruña, Spain, E-mail: miguel.alonso@udc.es

tel. +34 981 167 000 ext. 1338

Fax. +34 981 167 160

Carlos Gómez-Rodríguez

Research Group on Language and Information Society, Departamento de Computación, Universidade da Coruña, Campus de Elviña, 15071, A Coruña, Spain, E-mail: carlos.gomez@udc.es

tel. +34 981 167 000 ext.1377

Fax. +34 981 167 160

Abstract

Millions of micro texts are published every day on Twitter. Identifying the sentiment present in them can be helpful for measuring the frame of mind of the public, their satisfaction with respect to a product or their support of a social event. In this context, polarity classification is a subfield of sentiment analysis focussed on determining whether the content of a text is objective or subjective, and in the latter case, if it conveys a positive or a negative opinion. Most polarity detection techniques tend to take into account individual terms in the text and even some degree of linguistic knowledge, but they do not usually consider syntactic relations between words. This article explores how relating lexical, syntactic and psychometric information can be helpful to perform polarity classification on Spanish tweets. We provide an evaluation for both shallow and deep linguistic perspectives. Empirical results show an improved performance of syntactic approaches over pure lexical models when using large training sets to create a classifier, but this tendency is reversed when small training collections are used.

Introduction

Analysing and comprehending subjective information expressed in social media by users of different countries, cultures and ages has become a key asset in order to monitor public opinion about products, events or services. Before the appearance of the Web 2.0, a common solution for obtaining information about a topic was using surveys and polls. However, these strategies were typically expensive, had a limited scope and were only valid for a short period of time. Currently, social media could provide an effective way to poll users (Wang, Can, Kazemzadeh, Bar & Narayanan, 2012), plan business strategies (Li & Li, 2013) and make marketing decisions (Bae & Lee, 2012). However, human monitoring of web reviews presents important obstacles. The vast amount of opinions expressed every day in blogs, forums or social networks makes manual observation unfeasible. In addition, Pang, Lee and Vaithyanathan (2002) proposed an experiment to show how applying corpus-based techniques to extract good sentiment features presents some advantages with respect to relying on intuitions, such as exhaustiveness of the resulting list of subjective words and the capacity to capture implicitly subjective constructions. In this context, *sentiment analysis* (SA) has arisen as a field of research that deals with the automatic analysis of subjective content (Pang & Lee, 2008; Liu 2012; Feldman, 2013). At present, sentiment analysis is also known as *opinion mining* (OM), although this term was initially associated with web search and information retrieval whilst sentiment analysis referred to the automatic analysis of subjective texts. Following Pang and Lee (2008), we are using these terms interchangeably. Many subtasks can be located within this field of research. The most popular one is *classifying the sentiment* or *polarity* of a text as positive or negative, although it is also common to include additional categories to distinguish purely informative texts, and to differentiate the strength of the opinions.

Sentiment Analysis has become a very active field of research in the last decade. Nearly all SA approaches have focussed on long reviews (Pang et al., 2002, Turney, 2002), mainly from forums such as epinions[1] or tripadvisor[2]. However, the recent success of micro-blogging social networks, such as MySpace, Facebook, and remarkably Twitter; has increased the interest in monitoring short texts (Thelwall, Buckley & Paltoglou, 2012; Martínez-Cámara, Martín-Valdivia, Ureña-López & Montejo-Ráez, 2013). Twitter is a micro-blogging social network where users share their views, experiences or simply trivia in messages (called *tweets*) of up to 140 characters. At present, this social medium has 215 million monthly active users and more than 100 million daily active users. These users, located all around the world, include influential individuals and organizations, such as world leaders, government officials, celebrities, athletes, journalists, sports teams, media outlets and brands, which create approximately 500 million tweets every day (Twitter, 2013).

Although major efforts in SA have focussed on tweets in the English language, currently less than 50% of tweets are written in English, with a significant and growing presence of Spanish, Portuguese and Japanese (Carter, Weerkamp & Tsagkias, 2013). Thus, monitoring opinions in such languages is crucial in order to obtain a global vision.

In this article, we propose several methods for classifying the polarity of Spanish tweets by using linguistic knowledge. The use of linguistic approaches on Twitter is a debatable issue given the limited number of characters allowed per message and the presence of non-grammatical elements. The main contribution of the article consists of building models which combine lexical, syntactic, psychometric and semantic knowledge to illustrate the performance that linguistic perspectives can achieve, ranging from shallow to deep knowledge. In particular, generalised dependency triplets, a syntactic feature representation originally used by Joshi and Penstein-Rosé (2009) for identification of opinionated sentences on long reviews, are adapted and enriched to carry out polarity classification tasks over different sets of classes. We also explore how the size of the training set is relevant to properly exploit different linguistic-based models. In addition, an existent symbolic analyser proposed by Vilares, Alonso & Gómez-Rodríguez (2013), initially intended for long reviews, is used to enrich the models described above.

We also undertake a wide experimental evaluation, suggesting that a syntactic perspective outperforms pure lexical-based methods if the training collection is large enough. Most of the results only focus on classifying tweets as positive, negative or objective, but we also provide some conclusions regarding a finer classification that distinguishes sentiment strength.

The remainder of the paper is organised as follows. The next section presents the background and related research on the topic at hand, polarity classification. Next, we motivate the research and we detail the foundations of our lexical and syntactic approaches to polarity classification. We then show and discuss empirical results. Finally, we present our conclusions and introduce future research directions.

Background and related work

Polarity classification has mainly been tackled from two different perspectives, namely the supervised (Pang et al., 2002) and semantic (Turney, 2002) approaches.

Supervised polarity classification

Supervised methods for polarity classification are characterized by using machine learning (ML) techniques to classify the sentiment of a text as positive or negative. Pang et al. (2002) introduce this approach in order to classify documents by their overall sentiment instead of by topic. They show the effectiveness of standard machine learning techniques on movie reviews, using unigrams, bigrams and part-of-speech as features. Moreover, they indicate that automatic sentiment classification seems to outperform human-generated results.

Within the ML perspective, there are other approaches that try to use deep linguistic knowledge. Gamon (2004) evaluates the role of linguistic features such as PoS-tag tri-grams and constituent structure of phrases in sentiment classification. Empirical results show that, although features of this kind obtain a low performance by themselves, they contribute positively to accuracy when they are included in word n-gram models. In a similar line, Joshi and Penstein-Rosé (2009) explore the effectiveness of dependency-based features on identifying opinionated sentences. They introduce the concept of *composite back-off features*, or *composite generalised features*; which is the term we are using to refer this method in the paper: given a dependency triplet of the form (head_i, arc_{ij}, dependent_j) they propose generalising either the head or the dependent to their respective part-of-speech tag. It is important to note that (head_i, arc_{ij}, dependent_j) terms are called triplets in the literature related to dependency parsers, although they really represent pairs of words connected by a dependency type. For example, in the sentences '*He is a smart boy*' and '*It's a smart washing machine*', their approach generalises the triplets (boy, modifier, smart) and (washing-machine, modifier, smart) to a single triplet of the form (noun, modifier, smart). In this way, two triplets that have the same meaning are unified into one, while relations can still be captured. Their approach obtains a statistically significant improvement when some of these generalised features are used in conjunction with word unigrams.

Specifically, they obtained the best performance when applying generalisation over the head term. They concluded generalising the head is a better option because makes it possible to identify patterns about opinions about products, features or services. The dependency type does not play a role, in terms of generalisation, in the work by Joshi and Penstein-Rosé (2009). In any case, keeping information about the dependency type which connects a pair of words could be useful, as a way to capture how people connect terms.Greene and Resnik (2009) introduce *observable proxies for underlying semantics* to approximate the relevant semantic properties automatically as features in a supervised learning setting, on the basis that the connection between structure and implicit sentiment is mediated by semantic properties characterizing the interface between syntax and lexical semantics. However, their experiments are not directly comparable to conventional labelling for opinionated tests.

Wu, Zhang, Huang and Wu (2009) define a phrase-based dependency parsing approach and propose a tree-kernel based SVM as a model for polarity classification. Nakagawa, Inui and Kurohashi (2010) also employ dependency trees for sentiment classification. The authors represent the polarity of each dependency sub tree by a hidden variable and perform sentiment classification by means of Conditional Random Fields to finally compute the polarity of the whole sentence. More recently, Socher et al. (2013) proposed a new model based on recursive neural tensor networks. To train their model, the authors take the corpus of movie reviews presented in Pang and Lee (2005) and parse the sentences, to finally rely on the Amazon Mechanical Turk crowdsourcing service to manually label the resulting phrases. Empirical results show that this model outperforms the accuracy of a pure bag-of-features approach, reinforcing the idea that using the syntactic structure of sentences to capture context is helpful in SA.

The main drawback of supervised classifiers is their high domain dependency. Systems of this kind do not perform well when a single classifier is used across different domains, to the point that their accuracy can even drop almost to chance level (Brooke, Tofiloski & Taboada, 2009). The issue is that ML methods excel at learning the perception of a word for a specific domain, but the sentiment of that term can be different in other areas. Moreover, training a classifier usually requires creating large data sets so that an acceptable model can be built. In this respect, Sidorov et al. (2013) explore how different settings such as corpus size or evaluation over different domains affect the precision of several supervised classifiers. Self-training has also been proposed as a solution to annotate a large opinionated corpus (He & Zhou, 2011). In this framework, an initial classifier is learned by incorporating prior information extracted from an existing sentiment lexicon. After running this classifier on unlabelled data, documents that have been classified with high confidence are used as pseudo-labelled examples for automatic domain-specific feature acquisition.

Unsupervised polarity classification

Unsupervised approaches are characterised by the use of semantic orientation (SO) dictionaries or opinion lexicons (Devitt & Ahmad, 2013). To classify polarity, these methods obtain the subjective expressions present in a text and aggregate their SO in a given way. Turney (2002) was one of the first to apply this angle. He performs a classification over movie reviews by means of an algorithm which handles the semantic orientation of subjective phrases. To calculate the semantic orientation of adjective and adverb phrases he uses PMI-IR (Turney, 2001), which measures the mutual information of a phrase with respect to the words 'excellent' and 'poor'.

An interesting lexicon-based approach is The Semantic Orientation CALculator (Taboada, Brooke, Tofiloski, Voll & Stede, 2011). This system deals with relevant linguistic phenomena in SA such as intensification or negation, identifying their scope of influence by defining rules based on shallow linguistic knowledge. The semantic knowledge of the system relies on manually annotated opinion lexicons. The initial approach was proposed to analyse English reviews, but they also adapted their method to work on Spanish language texts (Brooke et al., 2009). The authors also highlight the importance of irrealis mood on polarity classification tasks, ignoring the sentiment reflected by these sentences. However, some types of irrealis such as conditional relations can contribute to evaluation in discourse, as pointed out in Trnavac and Taboada (2012).

Syntactic and semantic knowledge have also been applied to unsupervised polarity classification. Moilanen and Pulman (2007) argue that it is possible to calculate, in a systematic way, the polarity values of larger constituents as some function of the polarities of their smaller constituents, in a way analogous to the principle of compositionality from the formal semantics literature. However, they found that, even in an ideal situation with a clean input, their model would fail to solve almost 20% of the cases, in which further information is required. Shaikh, Prendinger and Mitsuru (2007) integrate semantic processing of input texts by dependency analysis on semantic verb-frames, following a rule-based approach to assess the valence of each semantic verb-frame in a sentence, claiming results comparable to state-of-the-art systems at that time.

Semantic-based approaches also present some disadvantages. The creation of manual opinion lexicons is often expensive, and as a result we obtain dictionaries with a low recall (Zhang, Ghosh, Dekhil, Hsu, & Liu, 2011). These dictionaries can make it possible to obtain a decent baseline on different domains, but they are unable to consider the specific subjective elements for a particular field and social medium. A possible solution is to try to create polarity dictionaries automatically by means of annotated data sets, but, as in the case of supervised methods, such dictionaries would be dependent on the domain (Vilares et al., 2013).

Hybrid approaches to polarity classification

Hybrid approaches combine lexicon-based and machine learning techniques to fill the gaps that each of these approaches presents when used separately. Kennedy and Inkpen (2006) first introduce a term-counting approach, showing the utility of including contextual valence shifters. Then, the authors combine their method with an ML model based on bag-of-words, also taking into account contextual valence shifters by means of bigrams. Choi and Cardie (2008) consider the effect of interaction among words or constituents in the light of compositional semantics, presenting a learning-based approach that incorporates structural inference into the learning procedure. They find that negation plays an important role in determining expression-level polarity and that classification accuracy uniformly decreases as additional, potentially disambiguating, context is applied. Zhang et al. (2011) propose a lexicon-based approach to achieve high precision on sentiment analysis at the entity level. They then use the output labels of the unsupervised system to train a supervised classifier, improving both recall and F-measure. Perea-Ortega, Martín-Valdivia, Ureña-López and Martínez-Cámara (2013) propose a different hybrid strategy: a voting model based on majority rule to conflate the results obtained by an SVM trained on word n-grams with those obtained by a lexicon-based approach. The system was used to classify a set of film reviews from a bilingual parallel corpus as positive or negative. Experimental results show a slight improvement over pure machine learning approaches. Arora, Mayfield, Penstein-Rosé and Nyberg (2010) employ genetic algorithms to construct complex features from subgraphs extracted from an annotation graph. A constant number of these features are added to a unigram feature space, obtaining a small but consistent increase in performance. Going beyond syntax, Heerschop, Goossen, Hogenboom, Frasincar, Kaymak and de Jong (2011) propose to apply Rhetorical Structure Theory to obtain the discourse structure of texts, weighting the sentiment conveyed by distinct text spans in accordance with their importance. The weights are optimized by genetic algorithms, obtaining an improvement in accuracy with respect to baseline models that do not take discourse structure into account.

Finally, it is interesting to note that in the field of information retrieval, syntactic and semantic information has also been applied to find documents that express an opinion about a given query. In this context, Guo and Wan (2012) incorporate the syntactic tree structure of a sentence into a probabilistic retrieval model, evaluating the modifying probability between an opinion term and a noun within the sentence in order to capture query-related opinion scores more accurately.

Polarity classification in Twitter

The use of linguistic knowledge has been successfully applied on sentiment analysis, as detailed above. However, most of this research has focussed on long reviews, where users have enough space to express their views. This means that a broad context is available, from which it is possible to extract a large amount of information about phrase and discourse structure. But there is not too much related work about how the use of deep linguistic knowledge can improve accuracy on micro-opinions, especially for languages that present additional linguistic challenges with respect to English, such as European Romance languages (Boiy & Moens, 2009) or Chinese (Zhang, Zeng, Li, Wang & Zuo, 2009) (e.g., the more complex morphology and freer syntactic word order of Spanish; or the segmentation ambiguities, the subtlety and ambiguity of adverbs, and the complex dependency relation among words in Chinese).

Montejo-Ráez, Martínez-Cámara, Martín-Valdivia and Ureña-López (2012) define an unsupervised approach based on a random walk algorithm that weighs synsets (sets of synonymous words) from tweets with polarity scores provided by SentiWordNet (Baccianella, Esuli & Sebastiani, 2010), a lexical resource based on WordNet (Miller, Beckwith, Fellbaum, Gross & Miller, 1990) that maps each synset to a set of scores representing its notions of positivity, negativity and neutrality.

Pak and Paroubek (2010) propose a sentiment classifier which uses word n-grams and PoS tags to carry out ternary categorisation in Twitter; differentiating positive, negative and purely objective texts. In addition, the authors stress the importance of including a preliminary step for predicting whether a tweet is subjective, to then classify subjective texts as positive or negative. For this purpose, Batista and Ribeiro (2013) employ a cascade of binary maximum entropy classifiers for multiple polarity classification for the Spanish language. The knowledge of the system relies on a pure bag-of-features, and it seems to improve over the accuracy of similar systems that do not use a hierarchical structure of classifiers. Unigrams, bigrams and PoS tags are also used as features for a Bayesian classifier by Spencer, J. and Uchyigit, G. (2012). Positive unigrams, negative unigrams, positive bigrams, negative bigrams, and Twitter specific elements such as emoticons, hashtags and URLs; are used by Bakliwal, Arora, Madhappan, Kapre, Singh and Varma (2012) as features for training a SVM classifier that obtains good results on several English datasets.

Thelwall, Buckley, Paltoglou, Cai and Kappas (2010) define SentiStrength, a machine learning approach to optimise sentiment term weightings that exploit repeated-letter non-standard spelling for extracting sentiment. Subsequently, Thelwall, Buckley and Paltoglou (2010) use it to study a one-month stream of tweets, obtaining strong evidence that popular events are normally associated with increases in negative sentiment strength and some evidence that peaks of interest in events show stronger positive sentiment than the time periods before the peak.

Jiang, Yu, Zhou, Liu, and Zhao (2011) propose combining target-independent and target-dependent features to improve the performance of polarity classification on Twitter. To achieve this, they first apply dependency parsing to obtain the syntactic structure of tweets. Then, relying on a set of manually-defined rules, their algorithm identifies syntactic patterns that reflect a relation between a term and a specific target. To overcome sparsity, they use a set of binary features which reflect whether a syntactic pattern appears in the tweet or not. Empirical results showed a significant improvement over pure target-independent classifiers.

Hybrid approaches have also been tested on Twitter messages. Zhang, Ghosh, Dekhil, Hsu, and Liu (2011) adopt a lexicon-based approach to perform entity-level sentiment analysis, which achieves high precision but low recall. Additional opinionated tweets are then identified by applying a chi-square test. Finally, a binary classifier (whose training data is provided by the lexicon-based method) is trained to assign sentiment polarities to the newly-identified opinionated tweets. Kumar and Sebastian (2012) propose combining a log-linear regression classifier to find the semantic orientation of adjectives and a dictionary-based method to find the orientation of verbs and adverbs, computing the overall sentiment by means of a linear equation.

Recently, syntactic structure has also been used at a certain level to help in the task of normalizing micro-texts (i.e., to convert their contents to standard language) such as Twitter and SMS messages. For this purpose, Costa-jussà and Banchs (2013) and Kaufmann and Kalita (2010) use machine translation techniques to align phrases in micro texts with the corresponding phrases in Haitian-Créole and English, respectively. In this respect, Han and Baldwin (2011) suggest that conventional supervised learning will not perform well due to data sparsity, as Twitter data exhibits a long tail of out-of-vocabulary words.

Motivation

In this article we study how lexical and syntactic features can help to improve polarity classification accuracy over Spanish tweets. In addition to the word forms, there exist several ways to extract complementary information to obtain better classifications. Many terms are associated with psychological properties, such as anxiety, anger or happiness. In the same line, morphological information can help discriminate between subjective and objective texts. For example, adjectives, adverbs or first person pronouns are *a priori* good indicators of opinionated texts. All this information is used and combined to create different supervised classifiers, in order to improve standard bag-of-terms approaches.

Moreover, we hypothesise that by syntactically relating these kinds of information it is possible to capture more context, improving accuracy (Rocher et al., 2013). For this purpose, we use dependency parsing to identify relations between words in order to overcome the problem of many sentiment detection approaches, which take into account individual words, but not their context. To identify these relations we rely on a more relaxed concept of generalised dependency triplets. Our aim is to use dependency triplets to capture interesting patterns between terms, modelling common linguistic phenomena such as negation or intensification, and many others which are difficult to treat by symbolic and pure lexical-based approaches. In general terms, figures of speech such as oxymoron are good examples of complex constructions that are uncommon, but should be taken into account by sentiment classifiers.

The main idea of the concept of *generalised composite features* is presented by Joshi and Penstein-Rosé (2009), as we commented previously. However, we believe this perspective is in itself insufficient for performing polarity classification on micro texts, for several reasons. Firstly, the authors worked on product reviews from Amazon, where vocabulary is more restricted and reduced than in Twitter, and ungrammatical elements are not so frequent. In addition, they used their perspective on identifying opinionated sentences in that domain, but it was not intended nor evaluated for classifying sentiments, neither on long nor on micro-texts. In this respect, only generalising to coarse PoS-tags can involve a loss of very useful information. In order to facilitate understanding, we will use examples in English to illustrate the relevant syntactic constructions in this and following sections, although the approach we are describing is designed for Spanish. Consider the sentence '*He makes a delicious villain*', which we will use as a running example in this section. According to the method proposed by Joshi and Penstein-Rosé (2009), the triplet (villain, modifier, delicious) would be generalised as (noun, modifier, delicious) or (villain, modifier, adjective). However, this is not an optimal generalisation. For example, selecting the option (villain, modifier, adjective) we are losing useful information because '*delicious*' provides sentiment by itself. However, if we try to use the original triplet, we will probably have sparsity problems because it is very unlikely that we have seen that specific combination of words and dependency relation in the training set. Finally, a base unigram approach would not be able to treat this sentence correctly, since the meaning of '*delicious villain*' can be different depending on whether these words appear together (which could be considered an oxymoron) or apart.

We adapt and enrich the initial concept of generalised dependency features, intended for detecting opinionated sentences, to improve the accuracy of lexical-based sentiment classifiers. We incorporate various levels of generalisation both for the head and the dependent term, instead just using part-of-speech information. We also contemplate deleting the dependency type, keeping only the head and the dependent term, which could be considered as a *syntactic n-gram.* In this way, given an original dependency triplet *($head_{i,}$ arc_{ij}, $dependent_{j}$)*, we apply a *generalisation function*, *g(x,G)*, where *x* represents the term to be generalised and *G*, the type of generalisation it will be generalised to: its part-of-speech information, its psychometric properties, its lemma, the term itself or even a blank slot (to completely remove the item). We also include a deletion function, *d*, to determine whether to keep the dependency type or not. As a result, a new dependency triplet *($g(head_i,G)$, $d(arc_{ij})$, $g(dependent_j, G)$)* is obtained. The goal here is to generalise composite features, but in such a way that we do not lose too much relevant semantic information. For example, the word '*villain*' could be assigned to the psychometric properties 'negative emotion' or 'anger', and the term 'delicious' could be classified as a 'positive emotion'. Thus, in the sentence '*He makes a delicious villain*' we could extract the triplets (negative emotion, modifier, positive emotion) and (anger, modifier, positive emotion) which are purely semantic dependencies, but more generalisable than (villain, modifier, delicious). Other examples of generalisation options could be (negative emotion, _, positive emotion) and (anger, _, positive emotion) if we omit the dependency type, or (common noun, _, positive emotion) if we apply a different generalisation for each term. To the best our knowledge, this is the first study on proposing this kind of composite generalised dependency triplets. Figure 1 sketches some of the theoretical advantages of this approach.

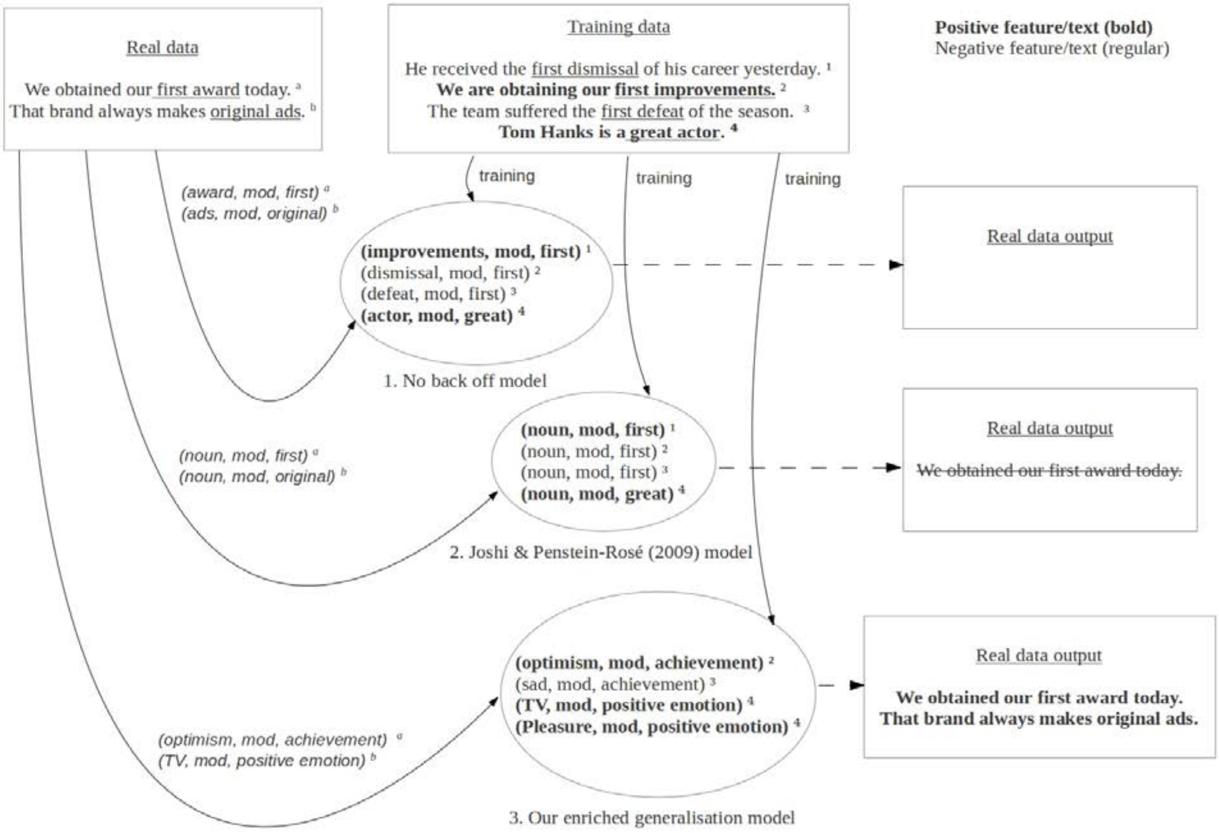

Real data
We obtained our first award today. [a]
That brand always makes original ads. [b]
Training data
He received the first dismissal of his career yesterday. [1]
We are obtaining our first improvements. [2]
The team suffered the first defeat of the season. [3]
Tom Hanks is a great actor. [4]
Positive feature/text (bold)
Negative feature/text (regular)
training
training
training
(award, mod, first) [a]
(ads, mod, original) [b]
(improvements, mod, first) [1]
(dismissal, mod, first) [2]
(defeat, mod, first) [3]
(actor, mod, great) [4]
1. No back off model
Real data output
(noun, mod, first) [a]
(noun, mod, original) [b]
(noun, mod, first) [1]
(noun, mod, first) [2]
(noun, mod, first) [3]
(noun, mod, great) [4]
2. Joshi & Penstein-Rosé (2009) model
Real data output
~~We obtained our first award today.~~
(optimism, mod, achievement) [a]
(TV, mod, positive emotion) [b]
(optimism, mod, achievement) [2]
(sad, mod, achievement) [3]
(TV, mod, positive emotion) [4]
(Pleasure, mod, positive emotion) [4]
3. Our enriched generalisation model
Real data output
We obtained our first award today.
That brand always makes original ads.

FIG 1: A naive example of different dependency triplet generalised models. We use both a hyphotetical training dataset and a small test dataset. The dependency type *mod* that appears in the dependency triplets is the short form of the syntactic function *modifier.* In the example, dependency triplets always contain this dependency type, because they are representing an adjective which modifies a noun. Our method generalises the words '*ads*' and '*actor*' to the category 'TV', according to the avaliable resources. This is an example of one of a number of generalisations that would be made by the new method. In fact, generalising '*actor*' to 'Films' and '*ads*' to 'Newspaper' or 'Magazine' would be two additional and acceptable options. Underlined phrases refer to the triplets that are taken as features. Boldface text refers to either a positive text or a generalised dependency triplet which implies a favourable sentiment. The real data output box contains the expected results for each model, where the strikethrough text indicates that the prediction was wrong and omitted texts mean that the model was unable to assign any label, given the corresponding input.

Polarity detection based on linguistic knowledge

In order to capture the semantic relations that convey information about the sentiment in a tweet, we will first apply a pipeline consisting of pre-processing, part-of-speech tagging and parsing.

Pre-processing tweets

Twitter is characterised by the use of a very informal language combined with specific Twitter elements, but also including formal expressions and figures of speech such as oxymoron, sarcasm or irony. The reply[3] '*@user I saw it* [the movie] *in London buddy...you'll freak out with Kevin! He makes a delicious villain! Hug!*' is a real example extracted from Twitter which reflects some of these particularities. The example contains an unreal Twitter user name for privacy reasons. We will use this sentence as our running example in this section, as far as possible.

We carry out an *ad-hoc* pre-processing of tweets, whose transformations we enumerate below:

Identification of the main compound expressions in the Spanish language: To deal with this issue, we extracted a list of compound terms present in the Ancora corpus (Taulé, Martí & Recasens, 2008). The Ancora corpus is a bilingual collection from newspapers for the Spanish and Catalan languages. It contains around 500,000 words, which are labelled with lemmas, part-of-speech information and dependency syntactic structure. We rely on this corpus to identify composite words and unite them into a single unit of meaning (e.g. '*not at all*' becomes '*not_at_all*').

Unification of punctuation marks: People often do not respect punctuation rules in web environments. This is a problem for the effectiveness of the segmentation and tokenisation steps, and thus for the rest of the processing procedure. Microtext-based social media such as Twitter are especially noisy. To solve this, we must homogenise the representation of punctuation marks, by adding blanks when required. To identify some of the most common problematic cases, we defined a set of regular expressions. For example, the typo *word1,word2* (which does not include a white space after the punctuation mark) is changed to *word1, word2*. In a similar line, the same process would be true for other punctuation marks (colon, semicolon, dots, question mark, etc). However, this change is often not desired if we are talking about numbers, where users often employ commas or colons to represent decimal points or high numbers. More complex cases are also considered, such as when a user concatenates a sentence which ends with a number with the beginning of a new sentence (e.g. *'[...] 13.I'm happy[...]'* should be *[...] 13. I'm happy[...]'*). We also try to normalise the number representation to Ancora format. If we consider the running example, the change is minimal; only a white space is added after the dots: '*[…] saw it in London buddy...you'll freak out with Kevin! [...]*' becomes '*[…] saw it in London buddy... you'll freak out with Kevin! [...]*'.

Treatment of Twitter special symbols ('@' and '#'): The use of Twitter special symbols is an important issue, not only for text analytics, but also for segmentation and tokenisation, as they can affect the performance of these processes. We deal with user mentions by removing the '@' symbol and capitalising the first letter (e.g. '*@user*' becomes '*User*'), because we hypothesise that user mentions usually refer to a proper name. An effective treatment of hashtags ('#') is more complex. A hashtag can be formed by a concatenation of multiple words, and often it refers to a very specific event and includes unknown words. In this case, we have followed a simple strategy: If the hashtag appears at the beginning or the end of the tweet we just remove it completely: we suppose that, in these cases, users only want to label their tweets. Otherwise, we only delete the '#', because we hypothesise that the rest of the hashtag contributes to syntactic information. The present pre-processor cannot properly handle composite hashtags, such as '*#word1_word2*' or '*#word1word2*', which will be taken as a unique token during the whole pre-processing of the tweet.

URL normalisation: We identify web addresses that appear in tweets and we change their form to the string '*URL*'.

Laughs normalisation: We pre-process irregular ways to express laughs in Spanish (e.g. '*jjjaaja*', '*JEJJJJJE*', …) as *jxjx* where x ϵ {a,e,i,o,u}, so as to be able to treat laughs in an unified way. We use a list of regular expressions to match the most common ways to simulate laughs in web texts. The pattern of the regular expression could be expressed as *[jJx]{4,}*; where *x* represents a character of the set {a,e,i,o,u}, and their corresponding uppercase. Interjections such as *'ja'* or *'jaJ'* are skipped, because we hypothesise they don't represent actual laughs, being often part of sarcasms or complaints.

Emoticon preprocessing: The emoticon list of Agarwal, Xie, Vovsha, Rambow and Passonneau (2011) is used as a reference. This collection distinguishes five classes of emoticons: emoticon-strong-positive (ESP), emoticon-positive (EP), emoticon-neutral (ENEU), emoticon-negative (EN), and emoticon-strong-negative (ESN). The pre-processing algorithm replaces the form of the emoticon by a string which represents the class. The resulting phrase is placed as a separate sentence in the tweet (e.g. '*I am happy :)'* becomes '*I am happy. Emoticon-positive.*') in order not to interfere with the subsequent tagging and parsing steps.

Part-of-Speech tagging

Part-of-speech (PoS) tagging is the process of marking up a word in a text as corresponding to a part-of-speech, based on both its definition and its context. Part-of-speech tags can be coarse-grained (when they only represent the grammatical category: noun, verb, adjective, etc.) or fine-grained (when they include additional morpho-syntactic information such as gender, number, tense, etc.). Although some PoS taggers specifically designed for Twitter messages written in English have appeared recently (Gimpel et al., 2011; Owoputi, O'Connor, Dyer, Gimpel, & Schneider, 2012), there are no Twitter PoS taggers available for Spanish. Therefore, we have decided to use as PoS tagger a version of Brill's tagger (Brill, 1995) implemented in NLTK (Bird, Klein & Loper, 2009). Due to the lack of a corpus of Spanish tweets with morpho-syntactic annotation, we trained the tagger using the Ancora corpus, taking 90% of the corpus as the training set and the remaining 10% as the test set, achieving an accuracy of 95,81%. However, when analysing the output of this model on web texts, systematic appearance of several tagging errors was observed. We realized that one of the main tagging problems was that users often ignore diacritical accents (e.g 'el' ('the') instead of 'él' ('he')), which significantly decreased performance. According to Foster et al. (2012), who show that a substantial proportion of parsing errors can be attributed to PoS tagging errors, it is crucial to improve the actual performance of PoS tagging. For this purpose, the training set was expanded: we cloned each sentence of the original training set, adding a version of the sentence with the accents removed in addition to the original sentence, and we used it to train a new tagger. We created two additional PoS-tags to re-annotate emoticons and URLs: emoticon-tag and url-tag. The accuracy obtained was 95.71%, very close to that of the original tagger; but the practical performance over web reviews was much better.

Dependency parsing

We analyse the syntactic structure of tweets by means of dependency parsing. For each tweet, the parser obtains its corresponding dependency tree. Each such tree is a set of triplets $\{(w_i,arc_{ij},w_j)\}$ which establish binary relations between words where: w_i is the *head*, w_j the *dependent*, and arc_{ij} is the *dependency* connecting w_i to w_j, labelled with the syntactic function that relates both terms, called *dependency type*. Figure 2 shows an example of a valid dependency tree for the phrase of our running example: '*He makes a delicious villain*'. We have used MaltParser (Nivre et al., 2007) and the Ancora corpus to train a data-driven parser model based on the *Nivre arc-eager* algorithm (Nivre, 2008).

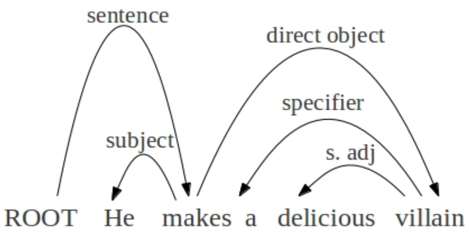


FIG 2: An example of a dependency tree.

A semantic orientation analyser

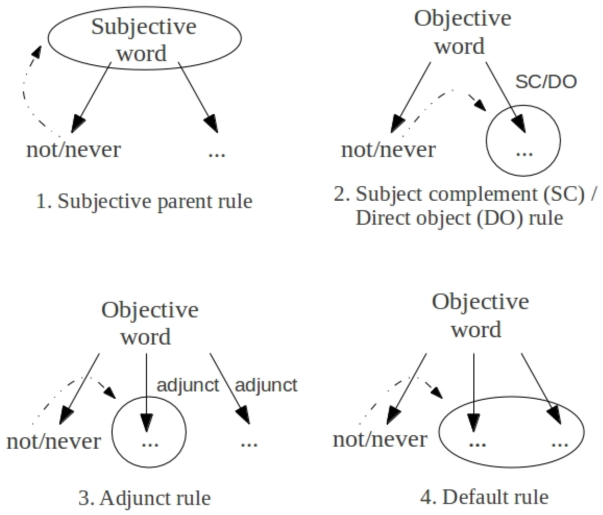
In Vilares et al. (2013) we detailed a completely unsupervised system for extracting the semantic orientation of long reviews. It is a generic system which establishes a good baseline in a different number of domains, as we showed in that study. The semantic knowledge of the system relies on the opinion lexicons presented in Brooke et al. (2009). Moreover, to obtain a reliable and robust output, the system defines a set of syntax rules to deal with phenomena such as intensification, negation or adversative subordinate clauses.

Treatment of intensification. The literature (Kennedy & Inkpen, 2006; Taboada et al. 2011) typically defines two types of intensifiers: *amplifiers,* which are used to maximize the sentiment of a word (e.g. '*very*', '*so*'); and a second group of valence shifters which diminish the semantic orientation of a term, called *downtoners* (e.g. *'a bit', 'not at all'*). Our algorithm detects these expressions, or combinations of both, and uses their syntactic head to identify the exact scope whose semantic orientation will be affected. The semantic orientation is then modified by a percentage, according to the opinion lexicons of Brooke et al. (2009).

Treatment of negation. Handling negation correctly is crucial to perform a robust sentiment analysis. Brown and Levinson (1987) and Brooke et al. (2009) discuss the human tendency to polite opinions. People tend to be polite when expressing their opinion in reviews, which often means the use of a negation of a positive term instead of using its corresponding antonym. Saying '*not good*' instead of '*bad*' and '*not very smart*' instead of '*stupid*' are typical examples of this phenomenon. To identify the scope of negation in a linguistically-motivated way, we define a set of four dependency-based rules. In this way, we deal with the most common negation terms in Spanish language such as '*no*' ('not') and '*nunca*' ('never'). These rules look for certain dependencies and dependency types to find a candidate scope and shift their semantic orientation. The rules, which are described in detail in Vilares et al. (2013), are processed in order and when one of them matches, the rest are discarded. Figure 3 illustrates this treatment. The first rule checks if the head of the negation term is subjective, in which case the semantic orientation of that node is changed. The second rule tries to find a branch at the same level of a negation term labelled with a *subject complement* or a *direct object* dependency type, and modifies that branch. If this rule does not match either, we then look for the nearest adjunct branch at the same level of the negation, shifting their semantic orientation. Finally, is none of the previous rules matches, the candidate scope to be shifted is composed by all the right branches of the negation term. An example for each rule would be:

- Figure 3.1: He didn't pass the maths exam.
- Figure 3.2: Sarah is not so awesome.
- Figure 3.3: Archibald does not work properly on Mondays.
- Figure 3.4: The police do not distinguish between radicals.

where underlined words correspond to the scope of negation, represented as circled words in Figure 3.



1. Subjective parent rule

2. Subject complement (SC) / Direct object (DO) rule

3. Adjunct rule

4. Default rule

FIG 3: Rules for scope identification of negation

Treatment of adversative subordinate clauses. Sentences of this kind can be considered in sentiment analysis as a type of intensifiers, because they allow ideas to be restricted. For example, in the sentence '*I like it, but it's expensive*', although both the main and the subordinate sentence reflect opinion, the second one seems to be more relevant. If we exchange both sentences and we build the micro review '*It's expensive, but I like it*', the adversative subordinate clause once again seems to be more relevant. A formal linguistic explanation regarding the restricting nature of subordinate conjunctions for the Spanish language can be found in Campos (1993). An issue that must be taken into account is that the Ancora corpus annotates this type of sentences heterogeneously, so we decided to adapt the syntactic structure of these clauses in order to treat them in a uniform way. Basically, we create an artificial Subordinate Adversative Clause node (called SAC) to identify the beginning of a composite sentence which contains an adversative subordinate clause. We then locate both the sub-graphs corresponding to the whole main sentence and the subordinate one, as children of the SAC node.

Figure 4 shows how our algorithm deals with some of these linguistic constructions in the sentence '*The transfer was cordial, but not complete*'[4], which is part of a real tweet. The main clause is '*The transfer was cordial*', which has a positive semantic orientation of 2 because the word '*cordial*' appears in the semantic dictionaries of Brooke et al. (2009). With respect to the subordinate clause, by applying the subjective parent rule to identify the scope of a negator '*not*', we obtain a semantic orientation of -2, due to the negation of the word '*complete*'. The SAC node computes and weighs both the semantic orientation of the main and the subordinate clauses, giving more importance to the second one. Finally, we obtain a negative SO for the whole sentence, which is coherent with human intuition, because the subordinate clause suggests that something failed on the transfer.

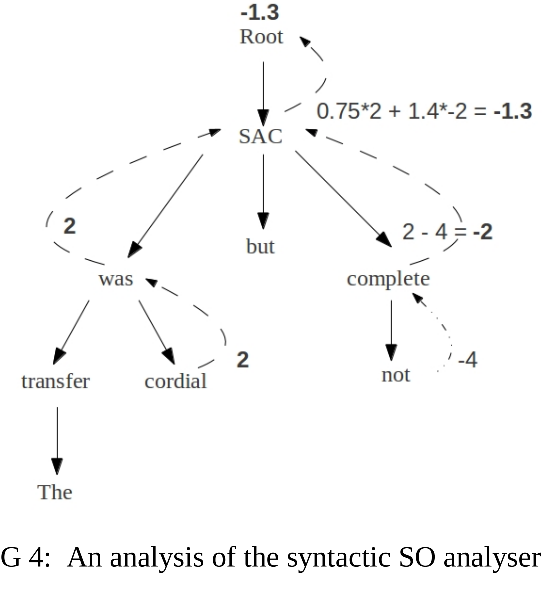


FIG 4: An analysis of the syntactic SO analyser

Experimental setup

The experiments will focus on answering unsolved questions about polarity classification over Spanish tweets, such as:

- Does it make sense to apply optimisation steps, such as lemmatisation, to a purely bag-of-words approach?
- Is it helpful to combine lexical, psychometric and semantic knowledge to improve performance on analysing tweets?
- Is it beneficial to incorporate syntactic information over pure lexical-based models? In this respect, is the concept of composite generalised features useful?

The research questions are addressed by implementing the involved features within a supervised classifier and then evaluating them on suitable test data. The experimental test set-up is described below.

Dataset description

The TASS 2013 General Corpus is a collection of tweets which has been specifically annotated to perform polarity classification at a global level, presented at the Workshop on Sentiment Analysis at SEPLN[5] (Villena-Román & García-Morera, 2013). It is a collection of Spanish tweets written by public figures, such as soccer players, politicians or journalists. Messages range from November 2011 to March 2012. The corpus is composed of a training set and a test set which contain 7,219 and 60,798 tweets, respectively. Each tweet is annotated with one of these six labels: *strong positive* (P+), *positive* (P), *neutral* (NEU), *negative* (N), *strong negative* (N+) and *none* (NONE). Neutral tweets refer to messages that contain both positive and negative ideas; whereas tweets labelled as NONE concern those that do not express any sentiment. In addition, each tweet is labelled with the set of topics it is talking about: *films, sports, economics, entertainment, soccer, literature, music, other, politics* and *technology*. A previous version of this corpus was introduced at the TASS 2012 (Villena-Román et al., 2013), where the training set was labelled manually, but it was revised in 2013 in order to correct certain labelling errors. The gold standard was generated by a pooling of the submissions of the workshop of 2012, followed by a human review by the TASS organisation, for the thousands of ambiguous cases.

Table 1 shows the polarity distribution of tweets in the collection, for both the training and test sets. As we can see, distributions are dissimilar between the two sets. This should arguably not be seen as a weakness of the corpus, but rather as a characteristic that is coherent with real-life settings, since the frequencies of the polarities of the tweets that are posted each day change depending on the topic. Regarding this issue, some studies (Brown & Levinson, 1987, Kennedy & Inkpen, 2006) highlight a general tendency of human language to positive classification, which could justify the presence of more positive reviews in training corpora.

TABLE 1 Statistics of the TASS 2013 corpus.

Category	#tweets (Official training set)	#tweets (Official test set)
P+	1,652 (22.9%)	20,745 (34.1%)
P	1,233 (17.1%)	1,488 (2.4%)
NEU	670 (9.3%)	1,305 (2.1%)
N	1,335 (18.5%)	11,287 (18.6%)
N+	847 (11.7%)	4,557 (7.5%)
NONE	1,482 (20.5%)	21,416 (35,2%)
Total	7,219 (100.0%)	60,798 (100.0%)

In order to perform a standard classification, we must be able to work on three-class categorisation: *positive*, *negative* and *none*. In this case, *neutral* tweets will be discarded and *strong positive* and *strong negative* tweets will be included in the *positive* and *negative* classes, respectively.

Supervised ML classifiers

We rely on the WEKA data mining software (Hall et al., 2009) to build machine learning models. As a classifier, we have chosen to work with SMO, an implementation of Support Vector Machines (SVM) proposed by Platt (1999). The use of an SMO is supported by the well-known good performance of SVM on classification tasks. Our preliminary experiments suggested that the SMO outperformed other implementations of SVM, and also other classification techniques such as Bayesian models or decision trees. In addition, we used the WEKA attribute selection tools to apply feature reduction. A lower number of features makes the training process faster and helps avoid irrelevant attributes, which is especially important in noisy media such as Twitter. Concretely, we relied on *information gain* (Mitchell, 1997) to decide the relevance of features in each model, selecting only those features with an information gain greater than zero.

Feature models

We have defined several feature models in order to test the effectiveness of relating features by means of dependency parsing when they are used in conjunction with models based on lexical and semantic knowledge.

These models are tested, trained and finally some of them are combined to obtain high-performing classifiers. Our aim is to be able to find out the real capacity of features to bring about a positive effect on performance.

Naive baseline: The trivial model is established by assigning all the instances to the majority class in the training set.

Sentiment information: In the present work, we propose to use the information provided by the unsupervised semantic orientation analyser described in Vilares et al. (2013), adapted to Twitter messages as indicated in the previous section. Concretely, the features the analyser provides to the classifier for each tweet are: 1) its global semantic orientation; 2) the number of positive words that appear in the tweet; and 3) the number of negative words that appear in the tweet.

Bag of words: A widely used supervised approach is to consider tweets as bags-of-words and to use them to feed a supervised classifier. Although simple, this strategy generally shows a good performance.

Bag of lemmas: A natural extension of the previous approach is to first apply lemmatisation which allows the number of features to be restricted. This can be useful in languages such as Spanish, where gender or number is expressed by declensions of nouns, adjectives or verbs. We rely on the collection of lemmas provided by the Ancora corpus to lemmatise words.

Lexical bigram features: In addition to unigrams, we also performed experiments using bigrams of words and lemmas.

Part-of-speech features: The use of PoS tags in polarity classification is a widely discussed issue in many studies (Pang et al., 2002; Spencer & Uchyigit, 2012). However, the utility of PoS tags by themselves is camouflaged because they are used in conjunction with other features (Pak & Paroubek, 2010; Zhang, Zeng, Li, Wang & Zuo, 2009). We test the effectiveness of both fine and coarse part-of-speech tags.

Psychometric features: We introduce a perspective based on psychological knowledge. We rely on the dictionaries presented by Ramírez-Esparza, Pennebaker, García and Suriá Martínez (2001). This lexicon distinguishes around 70 dimensions of human language. It provides information about *psychometric properties* of words (cognition mechanisms, anxiety, sexuality, etc.), but also considering *topics* (TV, family, religion, etc.) or even *linguistic information* (past, present and future tense, exclamations, questions, etc.). In this way, the verb 'imagine' would represent a cognition mechanism and insight. This psychological linguistic resource is found in the LIWC software (Pennebaker, Francis & Booth 2001).

Dependency type features: We take only the identifiers of the dependencies appearing in the parse tree of each tweet. Thus we are not considering any information regarding the words linked by dependencies. In this case, we try to test if dependency types can be helpful by themselves to solve polarity classification tasks.

Syntactic features: The models described above these lines will serve as a starting point from which to incorporate syntactic knowledge. Concretely, we represent syntactic information by means of *generalised dependency-based features.* The aim is to measure the effectiveness and sparsity problems of this type of features when they are used both separately and in conjunction with lexical-based models. We test different levels of generalisation over the head and the dependent word of a dependency triplet, including lemmas, psychometric properties and fine PoS-tags.

Evaluation metrics

The performance of our experiments is evaluated by means of standard metrics for sentiment analysis and text classification:

$$\text{Accuracy} = \frac{\text{\# instances classified correctly}}{\text{\# instances}}$$

$$\text{Precision}(i) = \frac{\text{\# correct instances classified in class } i}{\text{\# instances classified in class } i}$$

$$\text{Recall}(i) = \frac{\text{\# correct instances classified in class } i}{\text{\# instances of class } i}$$

$$\text{F1}(i) = \frac{2 \times \text{Precision}(i) \times \text{Recall}(i)}{\text{Precision}(i) + \text{Recall}(i)}$$

where *i* refers to a polarity category, such as Positive or Negative.

Experimental runs

The models proposed above are evaluated through two sets of experiments, in order to measure how the size of the training corpus can affect phenomena such as sparsity. In both cases, we perform a standard three class categorisation considering positive (P), negative (N) and without opinion (NONE) classes from the TASS 2013 corpus. This means that performance will not be directly comparable to the systems which participated at the TASS 2013 workshop, where only classification into 4 and 6 categories was proposed. To overcome this limitation, additional experiments on 4 and 6 classes are included for the best performing models.

'From small to large corpus' experiments: This first set-up relies on the training set of the TASS 2013 corpus to build the models, and we evaluate them against the test set. The training set of the TASS 2013 corpus only contains 6,549 tweets if we just consider those in the classes P, N and NONE.

'From large to small corpus' experiments: In this configuration, we use the test set of the TASS 2013 to train the models, and we evaluate them against the training set. In order not to cause confusion, we refer to the test set as the *reversed training set* and the training set as the *reversed test set*. The aim of this experiment is to measure the effect of sparsity on the different models proposed. The size of the reversed training set is 59,493; considering positive, negative and none tweets, so it is around 10 times bigger than the original training set. We have also trained models using incremental parts of the reversed training set, to show how its size may affect to the accuracy of different perspectives. We are aware that the reversed training set can present some annotation errors, because it was made by pooling, followed by a human revision. We hypothesise that this will manifest itself in the form of a somewhat lower yield on the reversed test set, but not in the practical utility of the perspectives proposed. Optimisation of models was made over this configuration, so we decided to split (fifty-fifty) the reversed test set into two parts: a development set, to analyse how properly combine different sets of features, and a test set to evaluate the real performance of selected models.

Experimental results

We show the performance obtained for the feature models defined in the previous section using two different configurations: '*from small to large corpus*' and '*from large to small corpus*'.

From small to large corpus configuration

Results are shown in Table 2. The bag-of-lemmas approach obtains the best performance, followed by the pure bag-of-words model. Table 3 shows how the performance improves over the initial learning-based settings when features are used in conjunction. We obtained the best performance by creating a model which combines lemmas, psychometric properties, and the information provided by the unsupervised system. Specifically, the semantic orientation and the number of positive and negative words that appear in a tweet. We take the accuracy obtained by this combined model as a good indicator of what can be achieved without considering relations between words. We then test the effect of including syntactic information over this lexical-based model, by adding generalised dependency triplets. We did not achieve any improvement incorporating syntactic features, following this experimental run, but Table 4 shows the results for some models which were able to improve performance when the collection is larger.

From large to small corpus configuration

Table 5 shows the results while Table 6 aims to show how their accuracy is improved when features are combined. As in the *'from small to large'* experiments, we obtain the best performing lexical model by creating a classifier which combines lemmas, psychometric properties and the information provided by the unsupervised analyser. This combined model is again taken as the base point from which to include syntactic information, in order to test the real effectiveness of generalised dependency features. The goal is to measure how relating terms, psychometric properties or part-of-speech information, by means of dependency parsing, can increase accuracy with respect to employing this knowledge in a purely lexical way.

Table 7 illustrates some improvements obtained on accuracy when different generalised dependency triplets and the features of the best performing lexical model are used together. Given the number of possible combinations of generalised features, we only provide results for those that obtained some degree of improvement. We also include results for the best models that we achieved by combining several types of generalisation. Asterisks indicate a statistically significant difference using chi-square tests. Unlike the configuration *'from small to large',* in this case syntactic information is useful to improve performance. This suggests that although useful, generalised dependency triplets suffer from sparsity and a larger training set is needed to properly exploit this type of feature. We show below some cases where we believe that generalised dependency triplets were helpful to correct the polarity assigned by the best lexical-based model on some difficult tweets:

'*@Maropopins5:jajaja creo que es peor este que vi yo. Otro incunable ;)*' ('*@Paropopins5:hahaha I believe this one I saw it is worse. Another incunable ;)*'). The best-performing lexical model determined that this tweet is positive, while it was annotated as Negative in the TASS 2013 corpus. Although the model identifies the negative word '*worse*' it also recognises the laugh '*hahaha*' and the emoticon '*;)*' as positive terms and finally decides to take the tweet as Positive. The main issue is that the lexical perspective does not differentiate between words forming part of the "core" of the sentence and those simply offering "auxiliary" information. In this respect, the employment of generalised dependency triplets helps to take into account the syntactic structure of tweets in order to assign greater relevance to main syntactic functions such as the *subject*, *direct* *object* or *subject complement*, on which most of the meaning of the sentence relies. In this case, the term '*worse*' is the subject complement of the sentence, so the model considers the triplet (*'is', subject complement, 'worse')*. By backing off the head of this feature to their psychometric properties, the best-performing syntactic model matches it with generalized triplets such as *('Present time', subject complement, 'worse')* or *('Reference to other', subject complement, 'worse')*, which are *a priori* negative, and finally classifies this tweet in its right category (Negative).

'*Cansada de la familia Livela*' ('I'm tired of Livela family'). The lexical model classified this tweet as objective, due to the word '*tired*', which when analysed in isolation does not express any opinion, but simply describes a lack of energy. However, it is important to note the difference between saying '*tired*' and '*tired of*'. The syntactic model is able to correctly deal with the triplet *('tired', prepositional object, 'of')* and assign this feature to more general ones such as *(Sleep, prepositional object, 'of')* or *(Physical, prepositional object, 'of')* where *Sleep* and *Physical* are both psychometric properties. Using these generalised features is better than employing the original non generalised feature, because in addition to '*cansado de*' ('*tired of*') they also encapsulate the meaning of similar Spanish phrases such as '*aburrido de*' ('*bored of*') o '*harto de*' ('*sick of* ').

Experiments on 4 and 6 classes

As we indicated in the experimental setup section, tweets in the TASS 2013 corpus are annotated with six labels and can thus be used to test performance on a more fine-grained scale of polarities. In this respect, Tables 8 and 9 present the performance for the most relevant feature models when they are used to classify polarity into 4 and 6 categories, respectively, using the '*from small to large*' configuration. Tables 10 and 11 show experimental results for 4 and 6 categories, respectively, this time according to the '*from large to small*' configuration.

TABLE 2. Performance of initial feature models, following the '*from small to large*' setup: #features refers to the number of features of each model with an information gain greater than 0. Po-F1, Ne-F1 and None-F1 refer to the value of F1 calculated for the positive, negative and none classes, respectively. Accuracy refers to the global accuracy, calculated over all the classes of tweets.

Features	#features	Po-F1	Ne-F1	NONE-F1	Accuracy
Lemmas	755	0.731	0.674	0.580	0.669
Words	851	0.701	0.655	0.557	0.645
Sentiment information	3	0.641	0.576	0.575	0.600
Psychometric	57	0.654	0.601	0.501	0.594
Fine-grained PoS-tags	86	0.611	0.561	0.474	0.559
Dependency types	33	0.575	0.497	0.447	0.519
Bigrams of lemmas	998	0.592	0.565	0.295	0.514
Coarse-grained PoS-tags	16	0.552	0.489	0.440	0.504
Bigrams of words	915	0.573	0.528	0.204	0.480
Naive Baseline	1	0.544	0.000	0.000	0.374

TABLE 3. Performance on combining sets of features of the initial learning-based methods, following the '*from small to large*' setup: *lemmas* (L) *psychometric* (P), *fine-grained PoS-tags* (FT), *dependency types* (DT), *sentiment information* (S).

Features	#features	Po-F1	Ne-F1	NONE-F1	Accuracy
L+P+S	601	0.765	0.702	0.609	0.700
L+P+FT`+S	696	0.764	0.701	0.608	0.698
L+P	598	0.749	0.688	0.592	0.684

TABLE 4. Performance on incorporating generalised dependency features, following the '*from small to large*' setup. We use the notation (*head*, *dependency*, *dependent*) for representing sets of generalised dependency triplets, where '_' is used to indicate omitted elements: *lemmas* (L) *psychometric* (P), *coarse-grained PoS-tags* (CT) , *fine-grained PoS-tags* (FT), *dependency types* (DT), *sentiment information* (S).

Features	#features	Po-F1	Ne-F1	NONE-F1	Accuracy
L+P+S (LPS)	601	0.765	0.702	0.609	0.700
LPS+(L, DT, P)	1,102	0.756	0.695	0.600	0.692
LPS+(L, DT, CT)	1,242	0.756	0.696	0.600	0.692
LPS+ (L, _ , P)	1,131	0.757	0.697	0.600	0.692
LPS+(L, _ , CT)	1,244	0.712	0.696	0.600	0.691
LPS+(L, _ ,L)	1,319	0.751	0.692	0.590	0.686

TABLE 5. Performance of initial learning-based methods, following the '*from large to small*' setup.

Features	#features	Po-F1	Ne-F1	NONE-F1	Accuracy
Words	4,288	0.767	0.691	0.622	0.702
Lemmas	3,192	0.769	0.691	0.622	0.701
Bigrams of lemmas	9,066	0.731	0.657	0.575	0.659
Bigrams of words	9,441	0.694	0.596	0.547	0.617
Sentiment information	3	0.635	0.548	0.523	0.577
Fine-grained PoS-tags	148	0.603	0.548	0.513	0.560
Psychometric	63	0.595	0.576	0.513	0.559
Dependency types	40	0.553	0.455	0.502	0.511
Coarse-grained PoS-tags	16	0.517	0.454	0.484	0.489
Naive baseline	1	0.611	0.000	0.000	0.440

TABLE 6. Performance on combining sets of features of the initial learning-based methods, following the '*from large to small*' setup. Although L+P+FT+S obtain an small improvement over the L+P+S model at the test set, the L+P+S approach was taken as the starting point to incorporate syntactic information, since it obtained the best performance at the development set.

Features	#features	Po-F1	Ne-F1	NONE-F1	Accuracy
L+P+FT+S	3,031	0.779	0.708	0.634	0.715
L+P+S	2,881	0.779	0.701	0.634	0.713
L+P	2,878	0.774	0.700	0.628	0.708

TABLE 7. Performance on incorporating generalised dependency features, following the '*from large to small*' setup. The model marked with an '*' shows a statistically significant difference ($p < 0.05$) with respect to the LPS method.

Features	#features	Po-F1	Ne-F1	NONE-F1	Accuracy
L+P+S (LPS)	2,881	0.779	0.701	0.634	0.713
LPS+(L, DT, CT)+ (L,_,CT)+(L,_,FT)+ (P, DT, L)+(CT,_, P)*	25,996	0.784	0.720	0.635	0.722
LPS+(L, _ , CT)	7,660	0.782	0.713	0.638	0.718
LPS+(L, DT, CT)	8,189	0.782	0.710	0.636	0.717
LPS+(L, DT, P)	8,671	0.783	0.702	0.638	0.716
LPS+(L, _ ,L)	11,057	0.779	0.706	0.635	0.714

TABLE 8: Performance of some relevant models obtained from the '*from small to large*' setup, evaluated over 4 categories: positive, neutral, negative and none tweets.

Features	#features	Po-F1	Neu-F1	Ne-F1	NONE-F1	Accuracy
LPS	428	0.76	0.124	0.684	0.609	0.677
Lemmas	485	0.715	0.086	0.641	0.568	0.636

TABLE 9: Performance of some relevant models obtained from the '*from small to large*' setup, evaluated over 6 categories: strong positive, positive, neutral, negative, strong negative and none tweets.

Features	#features	Po+-F1	Po-F1	Neu-F1	Ne-F1	Ne+-F1	NONE-F1	Accuracy
LPS	237	0.697	0.218	0.158	0.534	0.535	0.646	0.586
Lemmas	220	0.671	0.239	0.121	0.493	0.518	0.623	0.566

TABLE 10: Performance of some relevant models obtained from the '*from large to small*' setup, evaluated over 4 categories: positive, neutral, negative and none tweets. The model marked with an '*' shows a statistically significant difference (p<0.01) with respect to the LPS method.

Features	#features	Po-F1	Neu-F1	Ne-F1	NONE-F1	Accuracy
LPS+(L, DT, CT)+ (L,_,CT)+(L,_,FT)+ (P, DT, L)+(CT,_, P)*	17,876	0.748	0.178	0.647	0.603	0.643
LPS	2,127	0.739	0.098	0.652	0.592	0.639
Lemmas	2,366	0.728	0.118	0.650	0.587	0.633

TABLE 11: Performance of some relevant models obtained from the '*from large to small*' setup, evaluated over 6 categories: strong positive, positive, neutral, negative, strong negative and none tweets. The model marked with an '*' shows a statistically significant difference (p<0.01) with respect to the LPS method.

Features	#features	Po+-F1	Po-F1	Neu-F1	Ne-F1	Ne+-F1	NONE-F1	Accuracy
LPS+(L, DT, CT)+ (L,_,CT)+(L,_,FT)+ (P, DT, L)+(CT,_, P)*	12,671	0.669	0.225	0.154	0.484	0.495	0.598	0.525
LPS	1,649	0.652	0.141	0.093	0.485	0.465	0.578	0.507
Lemmas	1,726	0.649	0.157	0.093	0.469	0.479	0.578	0.504

Discussion of the results

From Tables 2 and 5, which show the performance of the basic feature models with both the *'from small to large'* and *'from large to small'* configurations, the tendency with respect to accuracy remains very similar in both runs. The bag-of-lemmas seems to be the most successful set of features, followed by the bag-of-words approach. In particular, in the *'from small to large'* run, the use of lemmas clearly outperforms the use of words. This shows the need to apply some type of normalisation of Spanish words, reducing the rich morphology of this language, but keeping the meaning of words.

With respect to the *'from large to small'* run, both bag-of-lemmas and bag-of-words obtain virtually the same accuracy, although lemmas employ a much lower number of features. It is important to note that a model based on a pure bag-of-words instead of on a bag-of-lemmas, implicitly captures features such as gender or number, which are good features by themselves, as we will discuss below. Thus, a bag-of-words model contains, to a certain extent, analogous information to that included in a combined model of bag-of-lemmas and fine part-of-speech tags. Models based on bi-grams show a low performance, probably due to the sparsity of these features in a small training set.

The psychometric approach also achieves a decent performance, strengthening the importance of taking semantic approaches as a starting point. Table 12 illustrates the top 5 features for these three approaches, based on their information gain, taking the *'from large to small'* configuration. The pair *'the/he'* would correspond to *'el/él'* in Spanish language. Actually, the second best discriminative was just *'el'*. However, as we commented previously, Spanish users often ignore acute accents when writing in web environments and furthermore articles are often omitted in microtexts. Therefore, we hypothesise the form *'el'* many times really refers to *'él'*. In this respect, third person pronouns are often good indicators of objective texts, since informative texts often present a distance from the sender, whilst opinions are more frequently expressed with first or second person pronouns.

TABLE 12. Ranking of best discriminating features for some of the initial learning-based methods, following the '*from small to large*' setup.

Ranking	Bag-of-lemmas	Bag-of-words	Psychometric properties
1	Positive emoticon	Positive emoticon	Positive-emotion
2	the/he	!	Affective
3	!	Not	Negative-emotion
4	Thanks	That	Positive-sentiment
5	Not	Thanks	Article

An interesting finding is the accuracy obtained by only using part-of-speech tags. Although it hardly provides any explicit semantic information, the fine-grained part-of-speech tags model obtains an accuracy similar to the psychometric approach. This suggests that features such as gender, number or some word categories (e.g., conjunctions) can be good classifiers in themselves. Table 13 shows the ranking of the top fine-grained PoS tags, according to their information gain in the training set, which reinforces this hypothesis. Labels such as the *close exclamation mark*, or the artificial *emoticon-tag,* are two of the most discriminative features, probably because they are good indicators of subjective tweets. In Spanish there also exists an open exclamation mark '¡', conventionally used to mark the beginning of an exclamation, but users often ignore it in web environments. The occurrence of the tag *subordinating conjunction* in the top five of the best part-of-speech features suggests the importance of identifying adversative subordinate clauses, as we have pointed out previously. Subordinating constructions often compare and oppose arguments, which represents a good point to identify subjective texts. The fine-grained PoS-tag *Verb 3 person singular present indicative* is intuitively a good indicator of objective texts, as has been noted by other authors (Pak and Paroubek, 2010; Spencer & Uchyigit, 2012): people giving an opinion tend to use first person pronouns, because they are probably talking about something that happened to them; but the same is usually not true for people who are merely reporting on a fact, where third person pronouns are more frequent. In Spanish, subject pronouns are usually eliminated, since inflected verb forms provide us with the information needed to determine the number of the subject, which can be helpful to differentiate between subjective and objective texts, as we have just described.

Dependency types, which represent the syntactic functions present in a tweet, seem not to be very helpful in themselves.

TABLE 13. Ranking of best discriminating fine-grained PoS tags, following the '*from small to large*' setup.

Ranking	Feature
1	Close exclamation mark
2	Verb 3 person singular present indicative
3	Negative adverb
4	Emoticon-tag (artificial tag)
5	Subordinating conjunction

Tables 3 and 6 show how we can improve performance in an effective way by combining different sets of basic features, obtaining a better lexical-based model. Combined models which incorporate unsupervised sentiment information (S), are not purely lexical-based, since our semantic orientation analyser uses heuristic syntactic rules. For both runs, the classifier whose features are the lemmas, psychometric properties, semantic orientation and the number of positive and negative words that appear in a tweet achieved the best performance. This allows us to establish a ceiling of effectiveness for dealing with terms in an isolated way. Moreover, with this model we reduced the number of features with an information gain greater than zero with respect to the best basic approach, the bag-of-lemmas perspective. Other lexical-based models which add linguistic information such as part-of-speech tags or dependency types did not increase the accuracy (difference not statistically significant. $p < 0.10$). Table 14 shows some of the most discriminative features for the best combined model which does not take into account generalised dependency triplets. The information provided by the unsupervised system seems to be highly relevant, validating the utility of that approach. The most discriminative lemma appears in the eighth position, although lemmas were the best approach when they were considered in isolation.

TABLE 14. Ranking of best discriminating features when we combine lemmas, psychometric properties and the information provided by our unsupervised system, following the *'from large to small'* setup.

Ranking	Feature	Provided by
1	Semantic orientation	The unsupervised system
2	Positive emotion	Psychometric approach
3	#positive words	The unsupervised system
4	#negative words	The unsupervised system
5	Affective	Psychometric approach
8	Positive emoticon	Lemmas approach

Tables 4 and 7 reflect the effect on performance when syntactic information is provided in the form of generalised dependency triplets; both for the *'from small to large'* and *'from large to small'* configurations. With respect to the *'from small to large'* runs, generalised dependency triplets do not improve the performance over the best lexical model. This is due to the high sparsity of this type of feature and the relatively small size of the training corpus, which is not even able to successfully exploit a model based on a bag-of-lemmas, as we have seen previously.

On the other hand, in the *'from large to small'* experiments, syntactic information are helpful to improve performance over purely lexical models. If we incorporate different types of generalised dependency triplets over the lexical model we obtain small improvements, but when several of these features are jointly aggregated we obtain an even higher accuracy. It is important to note that the best models were mainly obtained by including features which carry out a high level of generalisation on the dependent node, contradicting the approach proposed by Joshi and Penstein-Rosé (2009), who suggested that it is better to generalise the head node. However, when generalised dependency triplets were evaluated in isolation, performing a higher generalisation on the head node was more appropriate.

Table 15 presents a sample of interesting features for the model which obtained the best performance on the *'from large to small'* configuration. Some of these features show how Spanish users relate terms according to the frame of mind of society at large. For example, the term '*police*' appears directly associated with the psychometric category '*negative emotion*', probably due to the strikes and demonstrations occurring in Spain during the period in which the tweets were collected. Along the same lines, Spanish users relate the word '*economy*' with '*negative emotion*'. Picking topic terms is pointed out as a risk on building supervised sentiment classifiers. Given a training corpus, if a topic word such as *'econom*y' or *'police'* appears mostly in one class, those words should not be considered for analysing new tweets, due to the bias of the training set. Our approach is no exception to this limitation, because we are including unigrams of lemmas. However, the use of composite generalised features can diminish this phenomena, since we are able to relate topic words with psychological properties, which are fine complements for topic words, as is shown at the examples of the table 15. Moreover, we realized that generalised dependency triplets were able to catch, to a certain extent, the discourse structure on Twitter. As we can see in the same table, to classify the polarity of a tweet the use of the word '*thanks*' at the end of the sentence seems to be more relevant than explicitly thanking somebody (shown by the feature '*(thank, _, proper name)*').

TABLE 15. A sample of generalised dependency features for the best model, following the *'from large to small'* setup.

Ranking	Feature	Provided by
48	(Thanks, _, punctuation mark)	(Lemmas, _,Coarse tag)
349	(Thanks, _, proper name)	(Lemmas, _, Coarse tag)
447	(noun, _, Anxiety)	(Coarse tag, _, Psychometric)
6,863	(Negative emotion, s.a, police)	(Psychometric, dp, Lemmas)
19,417	(Negative emotion, suj, economy)	(Psychometric-dp-Lemmas)
19,421	(Reference to other, suj, Austerity)	(Psychometric, dp, Lemmas)

Models with a small number of features, such as psychometric or fine-grained part-of-speech tags, do not benefit from a larger training set, as expected. The same is not true for more complex models, which clearly improve their performance with a larger training corpus. Figure 5 illustrates how the size of the training set, following '*from large to small*' setup affects the performance of some of the models showed above. The X axis indicates the percentage of the reversed training set employed to build the model, and the Y axis corresponds the accuracy. It is important to remark that bag-of-lemmas outperforms the bag-of-words model when the training collection is small, but the performance between the two approaches converges as the training set grows. When the training set is not large enough (45% of the corpus equals 26,770 tweets) generalised triplets are not helpful to improve the accuracy of the model composed by lemmas, psychometric and sentiment information, which is the best one in these cases. But for larger training sets, the generalised-syntactic model outperforms the rest of perspectives.

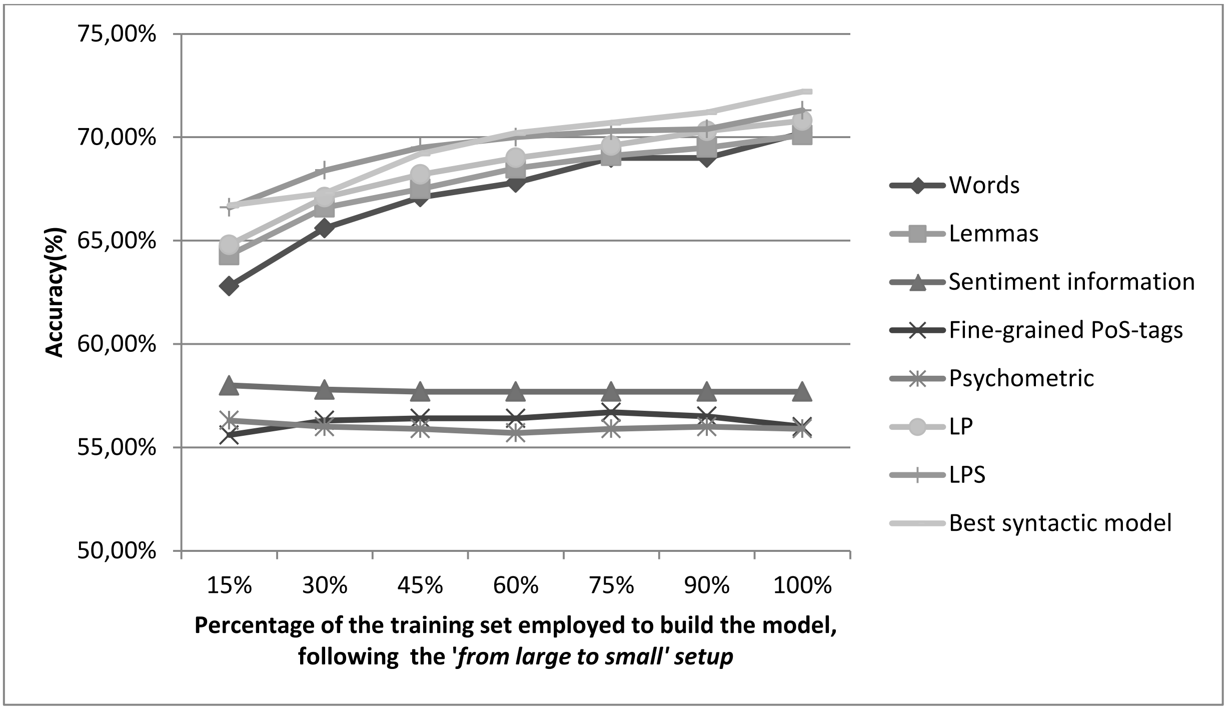


FIG 5. Performance following the *'from large to small'* setup for different models, using incremental pieces of the training collection to build them.

Regarding to the results over 4 and 6 classes, the tendency of the performance seems to be coherent with respect to the experiments over 3 classes; syntactic approaches once again outperform lexical models, and thus the discussion provided above could also be extrapolated to these runs.

In all experiments the best syntactic model obtained a statistically significant difference with respect to the best lexical model, when the training set is large enough, validating the utility of generalised dependency triplets over a wide range of polarity categories.

Conclusions and future work

This study focused on predicting the sentiment of tweets written in the Spanish language, by means of linguistics-based methods. We provided an evaluation which ranged from standard learning-based methods to shallow and deep linguistic approaches. The main contribution of the paper relies on testing how relating lexical, syntactic, psychological and semantic information affects polarity classification on tweets. To the best of our knowledge, this is the first article which performs a wide evaluation of the effectiveness of using these features, both in isolation and in combination, on a corpus of Twitter messages.

With respect to syntactic features, we rely on a more relaxed variant of the generalised dependency triplets proposed by Joshi and Penstein-Rosé (2009) to identify opinionated sentences. We adapt the method to perform polarity classification on tweets, enriching their angle by considering various levels of generalisation, ranging to part-of-speech to psychological and semantic abstraction. The utility of syntax on sentiment analysis is a widely discussed issue, but it has often been focused on long and specific domain reviews. To the best of our knowledge, this is also the first article which studies the effect of dependency parsing on Spanish tweets. Empirical results suggest that non-syntactic approaches obtain a better performance when the training set is small, but as the size of the training corpus grows, the incorporation of generalised dependency triplets helps to improve accuracy over the purely lexical perspectives.

With respect to future work, we believe that there is still room for improvement. Although syntactically relating linguistic knowledge has allowed us to achieve some improvement over purely lexical models, we believe it is possible to try to exploit syntax in a more general way. For example, the generalised dependency features are intended to identify how users relate very specific features and concepts, leading to sparse sets of features. In this respect, we propose to compute the specific polarity of each dependency subtree present in a dependency parse, to then employ this information as additional features for machine learning models, thus following a hybrid approach. As we explained, the current pre-processing of tweets is *ad-hoc*. This approach functions properly in the work presented in this article, because public figures often try to respect grammar rules in Twitter, but the same may not be true in other cases. We think text normalisation is an important issue (Vilares, Alonso & Vilares, 2013), even more so at the present time, with the rise of smartphones, where users often communicate and publish their thoughts via *texting* (text messaging from a mobile device). In this respect, we plan to adapt dependency parsing to the domain of micro-texts by applying approaches similar to those used in adapting parsers to other genres of web texts (Petrov & McDonald, 2012). Our present parser is trained from a newswire corpus and, although results suggest that even standard parsers behave rather well, we believe a specific parser for micro texts could be useful.

Acknowledgments

Research reported in this article has been partially funded by Ministerio de Economía y Competitividad and FEDER (Grant TIN2010-18552-C03-02) and by Xunta de Galicia (Grants CN2012/008, CN2012/319).

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In Proceedings of the Workshop on Languages in Social Media (LSM'11) (pp. 30-38). Portland: ACL.

Arora, S., Mayfield, E., Penstein-Rosé, C., Nyberg, E. (2010). Sentiment classification using automatically extracted subgraph features. In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. Los Angeles: ACL.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, Jan Odijk, S. Piperidis, M. Rosner, and D. Tapias (Eds.), Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta: ELRA.

Bae, Y., & Lee, H. (2012). Sentiment analysis of Twitter audiences: Measuring the positive or negative influence of popular twitterers. Journal of the American Society for Information Science and Technology, 63(2), 2521–2535.

Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., & Varma, V. (2012). Mining sentiments from tweets. In 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Proceedings of the Workshop (WASSA 2012) (pp. 11–18,). Jeju, Korea: ACL.

Batista, F., & Ribeiro, R. (2013). Sentiment analysis and topic classification based on binary maximum entropy classifiers. Procesamiento del Lenguaje Natural, 50, 77-84.

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Sebastopol, CA: O'Reilly Media.

Boiy, E. & Moens, M. F. (2009). A machine learning approach to sentiment analysis in multilingual web texts. Information Retrieval, 12(5), 526–558

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics, 21(4), 543–566.

Brooke, J., Tofiloski, M., & Taboada, M. (2009). Cross-linguistic sentiment analysis: From English to Spanish. In Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing (RANLP 2009) (pp. 50-54). Borovets, Bulgaria.

Brown, P., & Levinson S. (1987). Politeness: Some universals in language usage (Vol. 4). Cambridge: Cambridge University Press.

Campos, H. (1993). De la oración simple a la oración compuesta: curso superior de gramática española. Washington, DC: Georgetown University Press.

Carter, S., Weerkamp, W. & Tsagkias, M. (2013). Microblog Language identification: Overcoming the limitations of short, unedited and idiomatic text, Language Resources and Evaluation, 47(1), 195-215.

Choi, Y., & Cardie, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. In 2008 Conference on Empirical Methods in Natural Language, Proceedings of the Conference (pp. 793–801). Honolulu: ACL.

Costa-jussà, M. R., & Banchs, R. E. (2013). Automatic normalization of short texts by combining a statistical and rule-based techniques. Language Resources and Evaluation, 47(1), 179–193

Devitt, A., & Ahmad, K. (2013). Is there a language of sentiment? An analysis of lexical resources for sentiment analysis. Language Resources and Evaluation, 47, 475-511.

Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89.

Foster, J., Cetinglu, Ö., Wagner, J., Le Roux, J., Hogan, S., Nivre, J., Hogan, D., & van Genabith, J. (2011). #hardtoparse: POS tagging and parsing the Twitterverse. In The AAAI-11 Workshop on Analyzing Microtext. San Francisco, CA: AAAI.

Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In Proceedings of 20th International Conference on Computational Linguistics (COLING 2004) (pp. 841-847). Geneva: ACL.

Gimpel, K., Schneider, N., O'Connor, B., Das, B., Mills, D., Eisenstein, J., ... & Smith, N. A. (2011). Part-of-speech tagging for Twitter: annotation, features, and experiments. In HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers — Volume 2 (pp. 42–47). Portland:ACL.

Guo, L. & and Wan, X. (2012). Exploiting syntactic and semantic relationships between terms for opinion retrieval. Journal of the American Society for Information Science and Technology, 63(11), 2269–2282.

Greene, S., & Resnik, P. (2009). More than words: Syntactic packaging and implicit sentiment. In NAACL'09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 503–511,). Boulder: ACL.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter, 11(1), 10-18.

Han, B., & Baldwin, T. (2011). Lexical normalisation of short text messages: Makn sens a #twitter. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (pp. 368–378). Portland: ACL.

He, Y., & Zhou, D. (2011). Self-training from labeled features for sentiment analysis. Information Processing and Management, 47, 606–616.

Heerschop, B., Goossen, F., Hogenboom, A., Frasincar, F., Kaymak, U., & de Jong, F. (2011). Polarity analysis of texts using discourse structure. In Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM'11) (pp. 1061–1070). Glasgow: ACM

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent Twitter Sentiment Classification. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, USA. ACL.

Joshi, M., & Penstein-Rosé, C. (2009). Generalizing dependency features for opinion mining. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (pp. 313-316). Suntec, Singapore: ACL.

Kaufmann, M., & Kalita. J: (2010). Syntactic normalization of Twitter messages. In Proc. of 8th International Conference on Natural Language Processing (ICON 2010), Kharagpur, India.

Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence, 22(2), 110-125.

Kumar, A., & Sebastian, T. M. (2012). Sentiment analysis on Twitter. International Journal of Computer Science Issues, 9(4).

Li, Y.-M., & Li, T.-Y. (2013). Deriving market intelligence from microblogs. Decision Support Systems, 55, 206-217.

Liu, B. (2012). Sentiment Analysis and Opinion Mining, volume 16 of Synthesis Lectures on Human Language Technologies. San Rafael, CA: Morgan & Claypool Publishers.

Martínez-Cámara, E., Martín-Valdivia, M. T., Ureña-López, L. A., & Montejo-Ráez, A. (2013). Sentiment analysis in Twitter. Natural Language Engineering. Advance online publication. Doi: 10.1017/S1351324912000332

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. International Journal of Lexicography, 3(4), 235–244.

Mitchell, Tom M. (1997). Machine Learning. Mc-Graw-Hill.

Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., & Ureña-López, L.A. (2012). Random walk weighting over SentiWordNet for sentiment polarity detection on Twitter. In 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Proceedings of the Workshop (WASSA 2012) (pp. 3–10). Jeju, Korea: ACL.

Moilanen, K., & Pulman, S. (2007). Sentiment composition. In G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, & N. Nikolov (Eds.), International Conference Recent Advances in Natural Language Processing, Proceedings (RANLP 2007) (pp. 378–382). Borovets, Bulgaria.

Nakagawa, T., Inui, K., & Kurohashi, S. (2010). Dependency tree-based sentiment classification using CRFs with hidden variables. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010) (pp. 786-794). Los Angeles: ACL.

Nivre, J. (2008). Algorithms for deterministic incremental dependency parsing. Computational Linguistics, 34(4), 513-553.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., & Schneider, N. (2012). Part-of-speech tagging for Twitter: Word clusters and other adavances. Technical Report CMU-ML-12-107, School of Computer Science. Pittsburg, PA: Carnegie Mellon University.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., S. Marinov & Marsi, E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. Natural Language Engineering, 13(2), 95-135.

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010). Valletta, Malta: ELRA.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL'05) (pp. 115-124). Ann Arbor, MI: ACL.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (EMNLP'02) (pp. 79-86). Philadelphia: ACL.

Petrov, S. & McDonald, R. (2012). Overview of the 2012 Shared Task on Parsing the Web. In Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL), Montreal: ACL.

Platt, J. C. (1999). Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), Adavances in kernel methods: support vector learning (pp. 185-208). Cambridge, MA: MIT Press.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. Mahwah, NJ: Lawrence Erlbaum Associates.

Perea-Ortega, J. M., Martín-Valdivia, M. T., Ureña-López, L. A., & Martínez-Cámara, E. (2013). Improving polarity classification of bilingual parallel corpora combining machine learning and semantic orientation approaches. Journal of the American Society for Information Science and Technology, 64(9), 1864-1877.

Ramírez-Esparza, N., Pennebaker, J. W., García, F. A., & Suriá Martínez, R. (2007). La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. Revista Mexicana de Psicología. 24(1), 85-99.

Shaikh, M. A., Prendinger, H., & Mitsuru, I. (2007). Assessing sentiment of text by semantic dependency and contextual valence analysis. In A. Paiva, R. Prada, & R. W. Picard (Eds.), Affective Computing and Intelligent Interaction, volume 4738 of Lecture Notes in Computer Science (pp. 191–202). Berlin: Springer.

Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., I., Díaz-Ángel, S., Suárez-Guerra, A. Treviño & Gordon, J. (2013). Empirical study of machine learning based approach for opinion mining in tweets. In I. Batyrshin, & M. González Mendoza (Eds.), Advances in Artificial Intelligence (pp. 1-14). Berlin: Springer.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13) (pp. 1631–1642). Seattle: ACL.

Spencer, J. & Uchyigit, G. (2012). Sentimentor: Sentiment analysis of Twitter data. In M. M. Gaber, M. Cocea, S. Weibelzahl, E. Menasalvas & C. Labbe (Eds.), The 1st International Workshop on Sentiment Discovery from Affective Data (SDAD 2012) (pp. 56–66), Bristol, UK.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

Taulé, M., Martí, M. A., & Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In Proceedings of 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakesh, Morocco: ELRA.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544-2558.

Thelwall, M., Buckley, K., & Paltoglou, G. (2010). Sentiment in Twitter events. Journal of the American Society for Information Science and Technology, 62(2), 406-418.

Thelwall, M., Buckley, K. & Paltoglou, G. (2012). Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology, 63(1), 163-173

Trnavac, R., & Taboada, M. (2012). The contribution of nonveridical rhetorical relations to evaluation in discourse. Language Sciences, 34(3), 301-318.

Turney, P. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In L. de Raedt, & P. Flach (Eds.), Machine Learning: EMCL 2001 (pp. 491-502). Berlin: Springer.

Turney, P.D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (ACL'02) (pp. 417-424). Philadelphia: ACL.

Twitter, Inc. (2013). Form S-1 Registration Statement under The Securities Act of 1933. Washington, D.C.: United States Securities and Exchange Comission. Retrieved October 29, 2013, from http://www.sec.gov/Archives/edgar/data/1418091/ 000119312513390321/d564001ds1.htm

Vilares, D., Alonso, M.A., & Gómez-Rodríguez, C. (2013). Supervised polarity classification of Spanish tweets based on linguistic knowledge. In *Proceedings of the 2013 ACM symposium on Document engineering* (pp. 169-172). Florence, Italy: ACM.

Vilares, D., Alonso, M.A., & Gómez-Rodríguez, C. (2013). A syntactic approach for opinion mining on Spanish reviews. Natural Language Engineering, Advanced online publication. Doi: 10.1017/S1351324913000181

Vilares, J., Alonso, M.A. & Vilares, D. (2013) Prototipado Rápido de un Sistema de Normalización de Tuits: Una Aproximación Léxica. In A. Díaz Esteban, I. Alegría Loinaz and J. Villena Román (Eds.), Tweet Normalization Workshop at SEPLN 2013 (pp. 76-80). Madrid: SEPLN.

Villena-Román, J., & García-Morera, J., (2013). TASS 2013-Workshop on Sentiment Analysis at SEPLN 2013: An overview. In Díaz-Esteban, A., Alegría I., & Villena-Román J. (Ed.), Proceedings of the TASS workshop at SEPLN 2013. Actas del XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural. IV Congreso Español de Informática (pp. 112–125). Madrid: SEPLN.

Villena-Román, J., Lana Serrano, S., Martínez Cámara, E., & González Cristóbal, J. C. (2013). TASS workshop on sentiment analysis at SEPLN. Procesamiento del Lenguaje Natural, 50, 37-44.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12) (pp. 115–120). Jeju, Republic of Korea: ACL.

Wu, Y., Zhang, Q., Huang, X., & Wu, L. (2009). Phrase dependency parsing for opinion mining. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09) (pp. 1533–1541). Singapore: ACL.

Zhang, C., Zeng, D., Li, J., Wang, F. Y., & Zuo, W. (2009). Sentiment analysis of Chinese documents: From sentence to document level. Journal of the American Society for Information Science and Technology, 60(12), 2474-2487.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. Technical Report HPL-2011-89. Palo Alto, CA: HP Laboratories.

Footnote 1: www.epinions.com

Footnote 2: www.tripadvisor.com

Footnote 3: We offer a direct translation of the original Spanish tweet into English: '*@User la vi en Londres Amiguete...fliparás con Kevin !!! Hace un malvado delicioso!! Abrazo!!*.

Footnote 4: The original sentence in the Spanish language is '*La transferencia fue cordial, pero no completa*'.

Footnote 5: Sociedad Española del Procesamiento del Lenguaje Natural http://www.sepln.org/