Lexicon Expansion System for Domain and Time Oriented Sentiment Analysis

Nuno Guimaraes[1], Luis Torgo[2] and Alvaro Figueira[1]
[1]*CRACS - INESC TEC, Rua do Campo Alegre 1021/1055, Porto, Portugal*
[2]*LIADD - INESC TEC, Rua do Campo Alegre, 1021/1055, Porto, Portugal*

Keywords: Sentiment Lexicon, Lexicon Expansion, Twitter Sentiment Analysis.

Abstract: In sentiment analysis the polarity of a text is often assessed recurring to sentiment lexicons, which usually consist of verbs and adjectives with an associated positive or negative value. However, in short informal texts like tweets or web comments, the absence of such words does not necessarily indicates that the text lacks opinion. Tweets like "First Paris, now Brussels... What can we do?" imply opinion in spite of not using words present in sentiment lexicons, but rather due to the general sentiment or public opinion associated with terms in a specific time and domain. In order to complement general sentiment dictionaries with those domain and time specific terms, we propose a novel system for lexicon expansion that automatically extracts the more relevant and up to date terms on several different domains and then assesses their sentiment through Twitter. Experimental results on our system show an 82% accuracy on extracting domain and time specific terms and 80% on correct polarity assessment. The achieved results provide evidence that our lexicon expansion system can extract and determined the sentiment of terms for domain and time specific corpora in a fully automatic form.

1 INTRODUCTION

With the massive growth of Social Web, opinion data became much more accessible and in larger quantities. The use of social networks like Twitter or Facebook and the way users share their feelings regarding politicians, products, events, companies, and celebrities, through their personal profile has motivated the interest for further investigation on methods to automatically classify the associated sentiment.

Supervised and unsupervised approaches have been proposed in sentiment analysis classification and the inclusion of a sentiment lexicon is a common approach on both. These lexicons are mainly built using verbs and adjectives since they are the more common indicators of subjectivity. Although this may work relatively well in medium and large texts (e.g. reviews or articles), in small texts like tweets or comments, the task becomes more difficult due to their short length format (Kiritchenko et al., 2014). Small texts may not include any of the words in the sentiment lexicons and still express a sentiment. Tweets like *"Listening to Bowie. Still can't believe it"* do not include any opinion words but have a sentiment associated which is perceptible to who is aware of the death of the artist David Bowie. Furthermore, tweets with a sense of irony can also be misinterpreted by general sentiment lexicons. For example in the following tweet *"I used to think that Britain produced best comedy programs but where else but here could we watch a team like Sarah Palin and Donald Trump on TV?"* words like *"best"* could lead to a positive sentiment classification. However, the tweet is pointing to an overall negative sentiment "disguised" with irony. Our ability to detect the sentiment in both cases is due to: 1) the knowledge of events and persons which is achieved from news (seen on TV, newspapers and Internet) and 2) the knowledge on the public opinion and reactions to those news. However, this is a feature that current state of the art sentiment analysis methods do not consider when assessing the polarity of a text.

News have an important role in today's society. Up to date information of events in several different domains keep people aware of what is going on in the world. That awareness has grown with the rise of the World Wide Web since news have become much more accessible and in greater quantity. Furthermore, news are usually classified as relevant information and may transmit a change of opinion on certain entities or events. For example if an article is released on a ma-

Guimaraes, N., Torgo, L. and Figueira, A.
Lexicon Expansion System for Domain and Time Oriented Sentiment Analysis.
DOI: 10.5220/0006081704630471
In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) - Volume 1: KDIR*, pages 463-471
ISBN: 978-989-758-203-5

jor politician caught in a money laundering scheme, the public opinion on that person may change. Yet, there are also some cases where the opinion does not shift (e.g. advances in the cure for Alzheimer may not reverse the sentiment on the term "Alzheimer").

News headlines due to their short format, appear to be good sources for relevant terms extraction. However, the sentiment transmitted in them may not be the same as the sentiment from public opinion. To assess the public opinion on news, Twitter makes a good data source, since it includes millions of users from famous people to companies and presidents. The number of tweets and active users is also a factor. Since June 2015, on average, 500 million tweets are sent per day. The micro blogging site has also approximately 316 million users active per month (Twitter, 2015a). Moreover, Twitter provides a public API allowing the retrieval of tweets, getting user information and monitoring tweets in real time making it straightforward to retrieve large quantities of data for analysis (Twitter, 2015b). Summarising, Twitter is an updated and diversified source of information since millions of tweets are posted on a daily basis about different subjects from users with different opinions.

For this reason, we could use Twitter trending terms to build our sentiment lexicon. However, they do not always represent global relevance and are normally very specific to that social network. On the other hand, analysing directly headlines (or the full news article) may not provide a accurate sentiment on the terms it mentions.

Taking these facts into account, we developed a system for sentiment lexicon expansion that combines both. First, determines relevant and up to date terms from news headlines and then, for each term, it uses Twitter to determine the current public sentiment on it. Our main goals for this work are: 1) to assess the reliability in extracting domain and time specific terms for our lexicon expansion method using solely news headlines and 2) if the polarity assigned by the sample of tweets containing the terms corresponds to the polarity of the terms.

The rest of the article is organized as follows. First, we describe the state of the art on the subject. Next, we specify the workflow of our proposal. Then, we present the experimental evaluation of our system. Finally, we describe some conclusions and future work.

2 RELATED WORK

One of the most important parts for achieving high accuracy on sentiment analysis are "sentiment lexicons" (or sentiment dictionaries). Each of the words in these lexicons can have a binary (positive and negative), ternary (positive, neutral, negative) or numerical (e.g a -5 to 5 interval) sentiment value. Some studies also evaluate sentiment as emotions like fear, joy and sadness (Mohammad and Turney, 2010).

There are three main groups where sentiment lexicons creation methods can be included. The first is manual labelling where one or several volunteers/workers label a list of words with sentiment and then, use metrics to determine inter-worker agreement (Mohammad and Turney, 2010; Taboada et al., 2011; Hutto and Gilbert, 2014; Nielsen, 2011). However, this approach can be time consuming, increasing with the size of the word list and the number of different evaluations required for each word. It can also be expensive if we resort to services like Mechanical Turk (Amazon, 2016) or CrowdFlower (CrowdFlower, 2016) where a fee must be paid to each worker who completes the classification task.

Therefore, more automatic ways of creating sentiment lexicons were proposed. These require a small sample of sentiment labelled terms, normally named "seed words", and then expanding the lexicon using those words as base. Two different approaches have been used for expanding the lexicon in semi-supervised fashion: thesaurus based approaches and corpus based approaches.

Thesaurus based approaches rely on other syntactic resources like the General Inquirer (GI) (Stone et al., 1966) or WordNet (Fellbaum, 1998). WordNet is a large lexical resource containing noun, verbs, adverbs and adjectives grouped by synsets which are sets of cognitive synonyms. If the word is an adjective, a set of antonyms is also available. Some works like SentiWordNet used this features and a small number of labelled words to expand sentiment lexicons by assigning the same polarity of a word to its synonyms and opposite polarity to antonyms (Baccianella et al., 2010; Esuli and Sebastiani, 2006). However, the authors in (Mohammad et al., 2009) present better sentiment accuracy in words than SentiWordNet1.0 by using a Roget-like thesaurus. Several studies (Kim and Hovy, 2004; Hu and Liu, 2004a) also used WordNet to expand sentiment lexicon, making it one of the most used resources for lexicon expansion. One of the major problems on this thesaurus based approaches is the domain specific context on each opinion word. The word "loud" can have a negative orientation in a car review but positive sentiment in a speaker review. For more domain specific lexicon expansion, the corpus-based approaches are a better solution.

In (Hatzivassiloglou and McKeown, 1997) a corpus based lexicon expansion method is proposed us-

ing conjunction rules to infer new opinion words specific to the domain. For example, in the review *"The Samsung remote is awesome and **easy** to use."*, if we know that "awesome" has a positive sentiment then, due to the conjunction AND, we can infer that "easy" or "easy to use" has also a positive sentiment associated. In the same way, on the video game review *"The game has beautiful graphics but **easy** to complete."*, if we know that "beautiful" has a positive polarity we can infer that the conjunction BUT will reverse the polarity on "easy". The authors named this concept as "sentiment consistency".

Another proposal for corpus based lexicon expansion is presented in (Qiu et al., 2011). It uses a set of seed words combined with conjunction rules for extracting entities and opinion words. Then, through an iterative process, the new pairs of entities/opinion words are used for finding more pairs and ends when no new entities or opinion words are found. Evaluation on reviews dataset showed that this method outperforms other state of the art approaches (such as the one in (Hu and Liu, 2004a)).

However, not always opinion words have the same polarity, even in the same domain. For instance, in a laptop review, *"the battery is long"* is identified as positive whereas *"it takes to long to start"* is associated with a negative sentiment. So, to avoid erroneous sentiment classification, the use of entity level sentiment analysis techniques and the extraction of the ternary (word,entity, sentiment) was proposed for lexicon expansion (Ding et al., 2008).

Besides reviews, social networks have been explored for corpus based lexicon expansion. As a matter of fact, many social networks have specific opinion words that are normally not covered by the general sentiment lexicons (e.g. "ahahahah", "LOL", "OMG", "#hatemonday"). The study in (Bravo-Marquez et al., 2015a) present two models for creating a Twitter specific lexicon from a unlabelled corpus of tweets using tweet-centroid word vectors. The lexicon is classified into Positive, Neutral and Negative scores. Another work by the same authors (Bravo-Marquez et al., 2015b) presents a supervised algorithm for lexicon expansion using tweets label with emoticons and a combination of several seed word lexicons. Other supervised approach (Tang et al., 2014) uses SkipGram (for learning continuous phrases representation) and a seed lexicon (expanded with contents from the Urban Dictionary (Dictionary, 2016)) as training data for a sentiment lexicon expansion classifier. One more study (Du et al., 2010) shapes the information bottleneck method with cross-domain and inter-domain knowledge to extract a domain oriented lexicon.

A rather different approach is the one presented in (Feng et al., 2011). Whereas most of the methods presented focus on expanding sentiment lexicons with adjectives and verbs, Feng et al. study the influence of words with connotative polarity such as *cancer*, *promotion* and *tragedy*. Furthermore, they also use an unusual graph approach which incorporates with the PageRank algorithm and a seed of opinion words to propose a connotative lexicon creation system.

In fact, the majority of works study how to expand sentiment lexicons with verbs and adjectives. In some contexts, nouns may also imply opinion. For example in the mattress review *"Within a month, a **valley** formed in the middle of the mattress"* or in the tablet review *"It came with a **scratch** in the screen"*. The authors in (Zhang and Liu, 2011) study nouns that may imply sentiment in product features. The study relies on an seed lexicon to identify the sentiment on reviews and then select candidates for feature nouns that suggest opinion.

The detection of sentiment in words other than adjectives and verbs is yet an understudied research area. Therefore, in this work it is the exploration of assigning sentiment to connotative words, nouns that imply opinion, entities and topics that it will be highlighted. We intend to expand even more the sentiment lexicons in this studies by using public opinion as a measure of polarity, combining Twitter sentiment analysis and lexicon expansion methods to create new domain and time specific sentiment dictionaries.

3 WORKFLOW DESCRIPTION

In the following section we describe the workflow of our lexicon expansion proposal. We select terms from seven different domains: world, health, entertainment, politics, business, sports, and technology. For each domain we have a set of RSS URLs from several news websites in the English language (e.g. CNN, BBC, The New York Times). In each RSS feed, only the headline for each news is extracted since: 1) it summarizes the full article and 2) its short length provides an easier filtering of irrelevant words or terms. This way, we create a text *corpus* composed only of news headlines, from several sources, for each domain.

3.1 Term Extraction

For each domain corpus, we construct a term-document frequent matrix and retrieve the most frequent occurrences of 1-grams (words), 2-grams (two word terms) and 3-grams (three word terms).

The terms we define as "frequent" rely on the number of sources we have for the domain and the grams we are considering. The formula used to determined the frequency threshold (and therefore to decide if a term should be included in the lexicon) is presented in (1).

$$\text{frequency threshold}_{d,i} = n_d \times a_i \quad (1)$$

where n_d is the number of sources for domain d and a_i represents the percentage of the cut in each i-gram. In other words, if a term occurs more than the frequency threshold variable it is included in the lexicon. The values for a_i were reached experimentally and are presented in (2).

$$a_1 = 0.50; a_2 = 0.30; a_3 = 0.25 \quad (2)$$

It is important to filter some of the "noisy" terms (i.e. terms that are irrelevant for sentiment analysis) from the list extracted. The 1-gram are the ones that commonly have the most noisy data. Several filters are applied in order to reduce it. First, the words are classified with the OpenNLP Parts-of-Speech tagger (Apache, 2010). Then, only words classified as nouns, foreign words and adjectives are kept. Verbs are excluded due to the lack of context. For example, "wins" or "lost" are generally associated with a positive and negative sentiment, respectively. However, if we know to whom, or what, it refers to (e.g "Trump wins" or "Hillary lost") then the public sentiment of the term may not be the same. Then, a list of domains of specific words is used to remove 1-grams that do not infer any particular sentiment. We use Topic Dictionaries from (Oxford, 2016) to achieve that purpose. We left, however, words that refer to corporations and entities (e.g. "Apple" and "Microsoft" in technology domain). In addition, we also remove words that are common in the news (e.g. "review", "tech", "news"). Furthermore, words that are repeated in plural form ("syrian"/"syrians") and with apostrophe ("Trump"/"Trump's") are only kept in singular and non-apostrophe form. We also remove words that are in the AFINN (Nielsen, 2011) sentiment lexicon because those words by themselves already express sentiment.

Since the number of 2-grams and 3-grams terms obtained are less than the number of 1-grams and, because they appear frequently together (meaning that they already have relevance in the domain), only the plural and apostrophes filter is applied. We then send the terms to Twitter where a last filter on our final terms list is used. This filter relies on the number of tweets found on the terms. If it is lower than a defined threshold, it will not be included in the terms list.

The number of extracted terms is dynamic and highly depends on the relevance that they have in news media. In our work, we consider the extracted terms relevant since: 1) they appear multiple times in the same domain in several different news sources, and because 2) when querying Twitter there is a significantly high number of tweets regarding those terms on the same day they are extracted from the news sources. Our method requires that there is a minimum number of tweets that include the term. If that number is not fulfilled, the term is removed since it is likely to be irrelevant or syntactically incomplete (e.g. from the headline "Zika virus found in Montana", the term "found in" is not relevant).

3.2 Term Sentiment Analysis

To evaluate the sentiment of each term extracted from the headlines of RSS news feeds, we use the Twitter Search API (Twitter, 2016). Unlike similar studies who evaluate the sentiment of terms in the comments from users on news site (Moreo et al., 2012) or who select specific keywords from Twitter streams (Wang et al., 2012; Nguyen et al., 2012), our system uses a combination of both approaches. Using tweets, we guarantee that the opinions retrieved are not completely anonymous (like in the majority of news website) and therefore hate, advertise and insulting comments are less common.

In addition, several works have proved good results using Twitter for topic tracking (Wang et al., 2012; Amer-Yahia et al., 2012; Phuvipadawat and Murata, 2010) and Twitter users tend to react quickly to the occurrence of events which lead to several techniques for detecting real-time events on the social network (Atefeh and Khreich, 2015).

Moreover, using news headlines to search for terms to track in each domain, we guarantee that they are up to date and are relevant.

When the extraction procedure for all the domain finishes, the resulting terms are searched in Twitter and a sample is retrieved for each one of them. In our experiments we use a sample of 500 tweets for each term. In order to keep the term sentiment updated, the tweets extracted must be posted in the same day as the keyword extraction. In addition, we only get the more recent tweets. Furthermore, in order to remove tweets that can be from Twitter accounts who belong to news sites or newspapers, we apply a filter in our query that does not retrieve tweets containing links. This is due to the nature of tweets posted by news site accounts, which contain the link for accessing the full news in the correspondent website.

Using the Twitter API, the number of tweets extracted for each term is not always the same as request. This can also be used as a last filter for terms

extraction. In fact, if a keyword does not retrieve a minimum amount of tweets, this can be interpreted as non relevance or lack of meaning of the keyword. So in our system, if the keyword searched in Twitter does not retrieve a minimum number of tweets, it is discarded. In our experimental setup we used 33% of the sample as the threshold for not discarding the term.

The next step is to perform sentiment analysis on each tweet from the sample. We separate our method in two components: the syntactic analysis of the tweet and the identification and assessment of possible emojis and emoticons present in the tweet. Several studies (Go et al., 2009; Novak et al., 2015) provide evidence that emojis and emoticons are used as sentiment clues. In fact, emoticons were already used as classifiers of the polarity of the tweet since they are not specific to a certain domain (Hogenboom et al., 2013).

In order to evaluate emojis we use the results from (Novak et al., 2015) where the authors assess the sentiment of each emoji. As for the emoticons we consider the sentiment classification used in (Hogenboom et al., 2015). In our system, the emojis and emoticons in the tweet have the same sentiment impact independently of the position where they occur. All the emojis/emoticons are considered and repetitions are not discarded. The sentiment is calculated by simple average (summing all the identified emojis and emoticons and dividing by the number of occurrences).

The syntactic component of our analysis is to evaluate the sentiment on the text of each tweet. With the goal of not inducing wrong sentiment, we remove the term queried from each tweet. Hence, terms like "Trump wins" which already have a positive sentiment associated (due to the word "wins") do not skew our analysis. To determine the sentiment on the remaining words from the tweet, the general sentiment lexicon AFINN (Nielsen, 2011) is used. We choose this lexicon because, unlike more classical proposals (Hu and Liu, 2004b) in which sentiment words are classified only in a polarity fashion, AFINN provides 2477 words classified with sentiment in a $[-5,5]$ interval. In addition to the sentiment lexicon, we also use lists of amplifiers (e.g. "very", "extremely", "more") and attenuation (e.g. "few","little","rarely") words for better sentiment analysis. These assign a weight of 80% on the word polarity. Furthermore, we use a list a words that reverse the polarity (e.g. "not", "nobody", "never").

Due to the tweets limitation of 140 characters, the sentiment value of a word is affected if there is any element of the lists mentioned in the 4 words before and/or in the following 2. In other words, for each opinion word found in the tweet we create a cluster with the previous 4 words and the next 2. Then, we verify if any of them match the words in the amplification, attenuation or negation lists and assign the sentiment accordingly. The syntactic sentiment score of each tweet is calculated combining the lexicon and method previously mentioned. The final score for each tweet is a weighted average of 75% the text sentiment analysis and 25% the emoticon/emoji sentiment analysis. The assumption is that the average sentiment of the sample represents the overall sentiment of the searched term. Therefore, the term sentiment score is calculated by summing up the sentiment score of each tweet from the term *corpus* and dividing it by the number of tweets in that *corpus*.

4 EVALUATION

4.1 Tweets Sentiment Analysis

Since one of our goals is to assess if the polarity of the term can be obtained from the average sentiment of a sample of tweets containing the term, it is important to have a accurate tweet sentiment classifier. Therefore, we evaluate and compare our polarity classification method in several datasets provided by CrowdFlower (in their "Data for Everyone" library). Five datasets of tweets, classified with sentiment by human coders were used. A brief explanation on each dataset follows (for more details please refer to (Crowdflower, 2016)):

- **GOP:** contains over ten thousand tweets about the GOP debate in Ohio. Workers classified the sentiment of each tweet as Positive, Neutral or Negative.
- **SDC:** includes approximately 7000 tweets about self driving cars. Workers were asked to classify the sentiment as Very Positive, Slightly Positive, Neutral, Slightly Negative, Very Negative. We converted this to a Positive/Neutral/Negative scale.
- **USAIR:** dataset with around 16000 tweets about major US airlines. Contributors were asked to assign a Neutral, Positive or Negative sentiment to each tweet.
- **COACH:** dataset with 3847 tweets with reactions to the 2015 Coachella festival lineup announcement. Workers classified the sentiment of each tweet as Neutral, Positive or Negative
- **APPLE:** 4000 tweets containing references to the Apple company. Sentiment classification was done with a Negative, Neutral and Positive scale.

We are interested in determining the polarity in 2 classes (positive/negative) of each of the extracted terms. Therefore, we discarded the neutral tweets from the datasets. The results of that score in terms of, precision, recall, f1-measure and accuracy can be examined in Table 1.

Table 1: Results in terms of precision (Prec.), recall (Rec.), F1-Measure (F1), and accuracy (Acc).

Dataset	Prec. (%)		Rec. (%)		F1 (%)		Acc. (%)
	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.+Neg.
GOP	32.8	90.9	78.3	57.5	46.2	70.4	61.8
SDC	82.5	53.9	76.4	63.1	79.3	58.1	72.3
USAIR	39.4	97.4	94.9	56.9	55.7	71.9	65.6
COACH	85.3	42.5	74.2	59.7	79.3	49.6	70.7
APPLE	55.4	93.9	86.8	74.4	67.7	83.0	77.7

When analyzing each of the datasets, the sentiment component of our system seems to achieve better performance in tweets regarding the technology domain (SDC and APPLE). However, variation on accuracy values does not surpass 20% which gives a solid support that our method will perform well independently of the tweets domain. Accuracy reaches the lowest value in the GOP dataset. Similar conclusion was reached in (Thelwall et al., 2012) where the authors assess the low performance on some web extracted datasets due to political and controversial topics. In addition, we compare our sentiment component (SC) to other state-of-art methods to check if it was able to match them as 2-class (positive/negative) accuracy is concerned. A brief description on each system follows:

- **Emolex:** Manually created emotion lexicon using crowd-sourcing. The terms were extracted from a combination of Macquarie Thesaurus, General Inquirer and WordNet Affect Lexicon (Mohammad and Turney, 2010). Although the words were classified with emotion and polarity, only the second was used for this method.
- **SenticNet:** Assigns sentiment to common sense concepts to achieve a semantic sentiment analysis approach rather than the most common sentence level (Cambria et al., 2014).
- **SentiStrength:** Combines a manually annotated sentiment lexicon, machine learning algorithms and other important features like negation words and repeated punctuation for sentiment enhance. It provides the best results in gold standard tweet datasets (Thelwall et al., 2012).

The results can be seen in Table 2.

Although it is not the best system when compared to other state of the art approaches, our sentiment component still performs well on the different datasets achieving the best accuracy in 2 of them.

Table 2: Comparison of the sentiment component (SC) of our system with other state of the art approaches.

Sentiment System	2-Class Dataset Accuracy					
	GOP	SDC	USAIR	COACH	APPLE	Average
SC	61.8	**72.3**	65.6	70.1	**77.7**	69.6
Emolex	46.0	64.9	46.9	65.6	70.3	58.7
SenticNet	37.3	68.5	39.1	74.7	46.9	53.3
SentiStrength	**70.4**	70.1	**76.5**	**73.3**	74.5	**73.1**

In addition only is beaten by 4% margin by SentiStrength when assessing the overall accuracy.

In conclusion, these results provide a good support for the reliability on tweet classification of our system.

4.2 System Evaluation

In order to evaluate our system we carried out two experimental surveys. The first had the goal of assessing the effectiveness of our proposal in extracting relevant terms for each of the domains. The second survey was to evaluate if the sentiment assigned to each term was still accurate at present time.

The survey was conducted in a web application built for the effect. The question asked was ”Considering the present time (and current news), does the term x fits the domain y ?” where x and y were replaced randomly by the entries extracted from our system. Since our goal was solely to test our extraction method we allow users to classify an unrestricted number of entries. We obtained a total of 1414 entries classified by 57 different users consisting mostly of university students. When evaluating the fitness of the term in the domain we discarded all the entries of users whose response was ”I don’t know”. In the remaining 1336 entries we had an accuracy of 88.2%. This provides strong empirical evidence for our term selection method.

The second part of our study was to determine if Twitter sentiment on an extracted term reflects the current sentiment of the term. To assess this we used Crowdflower to conduct a sentiment survey. We used terms extracted from 2016-04-01 till 2016-04-03. The experience began on 2016-04-04 at approximately 3:15 pm and took 30 hours to complete. We submitted 101 pairs of terms/domain extracted randomly (but in equal number for each domain) from the daily retrieved dictionaries. Each of those terms was evaluated by 7 different workers with a level 3 performance. This level is assigned to workers who achieved high accuracy values in more than a hundred test questions (Crowdflower, 2014). The question asked in this CrowdFlower survey was: ”Considering the present time (and current news) and the domain x, please rate the sentiment associated with the

expression *y*" where *x* is the domain and *y* the term. The scale provided ranged from 1 (very negative) to 5(very positive). Although we are trying to assess the general polarity of the term, we used a likert scale to force workers to have a more careful decision on which sentiment to choose, avoiding a randomly (and easiest) choice.

We used the median as measure to determine our ground truth for each term since the average value could be highly influenced by possible outliers. For example, if six workers evaluate the term with a 2 and a worker with a 5 the average value would result in a final sentiment of 3 (neutral value). Using the median the final sentiment would result in a more realistic 2 (negative polarity).

We converted the results to fit our polarity scale. Values below 3 were classified as negatives and above as positives. Once again, we discarded the neutral values since our system assigns a positive/negative output for each term. We notice however, that the number of terms classified as neutral was significantly high (around 40% of our sample). This experimental results suggest that an implementation of a neutral classification must be accounted in future work.

As it was already mention, there are two types of automatic lexicon expansion methods: thesaurus and corpus based. However (and although we consider our approach to fit the corpus based category), our system cannot be compared to any of those methods. This is because traditional corpus based methods focus solely in one corpus and retrieve the sentiment words of it. However, our proposed method, generates a corpus for each extracted term. Furthermore, most of the state of the art approaches focus in retrieving opinion words classified majorly as adjectives and verbs. Consequently, any term comparisons with other methods is hard to achieve. Therefore, we compare our results against a random baseline (achieved with the best overall accuracy of 5 attempts) and a majority baseline (which classifies all terms as the the class who is more frequent). The results are presented in Table 3.

Table 3: Comparison of results of our system (SR) against a random baseline(Rbl) and a majority baseline (Mbl).

	Prec.(%)		**Rec.(%)**		**F1(%)**		**Acc.(%)**
	Pos.	**Neg.**	**Pos.**	**Neg.**	**Pos.**	**Neg.**	**Pos+Neg**
SR	74.36	90.00	93.55	64.29	82.86	75.00	79.67
Rbl	65.71	66.67	74.19	57.14	69.70	61.54	66.10
Mbl	52.54	NA	100	0	68.89	NA	52.54

Experimental results show good overall accuracy of 79.7%. A closer analysis on the predictions of the system has revealed a particularly low performance on political terms. This is presumably because several of the used terms have a rather controversial sentiment. As an example we have "abortion", "national living wage", and political candidates in US elections such as "Donald Trump", "Hillary Clinton" or "Bernie Sanders". In the entertainment domain the results are much better, missing solely in "batman v superman".

We are aware that our experiments involved a small number of terms. However, since we are evaluating time and domain specific terms, including more terms in our analysis from extractions further back in the past would not correspond to what we are trying to assess. We also considered extending to more domains but defining the "ground truth" sentiment in domains which have a narrowed scope could result in more neutral classifications due to unfamiliarity of the term to the workers.

5 CONCLUSIONS AND FUTURE WORK

In this work we have described a system for automatically extracting the more relevant terms from seven different domains and to classify their sentiment in a positive/negative scale. Our proposal retrieves the more frequent terms from news headlines using RSS feeds from several news sources. We then query Twitter with the same terms and infer their polarity using the average sentiment classification obtained from the sample of tweets.

Our experiments shown that the proposed term extraction component is rather effective, achieving a 80% accuracy. Some of the limitations of our method are due to the accuracy of the used NLP classifier that lead to some noisy unigram terms. Future research will try to explore more filters for a fine grain selection of unigrams in the different domains. Possible filters may include the use of different NLP classifiers to determine the part of speech tags and use name entity recognition techniques to infer terms that are referring to the same entity (e.g. "Obama"and "POTUS"). We also plan to uncover the relations between the 1-gram, 2-gram and 3-gram lists. For example, although the terms "april fools'", "fools' day" and "april fools day" are expected to have similar polarity, the terms "Syria" and "Syria ceasefire" are not.

Our sentiment classifier also produced good results in detecting the polarity of tweets from several different domains. Tests on labelled Twitter datasets achieved an overall accuracy ranging from 61.8% (GOP dataset) to 77.7% (APPLE dataset). Furthermore, when compared to other state of the art systems for sentiment analysis, it was only surpassed by SentiStrength by a minimal 4% margin.

The two preliminary evaluation experiments we have described have provided strong evidence on the validity of our approach. Experimental results using Crowdflower lead to an overall accuracy of 79.67% with positive terms achieving better f1-measure (82.86%) than the negative ones (75.00%). In future work, we will use the results from our system to complement and expand sentiment lexicons for domain and time specific contexts. We intend to assess if these lexicons can improve sentiment classification on dictionary-based approaches specifically on short informal texts (like tweets or website comments).

ACKNOWLEDGEMENTS

This work is supported by the ERDF European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT (Portuguese Foundation for Science and Technology) within project Reminds/UTAP-ICDT/EEI-CTP/0022/2014.

REFERENCES

Amazon (2016). Amazon mechanical turk. https://www.mturk.com/mturk/welcome. Acessed: 2016-08-21.

Amer-Yahia, S., Anjum, S., Ghenai, A., Siddique, A., Abbar, S., Madden, S., Marcus, A., and El-Haddad, M. (2012). MAQSA: a system for social analytics on news. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 653–656.

Apache (2010). Apache OpenNLP . https://opennlp.apache.org/. Acessed: 2016-08-21.

Atefeh, F. and Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164.

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

Bravo-Marquez, F., Frank, E., and Pfahringer, B. (2015a). From unlabelled tweets to twitter-specific opinion words. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 743–746, New York, NY, USA. ACM.

Bravo-Marquez, F., Frank, E., and Pfahringer, B. (2015b). Positive, negative, or neutral: Learning an expanded opinion lexicon from emoticon-annotated tweets. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '15.

Cambria, E., Olsher, D., and Rajagopal, D. (2014). Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 1515–1521. AAAI Press.

Crowdflower (2014). Introducing contributor performance levels. http:// crowdflowercommunity.tumblr.com/post/80598014542/introducing-contributor-performance-levels. Acessed: 2016-04-10.

CrowdFlower (2016). Crowdflower: Make your data useful. https://www.crowdflower.com/. Acessed: 2016-08-21.

Crowdflower (2016). Data for everyone. http://www.crowdflower.com/data-for-everyone. Acessed: 2016-04-10.

Dictionary, U. (2016). Urban dictionary. www. urbandictionary.com. Acessed: 2016-08-21.

Ding, X., Liu, B., and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining.

Du, W., Tan, S., Cheng, X., and Yun, X. (2010). Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 111–120, New York, NY, USA. ACM.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06*, pages 417–422.

Fellbaum, C., editor (1998). *WordNet: an electronic lexical database*. MIT Press.

Feng, S., Bose, R., and Choi, Y. (2011). Learning general connotation of words using graph-based algorithms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1092–1103, Stroudsburg, PA, USA. Association for Computational Linguistics.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12.

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 174–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., and Kaymak, U. (2013). Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th An-*

nual ACM Symposium on Applied Computing, SAC '13, pages 703–710, New York, NY, USA. ACM.

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., De Jong, F., and Kaymak, U. (2015). Exploiting emoticons in polarity classification of text. *J. Web Eng.*, 14(1-2):22–40.

Hu, M. and Liu, B. (2004a). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Hu, M. and Liu, B. (2004b). Mining opinion features in customer reviews. In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, pages 755–760. AAAI Press.

Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Adar, E., Resnick, P., Choudhury, M. D., Hogan, B., and Oh, A. H., editors, *ICWSM*. The AAAI Press.

Kim, S.-M. and Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kiritchenko, S., Zhu, X., and Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *J. Artif. Int. Res.*, 50(1):723–762.

Mohammad, S., Dunne, C., and Dorr, B. (2009). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 599–608, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mohammad, S. M. and Turney, P. D. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 26–34, Stroudsburg, PA, USA. Association for Computational Linguistics.

Moreo, A., Romero, M., Castro, J., and Zurita, J. (2012). Lexicon-based comments-oriented news sentiment analyzer system. *Expert Syst. Appl.*, 39(10):9166–9180.

Nguyen, L. T., Wu, P., Chan, W., Peng, W., and Zhang, Y. (2012). Predicting collective sentiment dynamics from time-series social media. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '12, pages 6:1–6:8, New York, NY, USA. ACM.

Nielsen, F. A. (2011). Afinn.

Novak, P. K., Smailovic, J., Sluban, B., and Mozetic, I. (2015). Sentiment of emojis. *CoRR*, abs/1509.07761.

Oxford (2016). Oxford Learner's Dictionaries topic dictionaries. http://www.oxfordlearnersdictionaries.com/topic/. Acessed: 2016-07-03.

Phuvipadawat, S. and Murata, T. (2010). Breaking news detection and tracking in twitter. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 3, pages 120–123.

Qiu, G., Liu, B., Bu, J., and Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Comput. Linguist.*, 37(1):9–27.

Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307.

Tang, D., Wei, F., Qin, B., Zhou, M., and Liu, T. (2014). Building large-scale twitter-specific sentiment lexicon : A representation learning approach. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 172–182.

Thelwall, M., Buckley, K., and Paltoglou, G. (2012). Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol.*, 63(1):163–173.

Twitter (2015a). Twitter Company about. https://about.twitter.com/company. Acessed: 2015-10-19.

Twitter (2015b). Twitter Company rest. https://dev.twitter.com/rest/public. Acessed: 2015-10-19.

Twitter (2016). Twitter Developers. https://dev.twitter.com/. Acessed: 2016-03-08.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., and Narayanan, S. (2012). A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 115–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhang, L. and Liu, B. (2011). Identifying noun product features that imply opinions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 575–580, Stroudsburg, PA, USA. Association for Computational Linguistics.