Article

JIS

Journal of Information Science
1–19

DOI: 10.1177/0165551510000000
jis.sagepub.com
SAGE

A linguistic approach for determining the topics of Spanish Twitter messages

David Vilares
Grupo LYS, Departamento de Computación, Facultade de Informática, Universidade da Coruña

Miguel A. Alonso
Grupo LYS, Departamento de Computación, Facultade de Informática, Universidade da Coruña

Carlos Gómez-Rodríguez
Grupo LYS, Departamento de Computación, Facultade de Informática, Universidade da Coruña

Abstract
The vast amount of opinions and reviews provided in Twitter is helpful in order to make interesting findings about a given industry, but given the huge number of messages published every day it is important to detect the relevant ones. In this respect, the Twitter search functionality is not a practical tool when we want to poll messages dealing with a given set of general topics. This article presents an approach to classify Twitter messages into various topics. We tackle the problem from a linguistic angle, taking into account part-of-speech, syntactic and semantic information, showing how language processing techniques should be adapted to deal with the informal language present in Twitter messages. The TASS 2013 General corpus, a collection of tweets which has been specifically annotated to perform text analytics tasks, is used as the dataset in our evaluation framework. We carry out a wide range of experiments to determine which kinds of linguistic information have the greatest impact on this task and how they should be combined in order to obtain the best-performing system. The results lead us to conclude that relating features by means of contextual information adds complementary knowledge over pure lexical models, making it possible to outperform them on standard metrics for multi-label classification tasks.

Keywords
Twitter; natural language processing; multi-label topic classification

1. Introduction

Social media are a meeting point where users can share their views about politics, events, technology, films and many other topics. Blogs, forums and social networks are some of the most popular examples where anybody can find opinions about virtually any subject. In this context, 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. By the end of 2013, its more than 100 million daily active users were producing 500 million tweets per day [1].

The task of analysing and comprehending all this information is becoming a need for companies in order to know directly from the source what is being said about them and their industry. For this purpose, they often rely on opinion mining applications for making better decisions, identifying key thoughts about their area of influence and even predicting their performance in the stock market [2, 3, 4]. One of the main issues is that many of the messages under analysis are not useful for the task because they deal with unrelated topics. This may not be a serious issue in specialised forums, but it becomes a real problem when monitoring media such as Twitter, where users publish comments about all kinds of topics. In this context, applying filtering steps is necessary to be able to exploit the messages in this social network, discriminating unrelated opinions and reducing the amount of traffic to analyse. For example, for a firm in the

Corresponding author:
David Vilares, Universidade da Coruña, Facultade de Informática, Campus de Elviña, 15071 A Coruña, Spain
david.vilares@udc.es
This version of the article has been accepted for publication, after peer review, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1177/0165551514561652

motion picture industry wishing to retrieve Twitter messages about a given movie, a search based on the film title can give as a result a lot of irrelevant messages in which the words appearing in the title are used in contexts unrelated to the movie domain [5].

Moreover, opinions can involve different topics, so traditional single-label classification systems are not appropriate to correctly deal with topic categorisation, as users often tend to relate different subjects in the same message. The following lines illustrate some real examples of tweets annotated with their topics[1] where a few topics are linked in different ways:

- '*The key to the new government: its structure. Will there be two deputy prime ministerships or not? The key, in the economic team*'.[2] This tweet explicitly relates two close topics: politics and economy.
- '*The intelligent public is on social networks. Education determines their use more than wealth. Impact on media*'.[3] This message contains information about technology, referring to social media, which also often represent a way of entertainment. Finally, the tweet was also annotated with the economy label due to making a reference to the concept of wealth.
- '*Hii my tweeps! A wonderful day to do a twitcam as I promised it will be at 6:00 pm (Mexico time) see you soon!!!*'.[4] The tweet was assigned to the music and "other" topics. Although a priori it has nothing to do with music, we must take into account this tweet was addressed to his Mexican fans by the Spanish artist Alejandro Sanz (@alejandrosanz).

In order to address these issues, we propose a multi-topic classification approach for Twitter messages. We rely on linguistic information, managing lexical, syntactic, psychological and semantic knowledge by means of a Natural Language Processing (NLP) pipeline that includes pre-processing, tagging and parsing steps. Linguistic processing of Twitter messages is particularly challenging, as they are characterised by the use of a very informal language combined with specific Twitter elements (e.g., @userMentions, #hashtags, RT, FF, etc.), but also including formal expressions, figures of speech (e.g., oxymoron) and rhetorical devices (e.g., sarcasm). Therefore, well-performing techniques for lexical and syntactic processing of regular texts do not behave as well when confronted with Twitter messages, needing an adaptation to this new kind of text genre.

The usefulness of our proposal has been tested at TASS 2013 [6], an evaluation workshop on sentiment analysis for Spanish language, where an initial implementation of our approach achieved the first place in the topic classification task. In this article we show how NLP techniques should be adapted in order to deal with the informal language present in micro-texts successfully. Then, we analyse our system in order to determine which kinds of linguistic information have the greatest impact on its success. Experimental results suggest that relating terms and concepts in a novel way allows us to improve the results for multi-topic classification; and using morphological and syntactic information obtained from NLP techniques provides an extra boost to accuracy over what can be attained with traditional models based on word presence, adjacency or proximity. An advantage of our approach is that it does not need any information external to the texts themselves, making it flexible enough to be applied to social media other than Twitter.

The remainder of this article is organised as follows. In Sect. 2 we present related research about topic classification, focussing on Twitter. In Sect. 3 we motivate the aim of this article, posing the research questions. We explain the details of our linguistic perspective for performing multi-topic classification over tweets in Sect. 4. Then, we establish the experimental setup in Sect. 5. Experimental results are presented and discussed in Sects. 6 and 7, respectively. Finally, we present our conclusions and discuss future work.

2. Related work

Text categorisation has been typically regarded as an application of supervised learning for classification, in which the features happen to represent documents [7, chapter 10]. In this framework, a general inductive process automatically builds a classifier by learning, from a set of preclassified documents, the characteristics of each category. Sebastiani [8] enumerates the advantages of this approach over the knowledge engineering approach (consisting in the manual definition of a classifier by domain experts): very good effectiveness, considerable savings in terms of expert labour power, and straightforward portability to different domains.

In the last few years, this framework has been used to classify micro-texts present in social networks, especially Twitter, where the short length (140 characters) and the informal language used in its messages poses new challenges. Moreover, classification techniques have been applied to Twitter messages mainly in order to perform tasks that differ from topic classification, such as determining whether a tweet refers to a given entity or not [9, 10], or determining

whether a tweet carries an opinion or not, and in the case of an affirmative answer, determining its polarity [11], i.e, if the opinion is positive, negative or neutral [12, 13, 14].

2.1. Categorisation of Twitter messages

Most work on categorizing Twitter messages is based on classical bag-of-words approaches without taking into account linguistic information other than surface forms appearing in texts. In this context, Sriram et al. [15] apply a bag-of-words model for classification of Twitter messages into five categories (News, Events, Opinions, Deals, and PrivateMessages), also incorporating author name and seven binary features: presence of shortened forms of words and slang, time-event phrases, opinionated words, emphasis on words, currency and percentage signs, @username at the beginning of a tweet, and @username within a tweet. They perform experiments on a corpus composed of 5 407 tweets from 684 authors and find that the author feature is the most discriminatory one.

Some authors propose to apply techniques for dimensionality reduction. Thongsuk et al. [16] propose a method for classifying Twitter messages into three categories (Airline, Food and Computer&Technology) by using Latent Dirichlet Allocation (LDA [17]) to cluster words extracted from tweets into a set of 50 topics, which are then used as features of an SVM classifier [18]. They perform some experiments on a purpose-specific corpus, reporting a significant improvement with respect to a baseline provided by a bag-of-words approach. LDA was also used by Hong and Davison [19] for topic modelling on Twitter. They apply the topic models to classify users and corresponding messages into 16 categories (Art&Design, Books, Business, Charity, Entertainment, Family, Fashion, Food&Drink, Funny, Health, Music, News, Politics, Science, Sports, and Technology) showing that models trained on aggregate longer texts (i.e., texts resulting from combining all messages generated by the same user or from aggregating all the messages that contain a given term) yield better performance than models trained on single messages. Therefore, their work is in a middle ground between short-text categorization and classical classification of longer texts.

The fact that a classifier can perform better than another on a given topic but worse on a different topic is exploited by Fiaidhi et al. [20] by training four different classifiers on 12 classes of trending topics (Politics, Education, Health, Marketing, Music, News&Media, Recreation&Sports, Computers&Technology, Pets, Food, Family and Other) and defining a method for choosing the best classifier for each class.

In a certain way, Twitter resembles an information network, and in the same way as the classification of news was a classic problem in professional journalism, tweet classification into news categories is a current relevant problem for online content creators. In this context, Bastos et al. [21] propose 12 Twitter topics, on a scale from hard to soft news: Politics, Altruism, Events, Technology, Games, Idioms, Music Personality, Movies, Celebrity, Lifestyle and Sports; and explore the distribution of messages across these topics, without performing linguistic processing. Lee et al. [22] classify Twitter trending topics into 18 categories (Arts&Design, Books, Business, Charity&Deals, Fashion, Food&Drink, Health, Holidays&Dates, Humor, Music, Politics, Religion, Science, Sports, Technology, TV&Movies, OtherNews and Other) using a bag-of-words model and a decision tree learner, testing it on a corpus of 768 trending topics. It is worth noting that they do not perform classification on individual tweets, instead, they save up to 1 000 tweets corresponding to each trending topic in a document and then classify this large document into one of the 18 predefined categories. As a consequence, their work has more similarities with classical text classification than with tweet classification.

Some works on classification of short messages integrate the textual content of each message with information from other sources such as hyperlinks or Wikipedia[5] pages. Gene et al. [23] develop a technique for calculating distances between messages based on the distance between their closest Wikipedia pages, regarding Wikipedia as a transform space in which measurements can be made more accurately. They perform experiments on a small corpus of 100 tweets into three categories. Gattani et al. [24] describe a real-time, end-to-end industrial system for Twitter processing involving entity extraction, linking, classification and tagging. In the case of topic classification, they define a set of 23 predefined topics corresponding to the child nodes of the root of the taxonomy of a knowledge base built out of Wikipedia, but also enriched with a variety of structured data sources. Their approach was tested on a small collection of 99 tweets, with a per-topic performance ranging from 100% precision and recall for topics such as Environment or Travel to 0% recall for composite topics such as Food&Agriculture or Home&Leisure. Kinsella et al. [25] investigate topic classification in social media by using textual content and metadata retrieved from external hyperlinks. Their work does not focus on Twitter messages but on longer ones. In particular, they perform experiments on classifying 6 626 user-generated posts submitted to a message board into general categories (Photography, Soccer, Musicians, Movies, Politics, MartialArts, Motors, Poker, Television, Atheism) and 1 564 posts in music categories (Rock, Electronic, Alternative, Hip-Hop, Punk) where the category of each post is determined by the forum in which it is published. They

find a large variation in classification performance across different categories, e.g., classification of Musicians is trivial since almost all posts have a link to MySpace[6] but classification of Television posts is challenging because messages in this category cover any topic which is broadcast. Scores for music categories are also lower than for general categories due to the higher similarity of the topics.

2.2. Categorisation of Twitter messages on the TASS corpus

The corpora used in the works cited so far exhibit a great variety in their size and topics, making a fair comparison of results difficult, if not impossible. The TASS corpus [26, 6] is a standard corpus for topic categorisation of Spanish Twitter messages, where each tweet is assigned at least one topic from a list including Politics, Entertainment, Economy, Music, Soccer, Films, Technology, Sports (other than soccer), Literature and Other.[7] Several authors have performed experiments on this corpus using a variety of approaches.

Batista and Ribeiro [27] rely on logistic regression classification models, which correspond to maximum entropy classification for independent events, creating a binary classifier for each topic that estimates the probability of a tweet pertaining to that topic or not. For each tweet, its unigrams and bigrams are submitted to all classifiers and the most probable topic is selected based on these classification probabilities. A similar approach is undertaken by Pla and Hurtado [28]: a cascade of binary SMO classifiers [29] trained for each topic is applied to each tweet, determining if a message belongs to it or not. As a result, they obtain as output the topics assigned by at least one classifier. In the case where no topic is assigned to a tweet, a cascade of libSVM classifiers [18] is used as back-off: as each libSVM classifier provides a weight for each possible assignment of a class to the tweet, the tweet is assigned the topic which provides the highest probability estimate. Martínez-Cámara et al. [30] also apply a SVM classifier to a bag-of-words enriched with selected hashtags and related words obtained from Google AdWords KeyWord Tool.[8]

Martín-Wanton and Carrillo de Albornoz [31] build for each topic a lexicon of words that best describe it, in terms of Kullback-Leibler Divergence (KLD [7, chapter 9]), thus representing each topic as a ranking of discriminative words. Moreover, a set of events is defined according to an LDA model. To determine which of the topics corresponds to each event, the topic with the highest statistical correlation is obtained by comparing the ranking of words of each topic and the ranking of words most likely to belong to the event. KLD is also used in the construction of language models by Castellano González et al. [32]. They address the challenge posed by topic categorisation from an Information Retrieval (IR [7]) perspective, where the training set of tweets is represented and indexed using language models. To determine the category of an unseen tweet, its content is used as a query against the index storing the previously built models. They extend their work in [33], where they analyse the type of tweet information to be used in the classification and which process should be followed to take this information into account, proposing different types of modelling as well as different ways of performing the information retrieval process according to the different types of information. The results suggest that named entities only help in the classification of a small number of tweets, so that it is necessary to use the overall bag of words to attain average performance. An IR perspective is also applied by Montejo-Ráez et al. [34] by defining a text processing component to convert tweets into vectors according to the Vector Space Model with tf.idf weighting [7, chapter 2] after a normalisation process. Then, an automatic extraction system is used to generate a term-feeling matrix from affective Twitter posts, whose dimension is reduced by means of Latent Semantic Indexing [35, chapter 18]. The approach shows a poor performance, due to the reduced size of the training set, according to its authors.

Fernández Anta et al. [36] apply several machine learning classifiers on a portion of the corpus in order to test the utility of n-grams, stemming, lemmatisation, word correction, presence of hashtags, presence of user references and presence of hyperlinks, concluding that none of them make a clear difference when introduced in the classification framework. Trilla and Alías [37] adapt a text classification scheme based on Multinomial Naive Bayes to deal with Twitter messages, using a binary-weighted feature space, which is unable to predict the class with the fewest number of training examples but performs well on other classes. Another variant of Naive Bayes, Naive Bayes Complement (NBC, [38]), is applied by Rufo Mendo [39]. In particular, he presents a co-training version of CNB consisting of four steps: 1) train a CNB classifier with the training dataset; 2) classify the test dataset and take tweets with a confidence value higher than 0.9; 3) train the CNB with the training dataset and the tweets obtained in step 2; and 4) classify messages in the test dataset. The results obtained are not very encouraging. In particular, results obtained by co-training CNB are far worse than the results obtained by CNB without co-training.

Cordobés et al. [40] present a technique based on graph similarity to classify Twitter messages as being related to a specific topic. Each vertex in their graphs is a word stem, with weighted arcs representing the frequency of joint occurrence of two given stems, which can be pondered by the distance between the two stems in parse trees. In the

training phase, a representative graph for each topic is built as the union of the graphs for the corresponding tweets. Then, for each tweet we wish to classify in the test phase, its graph is constructed and the most similar reference graph is retrieved using metrics inspired in PageRank [41] and HITS [42]. As a consequence, only one category can be assigned to each given tweet, and thus this method does not perform multi-topic classification. An interesting result they have found is that using synonyms in the construction and retrieval of graphs has a negative impact on performance.

It is interesting to point out that, although the problem of topic categorisation has been attacked from a great variety of angles on the TASS corpus, no approach has attained the level of performance shown by the system that we analyse in this article.

3. Motivation

Classification of Twitter messages is a radically different problem with respect to the classification of long texts that has been the focus of research effort during the last 15 years. Twitter messages have several characteristics that make them especially challenging. On the one hand, their very small size, limited to 140 characters, makes classification difficult because there is very little information to rely on - for example, a bag-of-word model for tweets will typically need to work with less than 10 input words, as opposed to hundreds or thousands of words in typical larger documents. On the other hand, tweets use a highly non-standard language (including informal abbreviations, spelling mistakes, and Twitter-specific elements like user mentions, hashtags, retweets, specific acronyms like FF, or URLs) that makes their processing more difficult, not only with standard machine learning approaches - where features will be sparser, especially without proper pre-processing - but even more so when linguistic processing is applied, as NLP for tweets is a difficult problem in itself that has only begun to be tackled in the last few years. Apart from the difficulty of the problem, the classification of Twitter messages is very relevant in practice, in the light of the growing importance that this social medium has achieved in recent years. For example, Twitter was decisive in recent political events, like those of the Arab Spring. In Western countries, it is a social thermometer that can be used to measure public opinion on a wide range of subjects, from the impact of political decisions to the success of TV programs.

It is in this context that TASS evaluation framework has emerged. TASS is the acronym of an experimental evaluation workshop for sentiment analysis and online reputation analysis focused on Spanish language [6]. Its main objective is to promote the design of new algorithms and techniques for the implementation of complex systems able to perform sentiment analysis and text classification on short text opinions extracted from Twitter, published by public figures (e.g. politicians, journalists or athletes). The setup is based on a series of challenging tasks that are intended to provide a benchmark forum for comparing the latest approaches in these fields. In addition, with the creation and release of the fully tagged corpus, it aims to provide a benchmark dataset that enables researchers to compare their algorithms and systems.

Table 1. Official scores of the best submission of each research group to TASS evaluation framework, comprising both 2012 and 2013 editions. A detailed description of some of these systems can be found in section 2.2.

Team	Official TASS score	Edition with best results
Our initial proposal [43]	**0.786**	**2013**
UNED-LSI (Castellano González et al.) [32]	0.777	2013
UPV (Plan and Hurtado) [28]	0.756	2013
ETH-ZURICH [44]	0.734	2013
FHC25-IMDEA (Cordobés et al.) [40]	0.719	2013
L2F-INESC (Batista and Ribeiro) [27]	0.654	2012
LASALLE-URL (Trilla and Alías) [37]	0.602	2012
TUDELFT	0.563	2012
UNED-JRM (Rufo Mendo) [39]	0.479	2013
SINAI-UJAEN (Martínez-Cámara et al.) [30]	0.394	2012
LSI UNED (Martín-Wanton and Carrillo de Albornoz) [31]	0.310	2012
SINAI-CESA (Montejo-Ráez et al.) [34]	0.160	2013

In its 2013 edition, our submission achieved the best performance in the topic classification task [6]. The challenge consisted of finding out the subject(s) that a tweet is talking about. We presented an initial topic classification system [43], based on a linguistic perspective, which did not take into account any external meta-data information, so it could also be applied to other social media. Since a tweet can refer to various topics, the task was addressed as a multi-label classification problem. Table 1 summarises the official results provided by the TASS organisation for the 2012 and 2013 editions.

The results for the initial implementation of our system were briefly presented in [43], but there was no analysis of why the system obtained such a high performance. We believe that performing a detailed analysis of the contribution of each technique will allow us to gain insight into understanding the problem of topic categorisation of Twitter messages and that it will allow other researchers to improve the performance of their own systems. Moreover, in this article we are going deeper into the treatment of linguistic information, showing the utilities, advantages and drawbacks of this approach.

In particular, in this article we address the following research questions with respect to topic categorization of Twitter messages:

- What is the contribution of morphological, syntactic and psycho-metric knowledge to this task?
- Does the use of context information help improve performance?
- Is it advisable to apply feature selection techniques when training a classifier for this task?

4. Natural language processing for Twitter messages

We rely on a linguistic-based approach to extract features for a supervised classifier, considering lexical, syntactic, psychological and semantic information. For this purpose, we first apply a Natural Language Processing (NLP) pipeline to each tweet, including pre-processing, tagging and parsing steps. No external information is used, so the approach could also be applied to microtext-based social media other than Twitter. Figure 1 illustrates the high-level structure of our topic classification system.

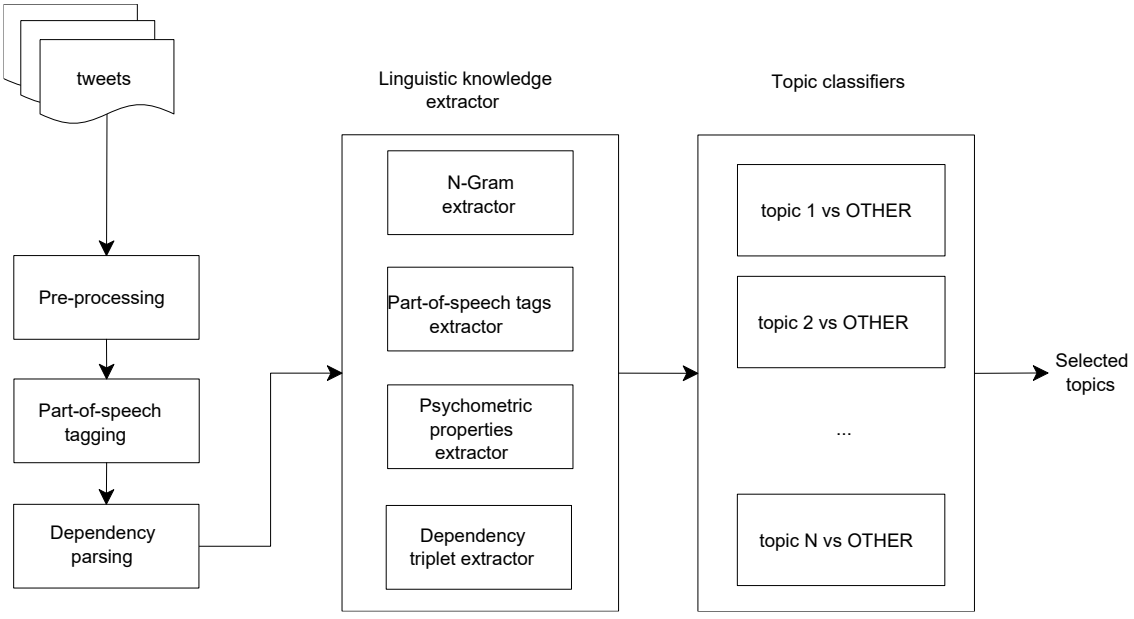

Figure 1. High-level architecture of the topic classification system.

4.1. Pre-processing

Twitter is characterised by the use of a very informal language, and so we need to pre-process each tweet in order to normalise some lexical text features that are usually omitted or not respected. In particular, we homogenise punctuation marks (period, semicolon…) by adding blanks when required. We must take into account that Twitter users often avoid spaces between a punctuation mark and the next word. This may affect the subsequent processing steps such as segmentation, tokenisation or PoS-tagging. For example, following our pre-processing algorithm; the phrase '*Hi,Jack!*' should be represented as '*Hi, Jack!*'

Some special Twitter characters such as '@' (user names) or '#' (hashtags) may negatively affect the performance of the rest of the natural language processing steps. This kind of symbols is especially confusing for tools such as the tokeniser or the parser. Thus, if the beginning or the end of the tweet contains a hashtag, we delete it completely. We hypothesise that this type of hashtag is only used to label the tweet, without providing any type of syntactic information.[9] If the hashtag is at some other position in the tweet, the pre-processor simply deletes the '#' because it considers the hashtag refers to a relevant word or entity in the sentence (e.g. the sentence '*The #screen is really good*' would be converted into '*The screen is really good*')

With respect to usernames, we just delete the '@' and capitalise the first letter, so that they will be treated as if they were real proper nouns.

4.2. PoS 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.). For the purpose of PoS tagging, we rely on a Brill tagger [45] implementation to label tweets. Concretely, we employed the version included in NLTK[10] a framework for natural language processing written in Python.[11] To train it, we used the Ancora corpus [46], a collection of tagged texts extracted from newspapers. Following common practice in the natural language processing community, 90% of the corpus was used as the training set and the remaining 10% as the test set.

The informal language used in Twitter messages is a source of confusion for a tagger. In this respect, an important issue in Spanish language is the use of acute accents, which are frequently omitted in web environments. For example, the noun '*cálculo'* ('*calculus'*) can be confused with the verb forms '*calculo'* ('*I figure*') or '*calculó'* ('*She computed*') when the acute accent is removed. The success of Twitter has prompted researchers to develop specific part-of-speech taggers for labelling tweets, mainly for English [47], but to the best of our knowledge this kind of tools are not available for the Spanish language. As a consequence, we have decided to consider a novel approach based on applying transformations to the training corpus in order to make it more similar to the kinds of text we found in Twitter. This approach has been recently tested successfully on user-generated web content: user reviews on cars, washing machines, books, cell phones, music, computers, movies and hotels [48].

In particular, we expanded the training set of the corpus in such a way that each text was duplicated without including any accent. We trained two models: the first one considering the original training set, and the second one taking into account the expanded set. The expanded training corpus makes the tagger able to handle words whose diacritical accents have been omitted due to the characteristics of the language used in web environments. Thus, the new tagger is able to solve the ambiguities created by omitted accents, as it has seen training examples of them, and it can rely on the word context to tell between correctly spelled forms and those where an accent has been omitted. In contrast, the regular tagger has a stronger dependence on the form of the word, which is a weakness when the diacritical accent is not found. For example, the phrase '*el medio jugo tan bien como siempre*' could be translated literally as '*the half juice, as good as usual*', completely unrelated to the sports category, if we consider '*jugo'* as a noun ('*juice'*) and '*medio'* ('*half*') as an adjective. However, the intended meaning corresponds to '*the midfielder played as good as usual*', since '*jugo'* is a misspelled form (without diacritical accent) of '*jugó'*, past tense form of the verb '*jugar'* ('*to play*' sports of games, but not performing arts), and '*medio'* is the noun for 'midfielder'.

The evaluation on the test set showed that both taggers perform very similarly on regular texts (with accuracies of 95.86% and 95.71% respectively), but practical performance on tweets was much better for the tagger built with the expanded training set, as detailed in [48, Section 3].

4.3. Dependency parsing

We obtain the syntactic structure of tweets by means of dependency parsing. Given a sentence $S = w_1...w_n$, where w_i represents the word at the position i in the sentence, a dependency parser returns a dependency tree, a set of triplets $\{(w_i, arc_{ij}, w_j)\}$ where w_i is the head term, w_j is the dependent and arc_{ij} represents the dependency type, which denotes the syntactic function that relates the head and the dependent. For example, for the simple sentence '*I eat peanuts*' we obtain the triplets ('*eat'*, subject, '*I*') and ('*eat'*, direct object, '*peanuts'*).

We rely on MaltParser [49], a data-driven dependency parser generator, to build our parser. The Ancora corpus was also used to train the model. We again employed 90% of the corpus as the training set, and the remaining 10% as the test set, achieving an LAS[12] of 83.75% and a UAS[13] of 88.16%, which is coherent with the state of the art for Spanish language.

4.4. Psychometric Knowledge

We also consider a perspective based on psychological knowledge. For this purpose, we rely on the dictionaries presented by Ramírez-Esparza et al. [50]. This lexicon distinguishes around 70 dimensions of human language. It provides information about psychometric properties of words (cognition mechanisms, anxiety, feelings, inhibition, sexuality, etc.), or even pragmatic information (words referring to oneself, 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 [51].

4.5. Extraction of linguistic features

The different kinds of linguistic information obtained after the application of the NLP pipeline will be used as features for automatic classification.

The basic model, usually called unigram model, considers each single word as a feature. This often becomes a simple baseline which obtains a decent performance, although it has an important disadvantage: it does not consider enough context. We will use the tweet '*The fifth bowl. Nadal saluting to the Argentine team. And now to the public. Oe, oe, oe*'[14] as running example to illustrate the theoretical advantages and disadvantages of some of the proposed set of features, as far as possible. We can think that the word '*team'* is a relevant feature to identify a tweet dealing with sports, but it could also refer economics or politics if the phrase '*economic team*' appears. Furthermore, we can easily propose an example where the words '*team*' and '*economic*' appear, but referring to a sports tweet: '*His favourite team does not give any economic benefit*s'. Therefore, considering each single word as a feature to feed the classifier is not the best strategy to correctly deal with topic classification.

A way to solve this issue is to employ larger n-grams[15], such as bigrams, i.e., sequences of two contiguous words. The main drawback of these types of features is their sparsity, which may decrease performance if the training set is not representative and large enough. Continuing with our running example, instead the unigram '*team*' we would have the bigram '*Argentine team*', which is probably a better indicator of a sports tweet. And the same would be true if we again take the phrase '*economic team*', where we now obtain a unique and more discriminative bigram, instead of separate features that may confuse the classifier, as we detailed above.

We also take part-of-speech tags as a possible complement to improve the performance of other sets of features. PoS-tags are not intuitively proper features, by themselves, to difference between various topics. However, some of them can be helpful when they are used jointly with other attributes. For example, the proper name tag can be more frequent in domains such as movies, where users often refer to the director, the actors or the producer to implicitly refer to the quality of the art, than in other domains such as technology, for example.

Moreover, we consider the extraction of the *psychometric properties* present in a text. We identify psychological features relying on LIWC dictionaries [51]. These lexicons allow us to significantly reduce the number of features, avoiding sparsity problems and assigning words to very specific areas. Their main drawback is their low recall, limited to the terms considered in the lexicons.

Finally, we propose relating terms by means of dependency parsing. More specifically, we rely on the concept of *generalised dependency triplets*, originally presented as *back-off dependency triplets* [52], and later enhanced and applied successfully to polarity classification in [14], where the authors obtained a significant improvement over a pure bag-of-words approach. Given a triplet (w_i, arc_{ij}, w_j), generalising it consists of abstracting the information provided by either the term i or the term j, obtaining a more general dependency triplet of the form $(b(w_i, x), arc_{ij}, b(w_j, x))$, where b is a generalization function and x the parameter indicating what to return: a part-of-speech tag, a psychometric property,

a lemma, a word form or even nothing (completely eliminating that element). We also consider the option of removing the dependency type arc_{ij} in order to further reduce sparsity, obtaining a syntactic bigram [52]: two words, not necessarily contiguous, related by a syntactic relation. If we again take our running example, the dependency triplet (*team*, modifier, *Argentine*) could be generalised in different ways, as shown in Table 2.

Table 2. Examples of generalisation for the dependency triplet (*team*, modifier, *Argentine*).

Desired generalisation	Result
(b(team, PoS-tag), delete(modifier), Argentine)	(common name, , Argentine)
(b(team, psychometry), modifier, Argentine)	(leisure, modifier, Argentine) (sports, modifier, Argentine) …
(b(team, psychometry), modifier, b(Argentine, PoS-tag))	(leisure, modifier, adjective) (sports, modifier, adjective) …

5. Experimental setup

5.1. Dataset description

The TASS 2013 General Corpus is a collection of tweets which has been specifically annotated to perform text analytics tasks. It was presented at the Workshop on Sentiment Analysis at SEPLN [6]. It is a collection of tweets in Spanish written by 150 public figures, such as soccer players, politicians or journalists from Spain, Mexico, Colombia, Puerto Rico, USA and other countries. Message dates 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 or more topics, which involve up to 10 categories: films, soccer, economics, entertainment, literature, music, politics, sports (other than soccer), technology and other. We take the "other" class as the default class (i.e., if our one vs. all strategy always discards the topic under study, the system will assign the default topic to the tweet). The gold standard has been generated by a pooling of the submissions, followed by a human review by TASS organisation for the thousands of ambiguous cases. Appendix A and Appendix B show the topic distribution of tweets in the collection, for both training and test sets. The classes of the training set are unbalanced. This may be interesting from a real-world environment and industry point of view, since some topics are often more popular than others, and therefore it may be difficult to build a balanced training set. In this situation, from a performance perspective, supervised methods tend to present biases when there are large differences in the number of training samples for each class. Thus, we decided to apply oversampling to the minority categories.

5.2. Evaluation metrics

We evaluate our approaches by means of the standard metrics for multi- label classification: Hamming loss (HL), label-based accuracy (LBA) and exact match (EM). They are calculated according to Equations 1, 2 and 3, where:

- L is the set of labels.
- D is the set of instances of the collection.
- Y_i is the set of the labels expected for an instance i.
- Z_i is the set of labels predicted for an instance i.
- $\triangle$ is the symmetric difference operation between sets.

$$Hamming\ loss = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i\,\Delta\,Z_i|}{|L|} \tag{1}$$

$$Label\ based\ accuracy = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \tag{2}$$

$$Exact\ match = \frac{\#instances\ exactly\ classified}{\#instances} \tag{3}$$

These metrics reflect different aspects whose relevance should depend on the type of application. We will illustrate the behaviour of these metrics with the following example. Suppose two tweets, *t1* and *t2*, where:

- $t1_a$ = {*politics*, *economy*}, represents the actual topics for *t1*.
- $t1_p$ = {*sports*, *economy*} indicates the predicted topics for *t1*.
- $t2_a$ = {*sports*, *films*, *entertainment*, *football*, *economy*} refers to the actual topics for *t2*.
- $t2_p$ = {*politics*, *films*, *entertainment*, *football*, *economy*} represents the predicted topics for *t2*.

Hamming loss is a loss function, thus its optimal value is zero. It measures the number of wrong labels with respect to the total number of labels, but does not appropriately reflect the percentage of the correctly predicted labels. Calculating the hamming loss for *t1* and *t2*, we obtain:

$$HL_{t1}\ \frac{|t1_a\ \Delta\ t1_p|}{|L|} = \frac{|\{sports,\ politics\}|}{|L|} = \frac{2}{|L|} \tag{4}$$

$$HL_{t2}\ \frac{|t2_a\ \Delta\ t2_p|}{|L|} = \frac{|\{sports,\ politics\}|}{|L|} = \frac{2}{|L|} \tag{5}$$

and so $HL_{t1} = HL_{t2}$, even though *t2* has a larger percentage of successful predicted topics.

Label-based accuracy is a measure able to harmonize the number of not assigned topics with respect to the wrongly selected ones. Taking again *t1* and *t2* as example, the LBA for each one would be:

$$LBA_{t1}\ \frac{|t1_a \cap t1_p|}{|t1_a \cup t1_p|} = \frac{|\{economy\}|}{|\{politics,sports,economy\}|} = \frac{1}{3} \tag{6}$$

$$LBA_{t2}\ \frac{|t2_a \cap t2_p|}{|t2_a \cup t2_p|} = \frac{|\{films,entertaiment,football,economy\}|}{|\{sports,politics,films,entertaiment,football,economy\}|} = \frac{2}{3} \tag{7}$$

concluding that the LBA for *t2* is better than for *t1*.

Finally, a special case of LBA is the *Exact match* metric, which is a more restrictive metric due to the fact that it only considers a multi-label classification as successful when $Y_i = Z_i$, that is $Y_i \cap Z_i = Y_i \cup Z_i$. If we calculate the exact match for *t1* and *t2* we would obtain in both cases a value of 0. Note that, if we take an example *i* where $Y_i = Z_i$, then the EM, LBA and HL would be 1, 1 and 0, respectively.

Additionally, we will also consider two more metrics, used by TASS organisers, calculated according to equations 8 and 10: *At least one* scores a classification as valid whenever at least one topic is right, whereas *Match all* considers a multi-label classification valid when a superset of the actual topic set has been predicted.

$$At\ least\ one = \frac{1}{|D|}\sum_{i=1}^{|D|} f(i) \tag{8}$$

$$where\ f(i) = \begin{cases} 1\ if\ Y_i \cap Z_i \neq \emptyset \\ 0\ if\ Y_i \cup Z_i \neq \emptyset \end{cases} \tag{9}$$

$$Match\ all = \frac{1}{|D|}\sum_{i=1}^{|D|} g(i) \tag{10}$$

$$where\ g(i) = \begin{cases} 1\ if\ Y_i \subseteq Z_i \\ 0\ if\ Y_i \nsubseteq Z_i \end{cases} \tag{11}$$

The main drawback of these two measures is that it is possible to obtain a 'perfect result' by assigning all possible categories to each tweet. Therefore, they are less robust with respect to academic misconduct.

5.3. Supervised classifier

To be able to assign several topics to the same tweet, we carried out a one vs. all strategy: given n topics, this perspective proposes to train n classifiers where each one makes it possible to differentiate a topic i, where $i \in n$, from the others. Thus, if we plan to discover if a text is talking about a topic X, Y or Z, or several of them at the same time, we would create three classifiers: topic X vs. Other, topic Y vs. Other and topic Z vs. Other.

We use the WEKA data mining software [53] to build machine learning models. As a classifier, we have chosen to work with SMO [29], an implementation of Support Vector Machines.

5.4. Feature Filtering

The number of features derived from n-grams, psychometric properties and dependency triplets results into a high multi-dimensional feature space. To address this issue, we test the use of an information gain filter, which measures the relevance of an attribute with respect to the class. Only features with an information gain greater than zero are selected to then feed a supervised classifier. Given the huge number of attributes and the low value of most of them, we cannot ensure the optimal selection of attributes, but experimental results suggest that applying an information gain filter improves performance with respect to omitting it. We use the WEKA attribute selection tools to apply the information gain filtering.

5.5. Feature models

As a starting point, we consider several basic feature models based on lexical, syntactic, psychometric and semantic knowledge. We will measure their performance when they are used independently, and then we will combine them in order to improve either the Hamming loss, the label-based accuracy or the exact match:

- *Unigrams of words*: Representing the text as a bag-of-words is a widely-used approach when performing text analytics. We are taking this model as the baseline. To apply this model, we only need to split sentences into words and then use each word as a feature for the supervised classifier.
- *Unigrams of lemmas*: Languages such as Spanish present a rich morphology which manifests itself in inflections for gender and number in nouns, adjectives and verbs. We think lemmatisation may help reduce sparsity and improve performance.
- *Bigrams of words*: In addition to unigrams, we also take into account bigrams of words. Bigrams make it possible to capture more context in a simple way. We hypothesise that this can be helpful to better discriminate non-related topics, since words are considered in context.
- *Bigrams of lemmas*: In a similar line, we also consider bigrams of lemmas as a possible optimisation of word bi-grams.
- *Psychometric properties*: We evaluate this type of features in an isolated way in order to measure and estimate the recall and scope of the LIWC psychometric dictionaries.
- *Part-of-speech tags*: Considering the employment of grammatical categories is probably not a crucial issue for the majority of topics, as we explained previously. But this might not be true in all cases. Certain PoS-tags, such as proper nouns, are likely to appear more often when someone is talking about music, books or films. People often refer to actor, director or producer names to implicitly indicate their opinion about a film. In a similar line, third person pronouns could be more frequent in not so personal issues such as politics or technology.

Finally, we will combine the best-performing features with Syntactic features: we represent syntactic information by means of generalised dependency-based features. We test different levels of generalisation over the head and the dependent word of a dependency triplet, including lemmas, psychometric properties and fine-grained PoS-tags.

6. Experimental results

Table 3 shows the performance of the initial sets of features proposed above. The information is ordered following the exact match metric, in descending order. Bigrams of lemmas obtained the best exact match, whereas the baseline (composed of only words) obtained the best Hamming loss and label-based accuracy values. For unigrams of words and lemmas we also included results without applying the information gain filter, in order to show the need for this step.

Table 3. Performance for basic feature models. The IG column indicates whether information gain (IG) was used to filter out features. The best result for each metric is shown in boldface.

Model	IG	HL	LBA	EM
Bi-grams of lemmas (BL)	Yes	0.077	0.626	**0.530**
Words (baseline) (W)	Yes	**0.073**	**0.658**	0.527
Bi-grams of words (BW)	Yes	0.080	0.613	0.524
Words (W)	No	0.079	0.634	0.498
Lemmas (L)	Yes	0.078	0.640	0.493
Lemmas (L)	No	0.085	0.611	0.460
Fine PoS-tags (FT)	Yes	0.289	0.262	0.032
Psychometric (P)	Yes	0.301	0.250	0.026

Table 4 illustrates how, by combining various sets of the features proposed above, we can obtain an even better classifier.

Table 4. Performance when combining lexical, psychometric and semantic knowledge: bi-grams of lemmas (BL), bi-grams of words (BW), psychometric properties (P), words (W), lemmas (L), fine-grained PoS-tags (FT), dependency types (DT). The best result for each metric is shown in boldface.

Model	HL	LBA	EM
W+BL	**0.068**	**0.671**	**0.573**
BL+P	0.076	0.632	0.539
BL	0.077	0.626	0.530
W+BW+P	0.078	0.647	0.530
W+BW	0.074	0.646	0.529
W+P+FT+DT	0.073	0.655	0.527
W+P+FT	0.073	0.656	0.528
W	0.073	0.658	0.527
W+P	0.073	0.655	0.526
W+L	0.073	0.656	0.525
BL+P+FT	0.082	0.612	0.495

Finally, we also included generalised dependency triplet features in order to find out if syntactic information helps us build more accurate topic classification models. Table 5 illustrates how syntactic knowledge can markedly improve the results of the standard multi-classification metrics over the bag-of-words model. Table 6 breaks down the results by categories, comparing the best contextual model with respect to the bag-of-words approach.

Table 5. Performance when improving the bag-of-words model by means of generalised dependency triplets: words (W), lemmas (L), coarse-grained PoS-tags (CT), psychometric properties (P), dependency types (DT). The best result for each metric is shown in boldface.

Model	HL	LBA	EM
W	0.073	0.658	0.527
W+(_,DT, L)	0.071	0.660	0.542
W+(L,DT,P)	0.071	0.661	0.551
W+(L,DT,L)	**0.067**	**0.674**	**0.579**

Table 6. Detailed performance per category both for the best syntactic model and the bag-of-words approach. Metrics are calculated as follows: P=TP/(TP+FP), R=TP/(TP+FN) and F-measure=2 RP/(R+P), where TP denotes the true positives, FP the false positives, and FN the false negatives. If a tweet discusses films and entertainment, but it is only classified in the films class, it would be taken as a true positive for the films category, and as a false negative for the entertainment class.

	F-measure		P (precision)		R (recall)	
Model	Best model	Words	Best model	Words	Best model	Words
films	0.359	0.306	0.317	0.216	0.414	0.523
politics	0.717	0.733	0.753	0.754	0.685	0.714
technology	0.343	0.344	0.429	0.252	0.286	0.540
entertainment	0.434	0.442	0.458	0.335	0.412	0.650
sports	0.271	0.271	0.349	0.224	0.222	0.341
other	0.678	0.689	0.578	0.611	0.820	0.790
economy	0.435	0.391	0.372	0.267	0.524	0.729
music	0.451	0.436	0.361	0.436	0.559	0.710
football	0.292	0.332	0.503	0.301	0.205	0.371
literature	0.380	0.348	0.395	0.255	0.366	0.548
Macro average	0.436	0.429	0.452	0.353	0.449	0.590

7. Discussion of results

Table 3 shows the results for the basic sets of proposed features. The use of n-grams, both of words and lemmas, clearly outperforms features based on part-of-speech and psychological knowledge, which are not helpful by themselves. It is important to remark that unigrams of words improve the metrics over unigrams of lemmas. However, this trend is not present when using bigrams, where lemmas perform better. We hypothesise that this is due to sparsity problems: words are sparser than lemmas. While this may not be a major problem when training with n-grams where n=1, it becomes an issue when using larger values of n: combinations highly increase the dimensional space of features, and probably a larger training set would be needed to even out the performance between bigrams of words and lemmas.

The results in Table 4 show that psychometric properties, although not being helpful by themselves when used in isolation, allow us to improve the exact match of other models based on n-grams. In this way, the LIWC psychological dictionaries seem to be able to provide additional information that a model based on terms cannot represent. In addition, this table outlines the performance that can be achieved by a multi-topic model which relates lexical information via contextual information, by using n-grams.

Table 5 shows how the use of syntactic information can be useful to obtain better contextual models taking as starting point the bag-of-words model. We obtain the best results using the non-generalised dependency triplets, but several generalised triplets achieved improvements over the bag-of-words approach. It is remarkable that the model composed by words and generalised triplets where the head term is removed, that is a model based on words and lemmas labelled with their dependency type, improves its corresponding lexical counterpart (composed only by words and lemmas). This reinforces the utility of considering the roles that words play in a sentence, since terms marked with syntactic functions such as attribute of direct object may have more impact on identifying the core parts of a sentence, and consequently their topics. After including syntactic information, we finally obtain a model that achieves the best performance in the three standard metrics for multi-label classification tasks. The model is presented in Table 5 and uses unigrams of words and (lemma, dependency, lemma) features to feed an SMO classifier. The result is a slight improvement over the best model (based on words and lemma bigrams) in Table 4. Therefore, while the bulk of the improvements over the baseline come from considering contextual information encoded in pairs of lemmas (be it as bigrams or as dependency triplets) in addition to word forms; the results suggest that using syntactic information can provide an extra edge, pushing performance slightly beyond that of bigram models. We hypothesise that this is because, while most syntactic dependencies are short-distance and therefore lemma bigrams largely overlap with lemma dependency triplets; the latter can capture structural relations between terms involved in longer-distance dependencies, providing some extra useful information to the classifier that is not present in bigrams. However, it is also worth taking into account that extracting syntactic relations from tweets is much more demanding than merely lemmatising and obtaining bigrams, both in terms of computational and linguistic resources needed. Thus, the gains obtained by using syntactic information can often not be worthwhile, depending on the resources at hand. In many cases, it will be sensible to stop the NLP pipeline at

lemmatisation, which is notably beneficial in the light of the results, and highlights the importance of adapting NLP techniques to the particular characteristics of language usage in Twitter.

Table 7 compares the performance of our best model with respect to the methods proposed at TASS 2013. Our syntactic model obtains the best value for the three standard metrics for multi-label classification. It also obtains the second best performance for the "at least one" and "match all" metrics. This reinforces the practical utility of our linguistic angle over the different models proposed at the workshop. It is important to remark that taking only into account the models that participated in TASS 2013, our initial system obtained the best value for the "match all" and "at least one" metrics, but the same is not true for label-based accuracy, exact match and Hamming loss, where FHC25-IMDEA outperformed our results. However, FHC25-IMDEA did not tackle the problem as multi-label, causing the exact match and the "match all" metrics to be equal. This strategy has benefits, given the topic frequency distribution of tweets (shown in Appendix A and Appendix B). In fact, if we evaluate our initial system, replacing the multi-label perspective with a single-label one, we would obtain an exact match of 0.589, a label-based accuracy of 0.656 and a Hamming loss of 0.07, outperforming FHC25-IMDEA and showing that our approach performs better under the same conditions. In any case, we do not consider it as valid, given the true nature of the topic classification problem.

Table 7. Performance on combining lexical, syntactic, psychometric and semantic knowledge: bi-grams of lemmas (BL), bi-grams of words (BW), psychometric properties (P), words (W), lemmas (L), fine-grained PoS-tags (FT), dependency types (DT). A detailed description of some of these systems can be found in section 2.2.

Model	HL	LBA	EM	Match all	Al least one
Best contextual model	**0.068**	**0.674**	**0.579**	0.663	0.771
FHC25-IMDEA (Cordobés et al.) [40]	0.072	0.637	**0.573**	0.573	0.702
Our approach at TASS 2013 [43]	0.086	0.614	0.456	**0.690**	**0.786**
UPV (Plan and Hurtado) [28]	0.084	0.608	0.468	0.659	0.756
UNED-JRM (Rufo Mendo) [39]	0.124	0.417	0.358	0.382	0.479
ETH-ZURICH [44]	0.098	0.370	0.291	0.385	0.455
LSI UNED (Martín-Wanton and Carrillo de Albornoz) [31]	0.185	0.197	0.070	0.364	0.406
SINAI-CESA (Montejo-Ráez et al.) [34]	0.182	0.126	0.093	0.093	0.159

Finally, Table 6 presented the results per category, both for the best syntactic model and the pure bag-of-words approach. Precision, recall and F-measure metrics are not intended for multi-label classification, but they provide useful information about how the system behaves in each category. Precision is higher for the syntactic model, but the same is not true for recall. This suggests that syntactic information is able to better discriminate unrelated topics while the bag-of-words approach assigns too many topics to each tweet, achieving a high recall but a lower precision; which is probably not what most users are looking for. This is also coherent with the results obtained with the standard metrics for multi-topic classification, explained above these lines. Our best scores were obtained on categories such as *politics*, *other, music* or *economy*, although we do not think this provides conclusive evidence that these classes are easier to classify. In this respect, there exist studies which discuss the differences on classifying different topics. Turney [54] poses this issue when dealing with polarity classification. He argued that text classification tasks become harder the more abstract the topic is. In this respect, in [48] it is shown how classifying the polarity on topics where criteria depend strongly on the person, such as movies or music, is harder than on categories where the quality criteria are more standard (e.g., hotels). A similar issue is also pointed out by Scharkow [55], where machine learning is used to identify the theme of German news. He concludes that some news, such as those about crime, are harder to classify by machine learning techniques, since they often rely on real-life knowledge. He also argued that supervised automatic coding is about 15% less reliable than human annotation, although exceptions are common. Computers have difficulties with categories that rely on real-life knowledge, although humans may present the same weakness when they are not familiar with the subject.

8. Conclusions and future work

We have presented a supervised topic classification system for Spanish tweets based on a linguistic perspective, employing morphological, syntactic and psychometric information. We address the problem as a multi-label classification task, since a single text can relate and refer to several different topics. We propose an approach which does not take into account metadata of any kind, but only considers the information provided by the text itself, and processes it with various NLP techniques. The approach has been applied on Twitter, given the present success of this medium, but it would be easily adaptable to other social networks, blogs or forums. The practical utility of this approach has been tested at the experimental evaluation workshop on sentiment analysis TASS 2013, where an initial model following this same angle was the best performing system in the topic classification task.

Our experimental results provide an exhaustive evaluation through several sets of features, showing how lexical, syntactic, psychological and semantic attributes allow us to improve different aspects that a topic classification system should take into account. The results lead us to conclude that relating NLP-extracted features by means of contextual information adds complementary knowledge over pure lexical models, making it possible to outperform them on standard metrics for multi-label classification tasks. On the other hand, it is important to apply feature selection techniques, as we observed that filtering features based on their information was a relevant factor for the good performance obtained in the experiments.

As future work, we consider that the techniques defined in this article could be used successfully in the filtering phase of a topic-related tweet retrieval system [57]. We also plan to explore how different kinds of metadata could be included into our linguistic proposal. The use of metadata is beyond the scope and aim of this article. However, this kind of knowledge has been applied with varying degrees of effectiveness in different tasks. In [43], we showed how by using user metadata, we were able to slightly outperform our original run at the TASS 2013 topic classification task. That information was not extracted through the Twitter API, but provided by the organisation, so it would be not effective in a real environment, and we have thus decided to avoid it. Nevertheless, the information provided by the Twitter API could be useful to make interesting findings about the authors of tweets. In a similar way, obtaining temporal and geographic information associated with tweets may be helpful on associating those messages with a specific event. The use of hyperlinks, which has been already applied by other authors such as [25], could also help to enrich our approach. On the other hand, we must take into account that user metadata provided by the Twitter API was not helpful on tasks related to Twitter classification, such as identifying influential authors in the CLEF Replab 2014 evaluation, where a pure linguistic approach attained the best performance [58]. Finally, integrating real-world knowledge in a machine learning system could be interesting to give a higher weighting factor to some topics, depending on the moment (e.g., it is more likely that a tweet containing the word 'Liverpool' is referring to the 'Soccer' category if it is published at the weekend during the soccer season).

Notes

1. These tweets are extracted from the corpus TASS 2013, which we are using to evaluate our approach.
2. The tweet is a translation of the Spanish tweet '*La clave del nuevo gobierno: su estructura ¿Habrá dos vicepresidencias o ninguna? La clave, en el equipo económico*'.
3. The corresponding Spanish translation would be '*El público inteligente está en las redes sociales. La educación determina su uso más que la riqueza. Impacto en medios*'.
4. This is a translation of the Spanish tweet '*Holaaa mis tweeps! Un dia maravilloso para hacer un twitcam como os he prometido sera a las 6:00 pm (hora de Mexico) nos vemos pronto!!!*'.
5. http://www.wikipedia.org/
6. https://myspace.com/
7. This corpus will be used in the experimental section of this article, therefore a detailed description can be found in Sect. 5 and a discussion of results in Sects. 6 and 7.
8. http://adwords.google.com/o/KeywordTool now replaced by Google Keyword Planner for registered users
9. We realise they could be useful on topic classification tasks focused on a very short period of time, but not applicable in a general way, since hashtags often refer very specific events or issues.
10. http://nltk.org/
11. We considered other PoS-taggers included in the NLTK, but they attained a similar performance on experimental evaluation.
12. Labelled Attachment Score: Percentage of words that have their head and their dependency type correctly assigned.
13. Unlabelled Attachment Score: Percentage of words that have their head correctly assigned, without considering the dependency type.
14. This is a translation of the Spanish tweet '*La quinta ensaladera, Nadal saludando al equipo argentino. Y ahora al público. Oe, oe, oe*'.

15. A contiguous sequence of *n* items from a given sequence of text.

Funding

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

[1] Twitter, Inc., 'Form S-1 Registration Statement under The Securities Act of 1933', Washington, D.C.: United States Securities and Exchange Commission, http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm (2013, accessed 20 May 2014).

[2] Jansen BJ, Sobel K and Cook G. Classifying ecommerce information sharing behaviour by youths on social networking sites, Journal of Information Science 2011; 37(2):120–136.

[3] Bontcheva K and Rout D. Making sense of social media streams through semantics: a survey, Semantic web (In press, accessed 20 May 2014).

[4] Junqué de Fortuny E, De Smedt T, Martens D, Daelemans W. Evaluating and understanding text-based stock price prediction models, Information Processing and Management 2014; 50(2):426–441.

[5] Rui H, Liu Y, Whinston A. Whose and what chatter matters? the effect of tweets on movie sales, Decision Support Systems 2013; 55(4):863–870.

[6] Villena-Román J and García-Morera J. TASS 2013 — workshop on sentiment analysis at SEPLN 2013: An overview. In: Díaz Esteban et al. [57], pp. 112–125.

[7] Büttcher S, Clarke CLA and Cormack, GV. Information Retrieval: Implementing and Evaluating Search Engines. Cambridge, Massachusetts: The MIT Press, 2010.

[8] Sebastiani F. Machine learning in automated text categorization, ACM Computing Surveys 2002; 34(1):1–47.

[9] Yerva SR, Miklós Z, Aberer K. Entity-based classification of Twitter messages, International Journal of Computer Science and Applications 2012; 9(2):88–115.

[10] Dan O, Feng J and Davison BD. A bootstraping approach to identifying relevant tweets for social tv. In: L. A. Adamic, R. A. Baeza-Yates, S. Counts (eds.), Proceedings of the Fifth International Conference on Weblogs and Social Media. Barcelona, Spain: AAAI, 2011.

[11] Martínez-Cámara E, Martín-Valdivia MT, Ureña López LA and Montejo-Ráez A. Sentiment analysis in Twitter, Natural Language Engineering 2014; 20(1):1–28.

[12] Martínez-Cámara E, Martín-Valdivia MT, Molina-González MD and Perea-Ortega JM. Integrating Spanish lexical resources by meta- classifiers for polarity classification, Journal of Information Science (In press, accessed 20 May 2014).

[13] Vilares D, Alonso MA and Gómez-Rodríguez C. Supervised polarity classification of Spanish tweets based on linguistic knowledge. In: DocEng'13. Proceedings of the 13th ACM Symposium on Document Engineering. Florence, Italy: ACM, 2013, pp. 169–172.

[14] Vilares D, Alonso MA and Gómez-Rodríguez C. On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages, Journal of the Association for Information Science and Technology (In press, accessed 20 May 2014).

[15] Sriram B, Fuhry D, Demir E, Ferhatosmanoglu H and Demirbas M. Short text classification in Twitter to improve information filtering. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010). Geneva, Switzerland: ACM, 2010, pp. 841–842.

[16] Thongsuk C, Haruechaiyasak C and Saelee S. Multi-classification of business types on Twitter based on topic model. In: The 8th Electrical Engineering/Electronics, Computer, Telecommunications and Information Technologies (ECTI) Association of Thailand — Conference 2011, Khon Kaen, Thailand, 2011, pp. 508–511.

[17] Blei DM, Ng AY and Jordan MI. Latent Dirichlet allocation, The Journal of Machine Learning Research 2003; 3:993–1022.

[18] Chang C and Lin C. LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems Technology 2011; 2(3):Article 27.

[19] Hong J and Davison BD. Empirical study of topic modelling in Twitter. In: Proceedings of the First Workshop on Social Media Analytics (SOMA 2010). Washington, DC: ACM, 2010, pp. 80–88.

[20] Fiaidhi J, Mohammed S, Islam A, Fong S and Kim T. Developing a hierarchical multi-label classifier for Twitter trending topics, International Journal of u- and e- Service, Science and Technology 2013; 6(3):1–12.

[21] Bastos MT, Travitzki R and Puschmann C. What sticks with whom? Twitter follower-folowee networks and news classification. In: AAAI Technical Report WS-12-01 Workshop on the Potential of Social Media Tools and Data for Journalists. Dublin, Ireland:AAAI, 2012, pp. 6–13.

[22] Lee K, Palsetia D, Narayanan R, Patwary MMA, Agrawal A and Choudhary A. Twitter trending topic classification. In: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW 2011). Vancouver, Canada: IEEE, 2011, pp. 251–258.

[23] Genc Y, Sakamoto Y and Nickerson JV. Discovering context: Classifying tweets through a semantic transform based on Wikipedia. In: D. D. Schmorrow, C. M. Fidopiastis (eds.), Foundations of Augmented Cognition. Directing the Future of

Adaptive Systems, Vol. 6780 of Lecture Notes in Artificial Intelligence. Heidelberg, Dordrecht, London, New York: Springer, 2011, pp. 484–492.

[24] Gattani A, Lamba DS, Garera N, Tiwari M, Chai X, Das S et al. Entity extraction, linking, classification, and tagging for social media: A Wikipedia-based approach, Proceedings of the VLSB Endowment 2013; 6(11):1126–1137.

[25] Kinsella S, Passant A and Breslin JG. Topic classification in social media using metadata from hyperlinked objects. In: P. Clough, C. Foley, C. Gurrin, GJF. Jones, W. Kraaij, H. Lee, V. Mudoch (eds.), Advances in Information Retrieval. 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings, Vol. 6611 of Lecture Notes in Computer Science. Heidelberg, Dordrecht, London, New York: Springer, 2011, pp. 201–206.

[26] Villena-Román J, Lana-Serrano S, Martínez-Cámara E and González-Cristóbal JC. TASS — workshop on sentiment analysis at SEPLN, Procesamiento del Lenguaje Natural 2013; 50:37–44.

[27] Batista F and Ribeiro R. The L2F strategy for sentiment analysis and topic classification. In: TASS 2012 Working Notes, Castellón de la Plana, Spain, 2012.

[28] Pla F and Hurtado LF. ELiRF-UPV en TASS-2013: Análisis de sentimientos en Twitter. In: Díaz Esteban et al. [57], pp. 220–227.

[29] Platt JC. Advances in kernel methods. Cambridge, MA: MIT Press, 1999, Ch. Fast training of support vector machines using sequential minimal optimization, pp. 185–208.

[30] Martínez-Cámara E, García Cumbreras MA, Martín-Valdivia MT and Ureña López LA. SINAI en TASS 2012, Procesamiento del Lenguaje Natural 2013; 50:53–60.

[31] Martín-Wanton T and Carrillo de Albornoz J. UNED at TASS 2012: Polarity classification and trending topic system. In: TASS 2012 Working Notes, Castellón de la Plana, Spain, 2012.

[32] Castellano González A, Cigarrán Recuero J and García Serrano A. UNED @ TASS: Using IR techniques for topic-based sentiment analysis through divergence models. In: TASS 2012 Working Notes, Castellón de la Plana, Spain, 2012.

[33] Castellano González A, Cigarrán Recuero J and García Serrano A. UNED LSI @ TASS 2013: Considerations about textual representation for IR based tweet classification. In: Díaz Esteban et al. [57], pp. 213–219.

[34] Montejo-Ráez A, Díaz Galiano MC and García-Vega M. LSA based approach to TASS 2013. In: Díaz Esteban et al. [57], pp. 195–199.

[35] Manning CD, Raghavan P and Schütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008.

[36] Fernández Anta A, Núñez Chiroque L, Morere P and Santos A. Sentiment analysis and topic detection of Spanish tweets: A comparative study of NLP techniques. Procesamiento del Lenguaje Natural 2013; 50:45–52.

[37] Trilla A and Alías F. Sentiment analysis of twitter messages based on multinomial naive Bayes. In: TASS 2012 Working Notes, Castellón de la Plana, Spain, 2012.

[38] Rennie JD, Shih L, Teevan J and Karger D. Tackling the poor assumptions of naive Bayes text classifiers. In: T. Fawcett, N. Mishra (Eds.), Proceedings of the Twentieth International Conference on Machine Learning. Washington, DC: AAAI, 2003, pp. 616–623.

[39] Rufo Mendo FJ. Are really different topic classification and sentiment analysis? In: Díaz Esteban et al. [57], pp. 206–212.

[40] Cordobés H, Anta AF, Núñez LF, Pérez F, Redondo T and Santos A. Técnicas basadas en grafos para la categorización de tweets por tema. In: Díaz Esteban et al. [57], pp. 160–166.

[41] Brin S and Page L. The anatomy of a large-scale hypertextual web search engine. In: Proc. of the Seventh International Conference on World Wide Web, Brisbane, Australia, 1998, pp. 107–117.

[42] Kleinberg JM. Authoritative sources in a hyperlinked environment, Journal of the ACM 1999; 46(5):604–632.

[43] Vilares D, Alonso MA and Gómez-Rodríguez C. LyS at TASS 2013: Analysing Spanish tweets by means of dependency parsing, semantic- oriented lexicons and psychometric word-properties. In: Díaz Esteban et al. [57], pp. 179–186.

[44] García D and Thelwall M. Political aligment and emotional expressions in Spanish tweets. In: Díaz Esteban et al. [57], pp. 151–159.

[45] Brill E. A simple rule-based part of speech tagger. In: Proceedings of the workshop on Speech and Natural Language, HLT'91, Stroudsburg, PA; ACL 1992, pp. 112–116.

[46] Taulé M, Martí MA and Recasens M. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis et al. (eds.), Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 2008.

[47] Gimpel K, Schneider N, O'Connor B, Das D, Mills D, Eisenstein J et al. Part-of-speech tagging for Twitter: annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2, HLT '11, Stroudsburg, PA; ACL, 2011, pp. 42–47.

[48] Vilares D, Alonso MA and Gómez-Rodríguez C. A syntactic approach for opinion mining on Spanish reviews, Natural Language Engineering (In press, accessed 20 May 2014).

[49] Nivre J, Hall J, Nilsson J, Chanev A, Eryigit G, Kübler S et al. Maltparser: A language-independent system for data-driven dependency parsing, Natural Language Engineering 2007; 13(2):95–135.

[50] Ramírez-Esparza N, Pennebaker JW, García FA and Suri Martínez R, La psicología del uso de las palabras: Un programa de computadora que analiza textos en español, Revista Mexicana de Psicología 2007; 24(1):85–99.

[51] Pennebaker J, Francis M and Booth R. Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 2001.
[52] Joshi M and Penstein-Rosé C. Generalizing dependency features for opinion mining. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, ACLShort '09. Suntec, Singapore: ACL, 2009, pp. 313–316.
[53] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P and Witten IH. The Weka data mining software: an update, SIGKDD Explorations 2009; 11(1):10–18.
[54] Turney, PD. 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. Philadelphia: ACL, 2002, pp. 417-424.
[55] Scharkow, M. Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. Quality & Quantity 2013; 47(2), 761-773.
[56] Cotelo JM, Cruz FL and Troyano JA. Dynamic topic-related tweet retrieval, Journal of the Association for Information Science and Technology 2014; 65(3):513–523.
[57] Díaz Esteban A, Alegría Loinaz I and Villena Román J. (eds.), XXIX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2013). TASS 2013 - Workshop on Sentiment Analysis at SEPLN 2013, Madrid, Spain: SEPLN, 2013.
[58] Vilares D, Hermo M, Alonso MA, Gómez-Rodríguez C and Vilares J. LyS at CLEF RepLab 2014: Creating the State of the Art in Author Influence Ranking and Reputation Classification on Twitter. In Linda Cappellato, Nicola Ferro, Martin Halvey, Wessel Kraaij (eds.), CLEF 2014. CLEF2014 Working Notes. Working Notes for CLEF 2014 Conference. Sheffield, UK, September 15-18, 2014, CEUR Workshop Proceedings, Vol. 1180, pp. 1468-1478.

Appendices

Appendix A. Training set statistics

Table 9. Training set statistics: percentage of tweets and number of tweets per category.

Categories	Training set		Test set	
	%tweets	#tweets	%tweets	#tweets
{films}	1.5	107	0.3	203
{films, economy}	0.0	1	-	-
{films, entertainment}	0.3	21	0.0	13
{films, entertainment, music}	0.0	1	-	-
{films, entertainment, other}	0.0	2	0.0	1
{films, entertainment, politics}	0.0	3	-	-
{films, soccer}	0.0	1	-	-
{films, music}	0.1	7	0.0	5
{films, other}	1.3	97	0.6	368
{films, other, politics}	0.0	1	-	-
{films, politics}	-	-	0.0	5
{films, technology}	0.1	4	0.0	1
{sports}	1.0	75	0.2	106
{sports, economy}	0.0	2	-	-
{sports, entertainment}	0.2	11	0.0	3
{sports, entertainment, music}	0.0	1	-	-
{sports, entertainment, other}	0.0	1	-	-
{sports, entertainment, other, politics}	0.0	1	-	-
{sports, entertainment, politics}	0.0	1	-	-
{sports, soccer}	0.1	5	0.0	1
{sports, literature}	0.0	1	-	-
{sports, music}	0.1	4	0.0	1
{sports, music, other}	0.0	1	-	-
{sports, other}	0.1	8	0.0	20
{sports, politics}	0.0	1	0.0	4
{sports, technology}	0.0	1	-	-
{economy}	3.7	267	2.0	1 209
{economy, entertainment}	0.4	32	0.0	4
{economy, entertainment, other}	0.1	5	-	-
{economy, entertainment, other, politics}	0.0	2	-	-
{economy, entertainment, politics}	0.5	36	0.0	1

{economy, entertainment, politics, technology}	0.0	1	-	-
{economy, entertainment, technology}	0.0	2	-	-
{economy, soccer}	0.0	2	0.0	1
{economy, literature}	0.0	1	-	-
{economy, literature, politics}	0.0	1	-	-
{economy, literature, politics, technology}	0.0	1	-	-
{economy, music}	0.0	1	-	-
{economy, music, politics}	0.0	1	-	-
{economy, other}	0.3	23	0.3	195
{economy, other, politics}	0.4	28	0.0	1
{economy, other, technology}	0.0	1	-	-
{economy, politics}	7.3	529	1.9	1 138
{economy, politics, technology}	0.0	1	-	-
{economy, technology}	0.1	5	-	-
{entertainment}	11.5	827	5.7	3 494
{entertainment, soccer}	0.4	30	0.0	6
{entertainment, soccer, music, other}	0.0	1	-	-
{entertainment, soccer, other}	0.0	2	-	-
{entertainment, literature}	0.3	19	0.0	9
{entertainment, literature, technology}	0.0	2	-	-
{entertainment, music}	0.5	39	0.0	6
{entertainment, music, other}	0.1	5	0.0	1
{entertainment, music, politics}	0.0	1	-	-
{entertainment, music, technology}	0.0	1	-	-
{entertainment, other}	4.5	328	2.4	1 486
{entertainment, other, politics}	0.2	13	0.0	3
{entertainment, other, technology}	0.1	4	0.0	3
{entertainment, politics}	3.3	241	-	-
{entertainment, politics, technology}	0.1	4	-	-
{entertainment, technology}	0.6	40	0.0	20
{soccer}	2.3	166	1.2	700
{soccer, literature}	0.0	1	-	-
{soccer, music}	0.1	8	0.0	2
{soccer, music, other}	0.0	1	-	-
{soccer, other}	0.4	27	0.2	95
{soccer, politics}	0.1	7	0.0	17
{soccer, technology}	0.0	1	0.0	1
{literature}	0.6	45	0.1	76
{literature, music}	0.0	2	-	-
{literature, other}	0.2	14	0.0	7
{literature, politics}	0.2	13	0.0	1
{literature, technology}	0.0	2	-	-
{music}	2.8	200	0.9	545
{music, other}	3.9	279	1.5	924
{music, other, politics}	0.0	1	-	-
{music, other, technology}	0.0	1	-	-
{music, politics}	0.1	5	0.0	13
{music, technology}	0.1	6	0.0	1
{other}	17.3	1 248	34.5	20 979
{other, politics}	3.0	215	6.7	4 081
{other, politics, technology}	0.0	1	-	-
{other, technology}	0.4	27	0.0	27
{politics}	27.5	1 982	40.2	24 416
{politics, technology}	0.4	29	0.0	16
{technology}	1.1	83	0.4	218