Lexicon Expansion System for Domain and Time Oriented Sentiment Analysis

Nuno Ricardo Pinheiro da Silva Guimarães
Mestrado em Ciência de Computadores
Departamento de Ciência de Computadores
2016

Orientador
Luís Torgo, Professor Associado, Faculdade de Ciências da Universidade do Porto

Coorientador
Álvaro Figueira, Professor Auxiliar, Faculdade de Ciências da Universidade do Porto

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, / /

To my parents

Acknowledgements

First, I would like to express my gratitude to my supervisor Luís Torgo, for always being available to help and for his guidance and mentoring throughout the course of this work. I would also like to thanks my co-supervisor Álvaro Figueira, not only for giving me the necessary tools to perform the evaluation component (which was essential to my work), but also for his advices and constant support.

In addition, I would also like to thank project REMINDS for my research grant. I would also like to express my gratitude to the remaining persons on project for the help provided during this work. Project REMINDS was financed by 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.

I would also like to thank all the "incredible" persons on the Department of Computer Science. Their presence has always be constant and helpful throughout the course of this thesis.

A special thanks to João Queirós, Joana Gonçalves, Jorge Silva, Patrícia Santos, Rui Fonseca, Tânia Rodrigues, António Pinto, Diogo Teixeira, Rui Monteiro, Rafael Carolo and Rui Gomes for the friendship and for the happiness they provided me, even through tough times. I would also like to thank Inês for her support, kindness and tenderness.

Last but definitely not least, I would like to express my gratitude to my family. They will always be there for me and I am eternally grateful for that.

Nuno Guimarães
Porto, 2016

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. Research has focused in these particular parts of speech. 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 an opinion without the use of words included 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, 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 a 90% accuracy on extracting domain and time specific terms and 80% on correct polarity assessment. In addition, an analysis on the sentiment dynamics and "trending" factor of a sample of terms (that were frequently referred in the news) was carried out. An association on the term polarity change and trend was not possible. However, the variation on the terms sentiment seems to be the expected through the time interval analysed. The achieved results provide evidence that our lexicon expansion system can extract and determine the sentiment of terms for domain and time specific corpora in a fully automatic form.

However, some flaws were detected during the evaluation. Namely, the large number of terms evaluated with a neutral score provided evidence that terms that appear on news may not have always a positive/negative value. Therefore, the implementation of a three class ensemble sentiment method was included in the workflow of our system. Evaluation using tweets datasets from different domains proved that our ensemble system ENS17 outperformed 19 other sentiment analysis methods. In addition, a tweet domain disambiguation process was included for the specific cases where the same term appears in more than one domain.

Preliminary evaluations were made by adding the resulting lexicons to state of the art sentiment systems and testing them on a dataset containing tweets and Facebook posts and comments. The results show that our expanded lexicons cannot be used directly on texts retrieved from the web (namely in factual or news texts). However, they improve the sentiment classification of all three methods on opinion texts, providing evidence of their usefulness on sentiment analysis classification.

Keywords: sentiment lexicon, sentiment lexicon expansion, sentiment analysis, term extraction, text mining

Resumo

Na análise de sentimento, a polaridade de um texto é frequentemente calculada usando léxicos de sentimento, que normalmente são constituídos por verbos e adjectivos classificados com um valor positivo ou negativo. No entanto, em textos curtos e informais (como *tweets* ou comentários) a ausência desse tipo de palavras não significa necessariamente que o texto não contém opinião. *Tweets* como "Primeiro Paris, agora Bruxelas... O que podemos fazer?" implicam uma opinião, não porque contêm palavras pertencentes aos léxicos de sentimento, mas sim devido ao conhecimento da opinião pública associada a alguns termos, num domínio e intervalo de tempo específico.

De modo a complementar os dicionários de sentimento gerais com esses termos, este trabalho propõe um sistema para expansão de léxico que automaticamente extrai os termos mais actuais e relevantes em diferentes domínios e avalia o seu sentimento através do *Twitter*.

Resultados experimentais mostram uma exactidão de 90% na extracção correcta dos termos relativos ao domínio e tempo e 80% na classificação correcta em termos de polaridade. Além disso, foi conduzida numa amostra de termos (que consistentemente apareciam nas notícias), uma análise da dinâmica temporal do sentimento e factor "*trending*". Não foi possível concluir uma relação entre a mudança de polaridade e factor "*trending*" de um termo. No entanto, a variação do sentimento detectada aparenta ser a expectável durante o intervalo de tempo analisado. Os resultados obtidos indicam que o sistema de expansão de léxico proposto consegue extrair e classificar termos de uma forma completamente automática.

No entanto, algumas falhas foram detetadas. Por exemplo, um elevado número de termos classificados como neutros sugere que nem sempre os termos relevantes e que aparecem nas notícias estão associados a um sentimento positivo ou negativo.

Portanto, foi incluída no nosso sistema, uma combinação de métodos (*ensemble*) para a classificação de sentimento. A avaliação usando *datasets* de *tweets* provaram que o nosso sistema por combinação de métodos, supera 19 métodos individuais de análise de sentimento. Além disso, foi criado um processo de desambiguação de *tweets* para os casos onde o mesmo termo aparece em domínios diferentes.

Foram realizadas avaliações preliminares adicionando os léxicos resultantes a 3 diferentes métodos de análise de sentimento, testando-os num *dataset* com *tweets* e *posts* e comentários do *Facebook*. Os resultados mostram que os nossos dicionários não podem ser aplicados diretamente em todo tipo de textos provenientes da *web*. No entanto, os dicionários criados melhoram a classificação de sentimento nos 3 sistemas testados em textos não-factuais, provando assim a eficácia do método proposto.

Palavras-chave: léxicos de sentimento, expansão de léxicos de sentimento, análise de sentimento, extração de termos, *text mining*

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Context

Throughout the years, opinion had significant importance on decision making across several fields. It is common knowledge that people ask and consider friends opinions before making decisions, whether it is on the quality of a certain product or a service (like restaurants and hotels). Therefore, people's opinion has an impact on the economy since, for example, client's opinion regarding a company products may influence the growth or downfall of sales [131] and in the release of a new product, negative opinion reviews represent a major influence on sales [25].

Opinion has also an important role in politics. During campaign periods several polls are conducted in order to predict the winner. Furthermore, it helps candidates to know which concerns to address or to avoid in different states or regions and how to manage campaign funds in order to get more voters. Moreover, the polls themselves can also influence voters [11], thus increasing the importance of public opinion.

Before the World Wide Web, companies spent time and resources in the elaboration of surveys to keep track of public opinion of their products. In same way, in a political domain and during the campaign periods, several polls were conducted in order to predict the winner and help candidates to plan the next steps.

Nowadays, information has become much more accessible. The emergence of blogs, reviews sites, and social networks have facilitated the decision making process. If a buyer wants to assess the quality of the product that intends to purchase, the Internet has provided means to make a much more informative decision.

Therefore, company's interest to evaluate and track the global opinion on their products or services has also grown, since now, million of users express their opinion online in social networks, blogs, review sites and forums, just to name a few. However, due to its unique characteristics, Twitter has been the more common and widely used service to access sentiment or opinion on different topics.

Twitter is a social network that allows users to send small pieces of text, called tweets, up to a limit of 140 characters. Besides normal text, tweets allow users to include hashtags, emojis and mentions to other users. Hashtags are tag words beginning with the character "#" whose purpose is to identify tweets with common topics and emojis are special figures that can be incorporated in the text. Other users can be mentioned in tweets by using the "@" character followed by the username. External links can also be included.

Twitter makes a good data source for public opinion analysis 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 active users per month [121]. Moreover, Twitter provides developers a public API allowing, among others features, the retrieval of tweets, getting user information and monitoring tweets in real time making it easier to retrieve large quantities of data for analysis [122]. Finally, Twitter also is an updated source of information since tweets are sent in real time about different topics from users with different opinions.

With such large amount of data publicly available and easily accessible, the interest on methods to automatically analyse and extract the opinion from that data has arise.

1.2 Definition of concepts

With the technological evolution, the volume of data that can be stored has increased significantly. This amount of information has reached a state where human analysis and comprehension is impossible without the help of data analysis tools that could transform raw data into useful knowledge.

Data Mining is the process of automatically extracting that knowledge. One of the most common example where Data Mining techniques can be applied is supermarket purchases. Several Data Mining techniques are used to analyse customers transactions and suggest other related products (recommendation systems).

This concept can be further divided in more specific areas. Text Mining is the one we will focus and consists on retrieving relevant information from texts. Common goals in Text Mining are for example topic extraction from a document, automatic summarising and automatic translation of texts. This work falls within another field of Text Mining called Sentiment Analysis or Opinion Mining.

The main goal of Sentiment Analysis is to determine the opinion or sentiment in a text. This sentiment can be relatively to the totality of the text or to the entities presented in the text. The sentiment classification can also be done in a polarity fashion (positive/negative sentiment) or in a ranged scale (e.g. from -5 to 5). Sentiment analysis can also be done using supervised methods (where a model is trained with sentiment labels and then tries to predict new unclassified data) or unsupervised methods (where no prior knowledge is given to the model).

In several sentiment analysis approaches, it is common to use sentiment dictionaries or lexicons. They consist of lists of words with a specific sentiment associated to each word and are used to identify the general sentiment of a text. For example, a common and simple unsupervised approach is to sum the sentiment of all the words found in the text that are present in the dictionary.

1.3 Research Hypothesis

One of the main problems related with sentiment dictionaries (or lexicons) is the source used and the words they contain. Currently, there are several automatic and manually labelled sentiment dictionaries in different sizes. However, the vast majority focus on opinion words such as adjectives like "beautiful" and "awful" or verbs like "lost" and "wins". Connotative words that are neither a verb nor an adjective, such as "cancer" and "terrorist", are not normally considered. Furthermore, when evaluating short informal texts, the absence of opinion words does not always implies the absence of sentiment. People often tend to use common knowledge to express an opinion without the use of opinion words. For example, in the sentence "After Paris, Brussels. When it will end?" there is a clear presence of a negative sentiment (due to the terrorist attacks that happened in both cities), but no opinion words to support it. This is because the opinion is expressed using facts regarding Paris and Brussels which are normally common knowledge due to the impact that they had on news and the way the public reacted to it. In addition, time gains a specific importance when we are dealing with this type of sentiment analysis with absence of opinion words. In fact, for most people, the fragment above would have no meaning by itself or sentiment associated prior to the terrorist attacks [37].

Besides time, these words must also have a domain associated to them. For example, if we consider the text fragment "listening to Prince, I still can believe it". "Prince" is, in a general context, a member of royalty. However, in this particular sentence (which we can associate to the entertainment/music domain) "prince" refers to the artist.

Therefore, it is plausible to say that the sentiment of terms, like the ones mentioned above, may vary through time and according to the domain. This is more visible if we consider entities like persons or enterprises. Again, in the example of the word "Paris", it is fair to presume that the sentiment of the word was different before and shortly after the terrorist attacks on the city as Figure 1.1 illustrates. 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.

Therefore, our research questions state the following:

- Can Twitter be a good source on the sentiment of time and domain dependent terms?
- Can domain and time dependent sentiment lexicons (consisting of parts of speech other than verbs and adjectives) improve the performance of sentiment analysis methods?

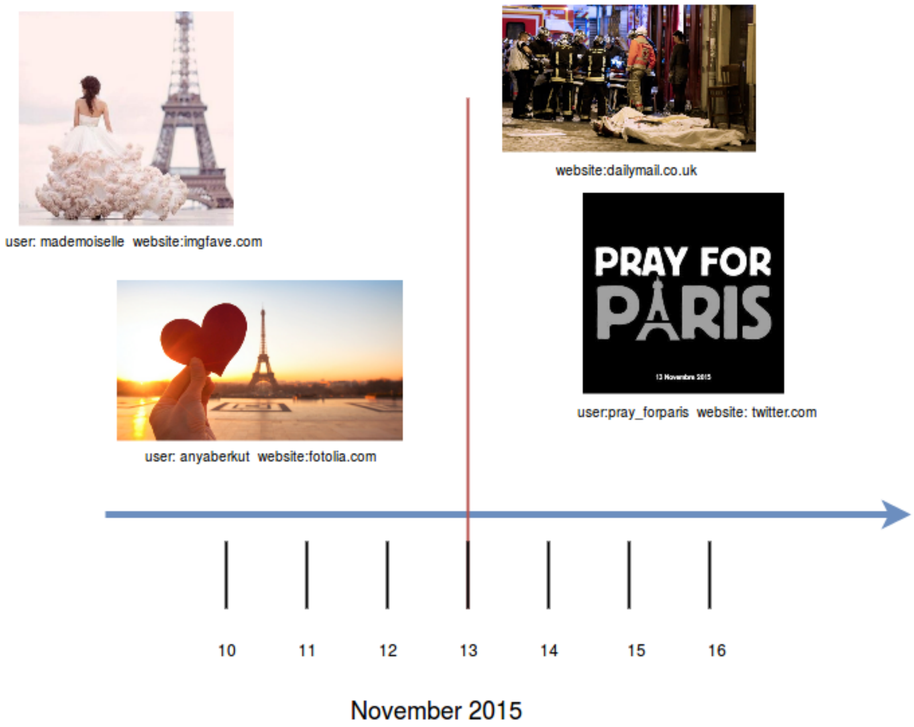

user: mademoiselle website:imgfave.com
website:dailymail.co.uk
PRAY FOR PARIS
13 Novembre 2015
user:pray_forparis website: twitter.com
user: anyaberkut website:fotolia.com
10
11
12
13
14
15
16
November 2015

Figure 1.1: Sentiment change in the word "Paris"

To assess these hypotheses, our proposal is to develop a system that automatically extracts terms which are domain and time specific. Next, using Twitter, we will assign a sentiment to those terms to build domain and time specific lexicons. In order to evaluate if our system assigns these terms the correct domain and sentiment, we will test it using volunteers. Next, we will use the resulting lexicons with sentiment analysis methods, to evaluate if they perform better in the presence of the dictionaries created by our system.

1.4 Thesis Structure

The thesis structure is the following: in the next section we will provide an overview on the state of the art of sentiment analysis. After, we will detail the workflow and implementation of our system. In Chapter 4 we describe the evaluation made to assess the quality of our generated sentiment lexicons. Next, we refer to the changes made in the system (resulting from the evaluation made). Finally, we will evaluate if the lexicons created improve the sentiment analysis task in Chapter 6 and draw some conclusions and future work in Chapter 7.

Chapter 2

State of the Art

2.1 Sentiment Analysis Overview

Sentiment Analysis and Opinion Mining interest has significantly grow in the last decade. The expansion of Social Web combined with the expression of opinion by its users on several different domains, has motivated the research on these topics. In fact, a query by year on Google Scholar[1] reveals that the number of studies published in Sentiment Analysis and Opinion Mining has increased with time. In Figure 2.1 we present a plot with the number of hits in Google Scholar using "Sentiment Analysis" and "Opinion Mining" as keywords and filtering the result by publication year.

It is important to notice that there is a significant number of publications that are included in both topics since they are, generally, used interchangeably. However, the plot intends to demonstrate that the research in sentiment analysis/opinion mining continues to grow and that it is an active research topic.

In order to comprehend the meaning of sentiment analysis, we first need to define the problem. Let us consider the following review extracted from IMDB[2]:

> *"Brilliant adaptation of the story of Bletchley Park and the cryptanalysis team, ran by Alan Turing, that cracked the code of the German Enigma Machine during World War II. Featuring an outstanding starring perfor-*

[1] `https://scholar.google.pt/`

[2] `http://www.imdb.com/title/tt02084970/`

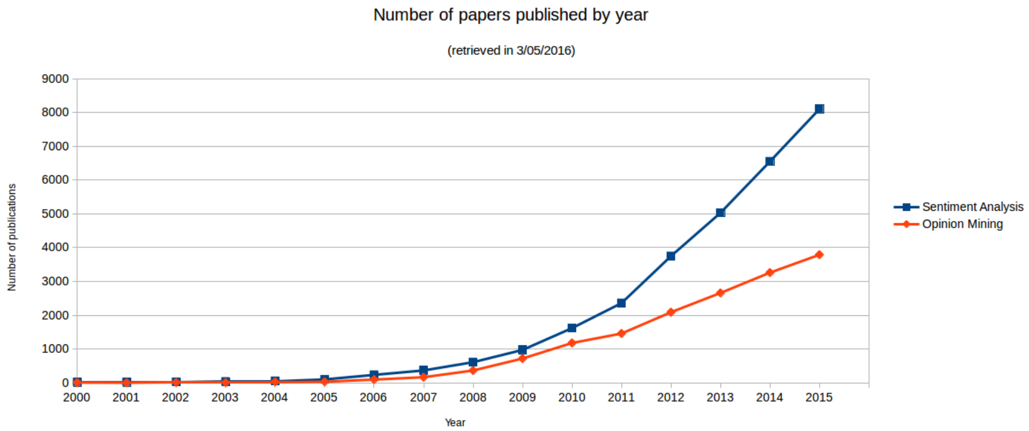


Figure 2.1: Number of papers published by year in Sentiment Analysis and Opinion Mining

> *mance from Benedict Cumberbatch as war hero Turning and supporting acts from a brilliant cast including Keira Knightley, Charles Dance and Mark Strong, 'The Imitation Game' is a powerful and eminently well-made biopic that illuminates the facts whilst respecting the story it is based upon. The English-language debut of 'Headhunters' director Morten Tyldum, this British World War II thriller is a highly conventional story about humanity that creates a fascinating character, anchored by a hypnotically complex performance."*

Sentiment analysis methods evaluate the sentiment polarity in fragments of text. In the example above, strong evidence of positive sentiment is hand over by the presence of words such as *brilliant*, *outstanding* and *fascinating*. Studies to determine the polarity of reviews have been frequently published. For example Turney [119] proposes assessing the sentiment in reviews using an approach based on the extraction of opinion words. The method consisted in extracting pairs containing adverbs or adjectives (normally used in sentences which express opinion) and estimate the semantic orientation of those pairs. For that purpose, an adaptation of the point-wise mutual information algorithm for measuring the distance of the pairs to the word "poor" and "excellent". This distance is achieved using the operator NEAR in queries on the Altavista search engine. However, depending on the text, not always the detection of polarity words is enough. Let us consider the following review from Amazon:

> *"The original Star Wars trilogy was a defining part of my childhood. Born as I was in 1971, I was just the right age to fall headlong into this amazing new world Lucas created. I was one of those kids that showed up early at toy stores [...] anxiously awaiting each subsequent installment of the series. I'm so glad that by my late 20s, the old thrill had faded, or else I would have been EXTREMELY upset over Episode I: The Phantom Menace ... perhaps the biggest let-down in film history."*

Although there are a significant higher number of positive opinion words, as a whole the review is negative. To address this issue, Taboada et al [106] analyse the sentiment of the text considering also the position of the opinion words.

Another way to tackle this issue is to identify the opinion or sentiment target. In other words, to which entity (movie, person, city...) is the sentiment referring to since sometimes different sentences may express sentiment on different topics. For instance, in a laptop review, we might want more details about what the user thinks about a specific aspect or feature (battery, storage, memory...). In this case, we can use sentiment analysis at an aspect level to have a more detailed opinion on certain component. Yohan et al. [59] propose an unsupervised model (ASUM) to discover pairs of (feature, sentiment) in reviews. Another study [115] uses specific techniques to perform sentiment analysis in movie reviews. The authors start by using a database of movie titles to differentiate sentiment words present in the titles from the ones present in the review. Then, analysis at a clause level is performed and positive or negative scores are assigned to words, having in consideration the dependency between them. Several other works [58, 20, 72] have studied methods for extracting sentiment words regarding specific entities.

Sentiment analysis can be performed at other levels besides the one mentioned previously. In fact, there are three distinct levels:

- Feature based/ Aspect level
- Sentence level
- Document level

Document level sentiment analysis refers to the task of identifying the overall sentiment in a document. Studies like [85, 119] use sentiment analysis in a document level to classify reviews based on its polarity. Another approach is sentence level sentiment analysis. Meena et al. [70] use linguistic analysis like sentence construction and conjunctions to infer the sentiment in the sentence. In another study [130], the authors try to differentiate the general polarity of a word from the polarity of the same word in a certain sentence context (contextual polarity).

Not only is important to identify the target of the opinion, but also in some contexts the source (or opinion holder). For example in a news domain, where the opinion holder can be a group, person or enterprise (e.g. "the president is pleased with the work done"). The authors in [61, 31] provide techniques for opinion holder extraction in online news texts.

2.1.1 Social Network Sentiment Analysis

Although early sentiment analysis studies are, in their majority, focus on reviews, with the grow of social networks, Twitter also became a main target source for sentiment data. The particular characteristics of this social network where users share their opinion on several different topics through short informal posts, has motivated sentiment analysis researchers to analyse and evaluate texts from it. Several works [80, 74, 46, 42] developed sentiment lexicons based on Twitter whereas in [112, 55, 95] the authors have used this social network data for evaluating sentiment analysis methods. Other approaches have used Twitter to track the sentiment on different topics such as elections [126, 118, 17] and stock market [13].

Fewer studies have been conducted on other well known social networks or websites. In [101] Youtube video comments are analysed in order to find relations between sentiment words in the comment and the rating that the same achieved. In the same study, the comment ratings are also used to evaluate whether a video has a polarity score associated to it or not. Furthermore, the study concludes that ratings are not equally distributed through different categories and that some categories (like music) have more positive ratings than others (such as vehicles, gaming and science).

Some studies have also used Facebook comments and status updates as source for sentiment analysis. The authors in [66] analyse sentiment in the comments of a Facebook fan page. They also try to retrieve the most important topic using keyword extraction with: term frequency, TD-IDF (measures the importance of a word in a set of documents), and Positive, Neutral, and Negative words. Another study tries to classify the different posts into three distinct categories: entertainment, posts from pages the user liked, and life events (labelling this last one according to the sentiment) [100].

2.1.2 Sentiment Analysis Features

Important features have been considered for better performance in sentiment analysis. Part of speech (POS) tags such as the Stanford Log-linear Part-Of-Speech Tagger [116] and OpenNLP[3] have been used to identify what words are nouns, verbs and adjectives with the goal of determining if a certain text is subjective and consequently classify it with sentiment. Other textual features include the use of valence shifters [88]. Valence shifters are words that reverse, amplify or attenuate the strength on sentiment words. For example:

- Negation: *"the food was delicious"* → *"the food was **not** delicious"*
- Amplification: *"the movie was bad"* → *"the movie was **very** bad"*
- Attenuation: *"the product was better than I expected"* → *"the product was **slightly** better than I expected"*
- Negation and Amplification: *"this album is good"* → *"this album is **not very** good"*
- Negation and Attenuation: *"the results are not bad ..."* → *"the results are **not only** bad ...."*

We can notice by the examples provided that negation combined with an amplification or attenuation reverses the polarity effect on the sentiment word. In other words, amplification words with negation become attenuation words and vice-versa.

[3] https://opennlp.apache.org/

Besides textual features, in a web context other indicators can be extracted to support sentiment classification. Emoticons, for example, have also been pointed out as major, non textual, intensifiers of sentiment. At this point, it is important to distinguish three concepts that are, most of the time, used interchangeably. Emoticons or smileys refer to a representation of a face using punctuation symbols whether emojis are pictures that can represent anything from a happy face to an air-plane or car [49]. Table 2.1 provides a sample of some emoticons and emojis.

Emoticons	:)	:(	:D	:'(
Emojis	😃	🎩	😡	💁‍♀️

Table 2.1: Example of emoticons and emojis

This feature can be particularly helpful in short length texts. If we consider the tweet:

> *"In the Lock Tavern waiting to see Andrew Bird. Last time I was here b4 a @RoundhouseLDN gig we were seeing Prince :("*

we can observe that the emoticon *":("* is crucial for sentiment detection since none of the traditional opinion words occurs.

To address emoticons as sentiment features, in [52] an emoticon lexicon is created and classified with sentiment by 3 annotators. Then, results of combining this particular lexicon with text-based sentiment analysis yield in an accuracy increase of 37% when compared with a text-based only analysis. Another work [92] infers that emoticons can be used in a semi-supervised method to assign polarity labels to text. Consequently, these texts are used as training data for supervised sentiment classifiers instead of the more traditional way of manual labelling by volunteers. Other studies [67, 57] also use emoticons as relevant feature in sentiment classification.

Some works have also studied emojis as sentiment cues. In [82] the authors assigned a sentiment classification to each emoji based on the polarity of the tweets where they are present. A more recent work [71] studies the perception of the same emoji in different platforms (since they are represented by different figures) in terms of sentiment and semantic.

The works on emoticons use, in it's majority, data from Twitter or Twitter-like social networks (e.g. Weibo[4]). Another specific characteristic from this social network that has been studied as a sentiment feature are hashtags. Works [127, 26] consider hashtags as features for sentiment classification.

2.1.3 Supervised and Unsupervised Methods

The features mentioned previously can be used in both supervised and unsupervised methods. The main difference lies on the presence/absence of labelled data for learning. For instance, if we want to predict a variable y with a set of features x_1, x_2, x_3, supervised learning uses classified entries to learn and predict the outcome in future test cases. Some classifier examples are Support Vector Machines [21], Naïve-Bayes [98], Max-Entropy [81] and Decision Trees [90]. On other hand, unsupervised learning does not require the labelled data making it a low resource method. One of the most common example is the k-means clustering [68].

One of the first approaches using supervised learning in sentiment analysis classification was the study in [86]. The authors proposed using supervised techniques, that prove to be effective in text categorisation, to classify a text with positive or negative sentiment instead of a topic like sport or politics. The authors used three algorithms: Naive-Bayes, Maximum Entropy and SVM's. Results showed that SVM's performs better and Naive-Bayes the worst in a movie reviews dataset. As a matter of fact, Pang et al. also prove how the presence and frequency of unigrams and bigrams can influence the accuracy on polarity classification of reviews. Results show that unigram presence is the more effective for learning than bigram presence or unigram and bigram frequency. However, in other work [26] the bigrams and trigrams achieved better performance. Works like [77, 42, 132, 129] also used supervised algorithms, however some research has also focus on alternatives approaches such as Dynamic Artificial Neural Networks [40] or even, more recently, a Deep Learning approach using a Recursive Neural Tensor Network and a sentiment tree bank which achieves 85.4% accuracy in short phrases sentiment classification [103].

In unsupervised methods, most of the work on sentiment analysis has been done using dictionary-based approaches. The general idea is to use list of opinion words such as

[4] http://www.weibo.com

"good", "bad" and "awful" rated with a sentiment and then match the words present in the text with the ones present in the list. It is important to mention that most authors agree that supervised methods in sentiment analysis are the ones which require labelled texts for a learning phase. Although sentiment lexicons can be considered labelled information, they are normally classify as semi-supervised or unsupervised methods. In this work, we will use the latter to categorise these approaches.

A simple sentiment classification can be done by summing all opinion words sentiment values and divide by the number of opinion words presented. This method can be complemented with the features mentioned before. For example, lists of amplification, attenuation and negation words can be included for increasing, decreasing and reversing the level of polarity respectively in a rule based form (i.e. amplification/attenuation words increase/decrease 0.3 the sentiment word, negation reverses it and so on). Emoticon and emojis sentiment dictionaries can also be integrated.

Systems have been developed for unsupervised sentiment analysis. For instance, SentiStrength unsupervised approach [114, 113, 111] not only combines the features mentioned earlier but also assigns different sentiment for words with repeated letters ("love" and "looooooove"), uppercase words ("this is bad", "this is BAD") or repeated punctuation ("awesome!", "awesome!!!!!"). It also includes a spelling correction procedure due to common errors in social web messages. Vader [55] also follows the same rule based approach which includes giving different sentiment to text with uppercase words, repeated punctuation and "emoticon extension" (i.e. from *:)* to *:))))))))*.

It is also worth mentioning that there are several works that implemented ensemble sentiment analysis systems by combining different methods (supervised and unsupervised) into one. For example the work in [6] uses an ensemble classifier based on 9 different lexicon based approaches to evaluate reviews from different domains. The assessed sentiment on each review is done by each method by simply counting how many positive words occur in the review from that lexicon and subtracting by the number of negative words. The authors prove that this ensemble system is faster and have similar or better accuracy than supervised approaches, using six different reviews datasets each one with a specific domain. In another work [128] the authors proved that ensemble learning algorithms like boosting [38], random subspace [50], and bagging [18] are more accurate in sentiment classification than a single SVM model. Similar conclusion are reached in [125] where an ensemble system by majority vote is built and compared with a lexicon based and several supervised methods in a dataset of tweets of airline companies. Results achieved provide evidence that assemble methods perform better than other individual methods in both Negative/Positive and Negative/Neutral/Positive classification.

Finally, although there is a large number of published works on sentiment analysis, they are normally evaluated on different datasets. Therefore, a direct comparison between methods is hard to achieve. As the authors in [117] refer, there is a need to aggregate the different sentiment methods into a framework in order to directly compare them across different text datasets. With this goal in mind, the authors in [1] make an exhaustive comparison of 20 different twitter sentiment analysis tools on five different domain datasets. They use metrics like accuracy and F1-score in each sentiment class as well as in each different domain. Another work [44] provides a detailed sentiment system comparison with 24 different systems (including paid sentiment analysis systems) in 18 different datasets, where some of them were used as a gold standard in previous publications. In addition, they provided a system (IFeel[5]) which can be used to replicate the results in the non-paid sentiment analysis methods.

2.1.4 Non-English Sentiment Analysis

Whereas the majority of studies on sentiment analysis are based on methods applied on data in the English language, some work has also been done in other vocabularies. For example, as Lee et al. [63] explain, sentences in Chinese are formed by a continuous stream of Chinese characters. Each character has a specific meaning but, when combined with other characters, that meaning may differ. In cases like this, simple sentiment analysis approaches (like Bag of Words) without taking this segmentation property into account may lead to inaccurate results. Due to the particularities of the Chinese language, besides segmentation treatment, the same study uses other steps (like conjunction rules and negation handling) combined with Max-Entropy to achieve high accuracy in Chinese online reviews. In [107] a comparative study is made using different feature selection methods and different learning models for the Chinese language. The results showed that the combination of using Information

[5] `http://blackbird.dcc.ufmg.br:1210`

Gain (a feature selection procedure that uses the changes on information entropy to determine the best features) to decide which set of features behave the better and SVM's as the sentiment classification model, perform the better amongst all other. Other works [94, 124] present lexical resources to perform sentiment analysis in the German language, while [102] presents a sentiment analysis system for user based political content (such as blog posts or news) in Portuguese vocabulary.

2.2 Domain and Time Specific Sentiment Analysis

In most of the cases, sentiment analysis is domain specific. However, time can also play an important role (for example when we analyse tweet data).

The authors in [92] state that, in the specific case of supervised techniques, exists a domain and time dependency on the models. A good example is the sentence "Go read the book" that has positive sentiment if the domain is book reviews but negative sentiment if it is movie reviews [85]. Adjective like "unpredictable" can also have opposite sentiment polarities on different domains. More specifically, the expressions "the plot was unpredictable" in a movie review and "the steering was unpredictable" in a car one [119]. In fact, several studies on sentiment analysis use datasets from different domains ("movies", "cars", "travel") to assess the performance of their proposal, since the results vary from domain to domain [111, 112].

Moreover, traditional data mining approaches make the assumption that the training data and test data are equally distributed and have the same feature space, something that may not be correct on the case of sentiment analysis on different domains. Transfer learning may help to tackle this problem by training a model based on a data from a domain and test it in several others [84]. The study in [133] classifies words regarding three attributes: polarity, domain and domain dependency/independence. For words that are domain independent only the polarity is identified. However, for domain dependence vocabulary, a polarity classification is assigned to each domain. Another study [41] introduces a deep learning approach for domain adaptation. Results show that domain adaptation was successfully achieved in 22 different domains using a two step procedure (consisting of an unsupervised feature extraction method in all domains and a SVM for sentiment classification).

As for time specific sentiment analysis, the majority of publications have focus on sentiment topic tracking through Twitter. This research subject studies how opinion varies on specific topics using Twitter for public opinion assessment. Approaches for sentiment topic tracking on Twitter generally use a term or set of similar terms as search query for retrieving tweets related to the topic. Then, sentiment analysis procedures are performed on the sample. In topic tracking systems, the output is normally displayed in a more graphical style (such as plots or histograms) for a quicker reasoning on results [27] and the evaluation is normally done by correlating real time events to sentiment changes. Several studies present sentiment topic tracking monitoring on different domains [126, 10, 56, 62, 3].

A peculiar controversy on sentiment topic tracking it's the capability of Twitter on predicting elections outcomes. While [8] defend that Twitter data can be a good predictor on election results, the works [17] refute that hypothesis. A more specific example is the study conducted in [118] where tweets are retrieved prior to the 2009 german election. The study concluded that Twitter can be used for real time analysis of political sentiment suggesting a predictive power close to polls. However, a later work using the U.S. 2011 congress election stated that no relation has been found between the predictions and elections results [39].

2.3 Sentiment Lexicons and Expansion Methods

One of the most important parts for achieving high accuracy on sentiment analysis is the "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.

Some sentiment lexicons were already discussed previously. However, they were not properly categorised. There are three main groups where sentiment lexicons creation or expansion methods can be included: manual labelling, dictionary based and corpus based.

2.3.1 Manual Labelling

This approach consists in the labelling of a list of words with sentiment by one or several volunteers/workers. Then, using metrics to determine inter-worker agreement, is established a ground truth for the sentiment of each opinion word [75, 105, 55, 79].

The lexicon used in Vader [55], for example, was created by concatenating several list of words from previous state of the art sentiment lexicons, and emoticons and acronyms. Then, using Amazon's Mechanical Turk [2], each word was classified with a sentiment value (ranging from -4 to 4) by ten different workers. In addition, four quality control tests were implemented with monetary rewards to the workers who performed better. This careful method on sentiment lexicon construction complemented with a rule based system makes Vader outperform several state of the art sentiment analysis methods.

Another example is the AFINN lexicon [80] which was built specifically for microblogs. The author began by adding a set of obscene words and common positive terms. Then, manual analysing a set of tweets, the lexicon was extended with other sentiment words and common slang used online. The resulting sentiment dictionary surpassed the ANEW on tweet sentiment analysis.

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 [2] or CrowdFlower [23] since a fee must be paied to each worker who completes the classification task.

2.3.2 Lexicon Expansion Methods

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 having as a base these words. Two different approaches have been used for expanding the lexicon in semi-supervised fashion[6]: thesaurus-based approaches and corpus-based approaches.

[6] To avoid confusion, here we use semi-supervised to classify the lexicon expansion method not the sentiment analysis itself.

2.3.2.1 Thesaurus-Based

Thesaurus-based approaches rely on other syntactic resources like the General Inquirer (GI) [104] or WordNet [35]. WordNet is a large lexical resource containing nouns, 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 to antonyms [7, 32]. However, the authors in [73] present better sentiment accuracy in words than SentiWordNet1.0 by using a Roget[97] like thesaurus, which is a tree schema dictionary with over a thousand branches and whose leaf nodes are cluster words aggregated by meaning.

Studies [60, 53] also used WordNet to expand sentiment lexicon making it one of the most used resources for lexicon expansion.

2.3.2.2 Corpus-Based

One of the major problems on 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.

Let us focus our attention in the following two reviews, extracted from a television and video game review, respectively.

- *The Samsung remote is awesome and **easy** to use.*
- *The game has beautiful graphics but **easy** to complete.*

Here the word "easy" has different sentiment polarity depending on the domain. Whether in television reviews points out to a positive sentiment, in video game reviews is associated with negative. For this type of problem, corpus based lexicons are more viable solutions.

In [47] a corpus based lexicon expansion method is proposed using conjunction rules to infer new opinion words specific to the domain. Using the television review above, 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, 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 lexicon expansion is presented in [89]. It uses a set of seed words combined with conjunction rules for extracting entities and opinion terms. Then, through an iterative process, the new pairs of entities/opinion words are used for finding more pairs; the algorithm 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 [53]).

However, opinion words polarity may vary, 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 [28].

Besides reviews and similar to sentiment analysis classifiers, 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 [15] 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 [16] presents a supervised algorithm for lexicon expansion using tweets label with emoticons and a combination of several seed word lexicons. Another supervised approach [108] uses SkipGram (for learning continuous phrases representation) and a seed lexicon (expanded with contents from the Urban Dictionary[7]) as training data for a sentiment lexicon expansion classifier. One more study [30] 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 [36]. Whereas most of the methods presented focus on expanding sentiment lexicons with adjectives and verbs, the Feng et al. [36] study the influence of words with connotative polarity such as *cancer*, *promotion* and *tragedy.*

[7] http://www.urbandictionary.com/

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. Consider the following extracted sentences from different reviews:

- Mattress Review: *"Within a month, a* ***valley*** *formed in the middle of the mattress"*
- Tablet Review: *"It came with a* ***scratch*** *in the screen"*
- Hotel Review: *"The bedroom walls had a lot of* ***stains****"*

The authors in [134] 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 thesis it is the exploration of assigning sentiment to connotative words, nouns that imply opinion, entities and topics that will be highlighted. We intend to expand even more the sentiment lexicons in this studies by using public opinion as a measure of polarity for domain and time specific terms, combining topic sentiment tracking and lexicon expansion methods to study if the dictionaries created can improve state of the art sentiment classifiers.

Chapter 3

The Proposal

3.1 Problem

Approaches on lexicon expansion are widely based on the exploration of sentiment in adjectives or verbs. This is due to the fact that they are the most significant cues on detecting if a text infers sentiment or not. Tweets like "We are winning!" or "@Nike these new tennis are beautiful", are associated to a positive sentiment from the author of those tweets. However, as mentioned before, not only verbs and adjectives are important for sentiment detection. Sometimes, other terms can also be an important factor. Consider the following sentences.

- "First Paris now Brussels... What can we do?"
- "Listening to Bowie. Still can believe it"
- "I'm not travelling in Egyptair the next few years!!!"

Although there are some subjectivity cues (such as repeated exclamation marks or the presence of suspension points), analysing the sentiment polarity based only on syntax is a difficult task, mostly due do the absence of verbs and adjectives that can infer positive or negative opinion. However, our capability to understand the sentiment on these sentences relies on the knowledge from the nouns, topics or entities we acquire on a daily basis. For example, we are aware of the negative sentiment on the first and third sentence due to the tragic events that occur on Paris and Brussels and the crash of the Egyptair plane [1]. The second sentence also points out to a negative sentiment appealing to a fact that is common knowledge for most of the people (the death of David Bowie [2]) without mention it.

Another difficulty in sentiment analysis is the detection of irony. Irony is when there is a difference between the literal meaning of the words and the truth meaning of the text. Irony can be a hard obstacle to sentiment analysis methods. Some examples follow:

- "Donald Trump will be such a great president! Hitler would be proud!"
- "'Your cheering for Leicester? Such a weak team.' I guess they were right though. We barely stayed on Premier League "

In the first set of examples, the detection of sentiment can be hard using traditional sentiment lexicons (even the domain specific ones) due to the absence of verbs or adjectives that usually infer sentiment. On the second set, although there are some adjectives such as "proud" on the first sentence and "weak" on the second, traditional approaches would classify the sentences as positive and negative respectively. However, once again, human perception can determine that there is irony present and, in fact, the polarity of each sentence is contrary to what individual adjectives point to. This is mostly because we are aware of the events occurring in the world and the opinion that such events generate in the public (i.e. public opinion). In this particular case, we are informed about Hitler as a negative word for the general public. Therefore, we can determine that although there are positive words in the text, irony is present and consequently the general sentiment of the sentence is negative. The same occurs in the second example where the presence of irony is detected due to the fact that the Leicester football team has won the Premier League [3] and therefore is associated to a positive sentiment .

To address the problem of lack of public opinion and general knowledge on sentiment lexicons, we must retrieve a sample of what that public opinion would be on each

[1] http://www.telegraph.co.uk/news/2016/05/19/egyptair-flight-ms804-disappears-what-we-know-so-far/

[2] http://www.bbc.com/news/entertainment-arts-35278872

[3] https://www.theguardian.com/football/2016/may/02/leicester-city-win-the-premier-league-title-after-fairytale-season

term in a particular time interval. We define term as a word or set of words which may be connotative words such as "cancer", "terrorism" and "tumor" or that are usually associated to a sentiment which globally and in a certain period of time, tends to a specific polarity ("Donald Trump", "Brexit", "Isis"). Assessing the sentiment on a specific term within a specific time automatically can be a difficult task. Let us consider the example when politician X, which public opinion is fairly divided (some people think s/he is a good politician and others the opposite) is caught in a money laundry scheme. It's fair to assume that the weight of public negative sentiment towards that politician will increase significantly. However, let us consider that some progress has been made on finding the cure of disease Y. Although this can attenuate the negative sentiment regarding the term Y this fact by itself should not be strong enough to change the polarity of the term. It is also plausible that most of public opinion sentiment changes emerge from news whether they are posted online, through television, or even in newspapers. For example a user reads the headlines on a news site and forms a personal opinion on certain terms (topics, events or entities) involved. If most of the public shares the same opinion, then there is a tendency on negative or positive sentiment. The assessing of those terms and sentiment associated can be important to improve sentiment classification, specially in short informal texts. However, for that effect, the terms and sentiment retrieved have to be constantly updated and aware of the events which may change public opinion on certain terms.

In summary, there are two important issues which sentiment analysis methods do not contemplate:

- The lack of sentiment words does not always means the lack of sentiment
- The miss classification of sentiment due to irony can be improved if we assess the public opinion on some domain and (specially) time specific terms.

3.2 Conceptual Design and Solution

In order to create a lexicon expansion method that includes nouns, connotative words and other words or terms whose sentiment may fluctuate, we must: 1) identify what terms are relevant at the time and are subjected to sentiment change and 2) assess the public opinion on those specific terms.

From the different lexicon expansion methods discussed in the previous chapter (manual, thesaurus/dictionary based and corpus based), the corpus based approach seems the more suitable. The manual approach is quickly discarded due to the time variable which would require an exhaustive and constantly labelling task from annotators. The dictionary based approach is also not appropriate because we are not only using verbs or adjectives but nouns which do not contain synonyms or antonyms. Therefore, the corpus based approach seems the more reasonable method.

With the lexicon expansion method selected, we must now determine what corpus or set of corpus we will use. Unlike other approaches where an unlabelled corpus is provided for detecting opinion words [15, 16] our approach will focus on detecting the more relevant terms and extracting a corpus based on public opinion for each of the terms. Then, our assumption is that the sentiment assessed in that corpus will be the general sentiment of the term.

In order to test that hypothesis and since we want to extract the more relevant and up to date terms, we will resort to news sources, more specifically the news headlines. As a matter of fact, headlines are the first part of the news scanned by readers and therefore, the decision of whether they will read or not the article relies on its relevance. For example, a study was conducted in [29] where the authors investigate how headlines relevance is important to capture readers attention. Since headlines summarise the information on the news, they should include the more relevant terms to hold the attention of the reader. So, searching relevant terms on news headlines should provide the more relevant ones without noisy words, which may be present in news body. Furthermore, since news are normally organised in different sections (world, sport, entertainment), we can create a domain specific sentiment lexicon by using ternary entries composed by <term,domain,sentiment>.

Since we are going to use headlines to determine the more relevant terms, we could apply sentiment analysis procedures on the same news corpus to determine the polarity of the term. However, headlines (and news content) tend to report facts which consequently lead to an absence of opinion. Even if there is sentiment expressed, it may not be the same as public opinion. As an example, if we consider the headline *"Afghanistan intelligence agency confirms death of terrorist leader"* [4], the presented sentiment is

[4] http://www.foxnews.com/world/2016/05/22/afghan-taliban-leader-mullah-mansour-likely-killed-in-us-airstrike-official-says.html

clearly negative (due to the presence of the "death" word) but, public opinion may not share the same polarity.

Therefore, to assess the public opinion on a term, we rely on Twitter and its API. Our approach has solid foundations on several studies who conclude that Twitter is good for assessing public opinion on a certain topic [126, 10, 56, 62, 3].

3.3 System Proposal

In this section, we describe the workflow of the proposed method. The system is divided in two main components: the term extraction procedure and the sentiment evaluation method.

To the best of our knowledge, no lexicon expansion method uses an automatic way to extract and infer sentiment analysis on a term based on tweets retrieved on that term. Our main assumption is that the polarity of a sample of tweets on that term represents the public opinion of the term and therefore, that must be evaluated before assessing how the terms from our lexicon expansion method influence dictionary based sentiment analysis. That evaluation is explained in the next chapter.

3.3.1 Term Extraction

The first component of our system is the term extraction procedure. Since we want to extract the most relevant terms at the present time, we use news headlines as our *corpus*. For that purpose, we use RSS feeds from news sources for our extractions. RSS (or Rich Site Summary) is a specific format for dynamic web content. It is often used in media websites and allows an aggregation of content from different sources through a RSS reader [12]. Although RSS seems an outdated technology, the fact is that Feedly, which is one of the most used RSS reader for Android has over 1.5 million downloads [34] and reached 60.000 of paid subscribers in 2015 [33]. Furthermore, Google News RSS feeds extension for Chrome browser has over 1.2 million users [69].

Several news sites still maintain RSS for their news feeds. In addition, the RSS feed, like the news present in website, are categorised by news sections, providing to the extracted terms a specific domain. In our approach, we select several major news sites as news sources. We also restrain our research to seven specific news domains: world, entertainment, technology, sports, business, health, and politics. This selection was based on the majority of the domains found for RSS news sources (e.g. we would like to include the domain science but the number of sources found was relatively small) and the wide scope of the domains. For example we could use global markets news feeds but achieving a public opinion on such narrow scope could be a hard task due to the specificity of the terms.

We only used news sources in the English language and whose origin countries are the United States or included in the United Kingdom. We restrain our research to these two location specific news (and consequently location specific terms). In domains like politics, it is fair to assume that in the U.S. terms like the names of presidential candidates (e.g. "Bernie Sanders", "Donald Trump", "Ted Cruz"), "White House", "GOP" or "Obama" are likely to appear. However, in the United Kingdom it should be expected more terms like "David Cameron" or "Prince Charles". This geographical dependency inside domain specific terms should be an interesting topic to explore in future research, but for now we focus on the aggregation of these two due to their influential status. In fact, a survey puts the US and UK as the two most influential countries in the world according to several different factors [48]. Therefore, we argue that international media coverage is bigger in these countries and consequently, public opinion data should also be vast and easier to acquire using terms from these geographical sources.

Table 3.1 shows the sources on each of the selected domains. We left out Google News RSS feed since their redirect their news from other sources and, in order to achieve at least nine sources on each domain we were forced to include a domain specific news site (MedicineNet).

Our term extraction procedure begins with the retrieval of news headlines from the different sources for each specific domain. It is important to mention that the number of headlines on each source is dynamic. Then, headlines are aggregated and a corpus is constructed for each domain.

With the constructed corpus, some filters were made to avoid "noisy" terms. We will refer to noisy terms as terms who are not relevant to be classified with sentiment in

Table 3.1: News Sources by domain

News Source	**Domain**						
	World	Entertainment	Technology	Sports	Business	Health	Politics
CNN	x	x	x	x	x		
BBC	x	x	x	x	x	x	x
The Economist	x	x	x		x		
The Wall Street Journal	x		x		x		
ABC News	x	x	x	x	x	x	x
CBS News	x	x	x		x	x	x
The Washington Post	x	x		x	x		x
NBC	x	x	x	x			x
The Guardian	x	x	x	x	x		
Reuters	x	x	x	x	x	x	x
Yahoo News	x	x	x	x	x	x	x
Sky News	x	x	x		x		x
Daily Mail	x	x		x	x	x	
The New York Times		x	x	x	x	x	x
Financial Times			x			x	
Forbes			x		x		
MedicineNet						x	
Total	**13**	**13**	**14**	**10**	**14**	**9**	**9**

our lexicon. First, punctuation is removed from the corpus and lower case is imposed to every word. Additionally, stop words (e.g. "a", "at" and "this") are also removed because they occur with relatively high frequency and are not relevant to be included in our sentiment assignment task.

The next step is to build three different term-document matrices. One for unigrams (i.e. individual words) other for bigrams (i.e. two words terms) and the last one for trigrams (i.e. three word terms). Through experimentation we realise that most of the times, terms above trigrams were unique (in other words, they only occur once) so we discarded them. Furthermore, it is important to establish a limit for each set of terms. One way is to define a minimum number of occurrences of the term depending on the number of sources in that particular domain. Therefore, this threshold must be a percentage value of the number of sources so that 1) we can easily adapt and scale the system to more domains and sources 2) the number of terms extracted is restrained avoiding possible noisy terms which are not filtered. The minimum number of terms required is determined by the following formula:

$$\text{threshold}_i = n_{\text{sources}} * a_i$$

where i represents the n-gram considered and

$$a_1 = 0.50$$

$$a_2 = 0.30$$

$$a_3 = 0.25$$

These values were also obtained by experimentation and considering an estimated time that each term would take to be classified with sentiment (since we wanted to retrieve daily dictionaries, a large number of terms could be hard to process).

Let us consider the sports domain which contains 10 sources. Therefore, the number of occurrences that an unigram term must occur in the sport *corpus* to be included must be 5. For bigrams is 3 and trigrams also 3 (since it's rounded to the unit).

At the end of this process, there are 3 term lists which will pass through another set of filters, before assessing the sentiment. The unigram list is tagged with a POS-Tagger (OpenNLP). This allows to identify what parts of speech (such as verbs, personal prenouns, and adjectives) are present in our list. Since we are targeting connotative terms and terms which may have a polarity associated in a given time interval, in our lexicon expansion method the words classified as nouns and foreign words by the POS-Tagger are kept. Adjectives were also maintained since: 1) sometimes nouns and adjectives were misclassified by the POS-Tagger and therefore we could be excluding some potential terms and 2) since we also have a filter to exclude sentiment words (which are mainly adjectives) their presence would be very reduced and would not severely impact the lexicon. We do not wish to include verbs due to the same reasons we do not wanted to include adjectives. However, due to the different forms that a verb may have, it would be difficult to exclude them all based only in our sentiment dictionary. In addition, like adjectives, the word alone can have a different sentiment than when included in a specific context. For example, "win/wins/winning/won" or "lost/loose/loses" are generally associated with a positive and negative sentiment, respectively. Therefore, we used the POS-Tagger to exclude words that were labbeled as verbs. Nonetheless, if they appear within a context ("Hillary wins"/"Sanders lost"), then the public sentiment of the whole may not be the same as the sentiment of the term alone. The next filter applied removes words that are already classified with sentiment in the AFINN lexicon[79]. We decided to use this particular lexicon in the next stage of our system where we classify the tweets due to the presence of not only common sentiment words but also because it includes Internet slang. Therefore, we also use it to avoid classify repeated terms. In addition, words that are repeated in plural form ("syrian/syrians") and apostrophe form ("Trump/Trump's") are kept solely in singular and non apostrophe form. Moreover, we also excluded common words used in each specific domain. For that purpose, Oxford Learners's Topic Dictionaries[5] were used . It consists in groups of words commonly used in a specific domain. The lists for our specific domains were extracted and manually analysed in order to prevent that some words which it would be interesting to classify, were not excluded from our term extraction procedure. For example words like "Apple" and "Microsoft" were withdrawn from the technology topic list , because they refer to companies that, in a specific time period, may have a negative or positive public opinion (and therefore are worthy to include in our lexicon). Likewise, common terms relative to news domain (e.g "news" "review", "tech", "week", "watch", "weekly podcast") and news entities (e.g. "ABC News" and "NBC Sports") were also excluded. This term list was constructed manually and terms were discovered by experimentation based on daily retrieval of words and frequency assessment.

The POS-Tagger filter is only applied to unigrams where the other filters are applied to all term lists. This is mainly because 2 or 3 word terms already have a context since they occur several times in different headlines from different sources. Even if they do not, we create a last filter to prevent that meaningless terms end in the sentiment lexicon. This last filter relies on the beginning of our sentiment analysis procedure. As it was mentioned before, we will use Twitter to assess the public opinion on a certain term. With that purpose, we will define a search query with the term extracted and require a specific number of tweets. In order to filter terms that are irrelevant, we define a 33% threshold on that sample meaning that if the tweets retrieved do not reach that threshold, then the term is discarded. Our assumption is that if tweets are not found on that specific term then it is because the term is irrelevant or does not contain a specific meaning. Some terms like "deal will approved", "receive time" or "floyd mayweather got" are not syntactically correct or meaningful when isolated and therefore are not relevant for our lexicon.

[5] `http://www.oxfordlearnersdictionaries.com/topic/`

To summarise, in our procedure the number of terms extracted is dynamic and highly depends on the relevance that they obtain in news. In our work, we consider the terms relevant if: 1) they appear multiple times in the same domain in several different news sources (and therefore are news worthy), and because 2) when querying Twitter with those terms, we reach a minimum number of tweets containing that term. If that number is not fulfilled, the term is removed since it is likely to be irrelevant (since Twitter users are not discussing or talking about it) or syntactically incomplete. An overview of our term extraction workflow is represented in Figure 3.1 whereas Table 3.2 shows an example of the resulting terms (extracted in 2016-04-03) for each domain.

Table 3.2: Sample of terms for each domain

Domains						
World	**Entertainment**	**Technology**	**Sports**	**Business**	**Health**	**Politics**
azerbaijan	tonight	microsoft	villanova	sales	valeant	sanders
migrant	ronnie corbett	google	thompson	money	cdc	cruz
syrian	zaha hadid	ipad pro	west indies	steel crisis	abortion pill	massive recession
tsunami	batman v superman	april fools day	bahrain grand prix	virgin america	nuclear waste	donald trump
islamic state	guns n roses	tesla model 3	ncaa title game	minimum wage	zika virus	state department

3.3.2 Term Sentiment Analysis

The second component of our system is the sentiment evaluation of the terms extracted using Twitter. Unlike other studies who track the sentiment of terms in comments from users on news site where there is a lot of spam, advertising and hate [76] or select a set of specific keywords for Twitter streams [126, 78], our system uses a combination of both approaches. Relying on headlines corpus we assure that the terms extracted are relevant (and may have a public opinion associated) and using tweets, we guarantee that the opinions retrieved are not completely anonymous and, therefore, hate, advertise and insulting comments are less common. In addition, as it was already mentioned, several works have proved good results using Twitter for topic tracking 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 [5].

Our hypothesis is that the average sentiment of a sample of tweets regarding the term, represents the public opinion or general sentiment of that term at that specific time. In order to assess that hypothesis we defined 500 tweets as the sample for each term. This number was achieved by experimental procedures which took into account the

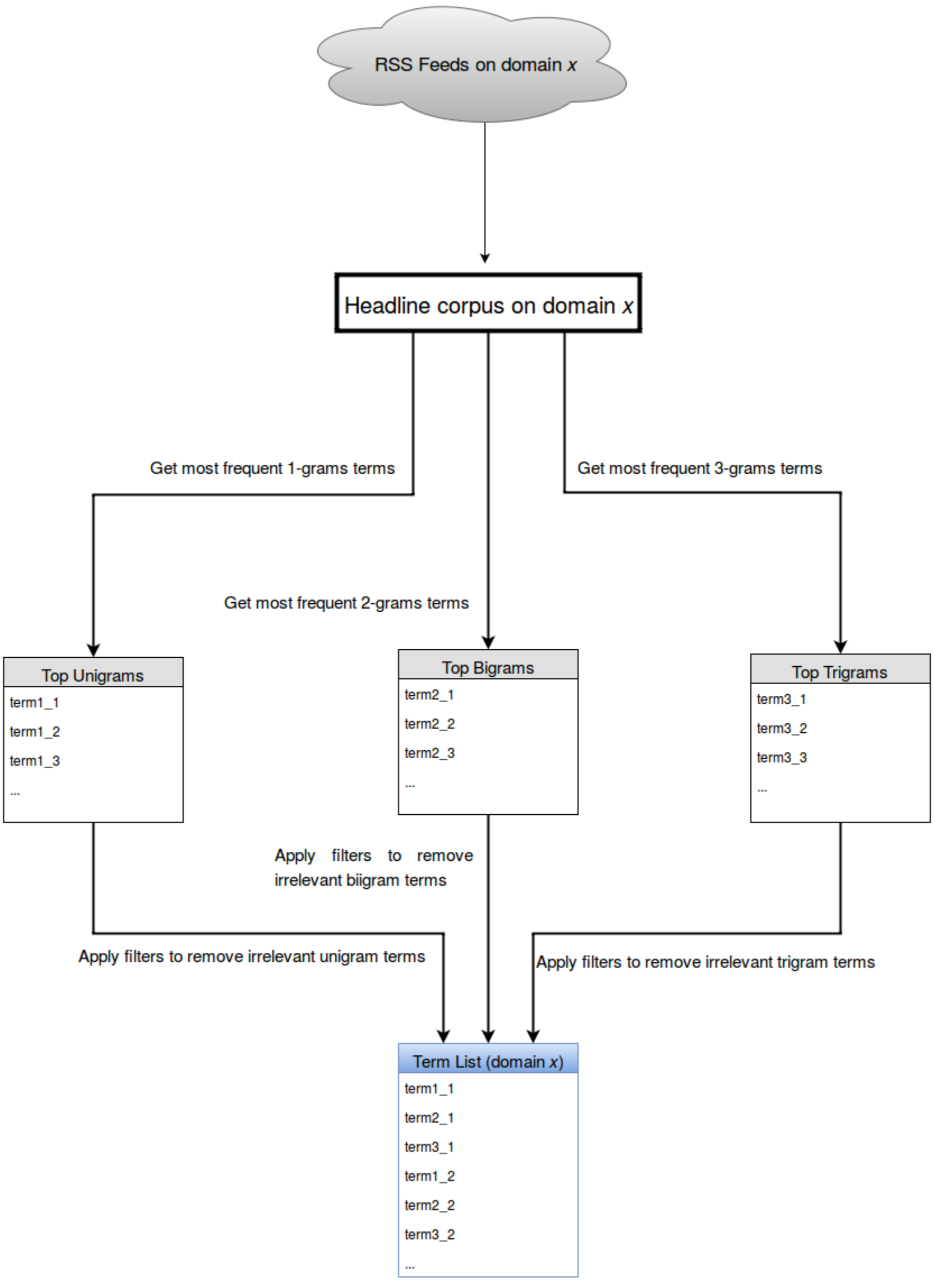

RSS Feeds on domain *x*
Headline corpus on domain *x*
Get most frequent 1-grams terms
Get most frequent 2-grams terms
Get most frequent 3-grams terms
Top Unigrams
term1_1
term1_2
term1_3
...
Top Bigrams
term2_1
term2_2
term2_3
...
Top Trigrams
term3_1
term3_2
term3_3
...
Apply filters to remove irrelevant biigram terms
Apply filters to remove irrelevant unigram terms
Apply filters to remove irrelevant trigram terms
Term List (domain *x*)
term1_1
term2_1
term3_1
term1_2
term2_2
term3_2
...

Figure 3.1: Terms Extraction Workflow

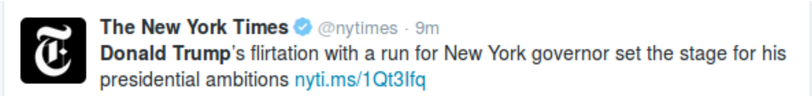
restrictions imposed by the Twitter API as well as the time to classify each tweet with sentiment. Twitter API allows retrieval of tweets in two different ways: using the REST API [122] or the Stream API [120]. The main differences between the two are presented in Table 3.3.

Table 3.3: Main differences between Twitter's Stream API and REST API

REST API	**Stream API**
Extracts a sample of tweets	Maintains a connection for continuous extraction
Focus the search on relevance	Focus the search on completeness
180 queries per 15 minutes	One persistent connection (query) per user

Although the Stream API is more suitable for our approach since it collects continuously the tweets posted, we realize that a lot of the tweets retrieved contained spam and advertising. Several twitter accounts use relevant topics to publish non related tweets so it can appear on searches. Using the search included in the REST API, Twitter already restrains through spam detection procedures, some of those non related tweets. In fact in the Frequently Asked Questions page[6], Twitter developers refer that *"Some results are refined to better combat spam and increase relevance"*. Therefore, we choose to use the Twitter REST API to avoid non related results and also use the most relevant tweets.

We also impose some restrictions on the tweets extracted for each term. Since we want to keep the sentiment up-to-date on each term, we only retrieve tweets posted in the same day as the term extraction procedure. In addition, we use the parameters provided by the REST API to retrieve the more recent tweets. Furthermore, in order to avoid extracting tweets by news sources (since we want to analyse the sentiment in common users and not in news media) we do not extract tweets that contain an external link. This is due to the fact that the majority of Twitter accounts that belong to the news industry refer to their web page in each news post (so the user can read the full article). An example of a tweet posted by a news source is provided in 3.2.



The New York Times @nytimes · 9m
Donald Trump's flirtation with a run for New York governor set the stage for his presidential ambitions nyti.ms/1Qt3lfq

Figure 3.2: Tweet posted in "The New York Times" official Twitter account

Finally, the number of tweets requested is not always the number of tweets provided by the REST API whether because there are no matches or because the term is not relevant enough. Therefore, we use this as a measure to determine if the term should

[6] `https://dev.twitter.com/faq`

be included in the lexicon. The required number of tweets was set to 33% of the sample (which in our experimentation, corresponds to a minimum of 165 tweets).

As soon as the tweet term corpus is built, we applied some cleaning procedures to it and begin the sentiment analysis in each tweet. We divide our analysis in two parts: the non textual and the syntactic. The non textual component handles the presence of emoticons and emojis in the tweets whereas the syntactic analysis deals with the syntactic construction of the text.

For each individual tweet, we start by converting the encoding of the text to ASCII. This process is essential to encode the emojis into a readable and unique character string so we could identify them in our non textual analysis. The next step is transforming to lowercase all the words in the tweet. Contrary to what was done in the headline *corpus*, no punctuation is removed due to the possible presence of emoticons. In addition, the term which composes the query is removed from the tweet. This is mainly because the bias that the possible presence of sentiment words (in the 2-gram and 3-gram lists) can have when assessing the sentiment of terms. For example considering the term "hillary wins", since the word "wins" is already associated to positive sentiment and it is present in all tweets from the corpus, it would skew our sample into a positive sentiment.

In order to separate the two components of our analysis, we must previously identify the emoticons and emojis on the tweet. For that purpose we use the results from [82] where the author assesses the sentiment of each one of the emojis. As far as emoticons are concerned, the work in [51] provides the classification of over 450 emoticons. Therefore, we will use this emoticon sentiment lexicon as groundwork for our classification.

We extracted the emoticons by matching characters in the tweet with the ones in the emoticon lexicon. With emojis, we first had to convert the emojis on the dictionary to ASCII to provide a proper matching. Since the sentiment values from both dictionaries were in a $[-1, 1]$ interval, no scaling procedure was applied. In our system, we consider emoticons and emojis to have the same sentiment impact regardless of the position they occupy on the tweet and we also don't discard repetitions because tweets with repeated emoticons enhance the sentiment of the tweet (for example "I like you :D" and "I like you :D:D:D:D:D"). The non textual sentiment is calculated by simple average. In other words, we sum all the identified emoticon and emoji sentiment values (provided by the dictionaries) and divide by the total number of occurences.

The second component is the syntactic sentiment analysis of the text. With that purpose, AFINN sentiment lexicon [79] was selected among several options. Although this particular lexicon has a smaller number of words than another lexicons it has also several advantages. First, unlike more classical approaches [54], AFINN provides 2477 words classified with sentiment values in a $[-5, 5]$ interval. This means, for example, that words like "good" and "excellent" do not have only the same polarity but also are classified with polarity strength. Furthermore, this lexicon also includes words that are normally excluded from other lexicons such as Internet slang words and obscene terms which are very common in Twitter posts. Finally, the lexicon was manually classified which removes potential errors caused from using automatic lexicon expansion methods.

In addition to the sentiment lexicon, a list of amplification, negation and attenuation words were used. Amplification words ("very", "real", "huge" or "more") or attenuation words ("barely", "rarely" or "slightly") increase or decrease the strength in 80% the value of the sentiment word to which they refer to. For example, the word "amazing" has a sentiment value of 3. Then, "real amazing" would increase that value to $3 + 3 * 0.8 = 5.4$. Negation or reverse words such as "no", "not", "don't" or "isn't" are used to reverse the polarity of the sentiment terms. Using the same word as an example, "not amazing" would score $3 * (-1) = -3$. However "not not amazing" would score $3 * (-1) * (-1) = 3$. In our experiment, due the 140 character limitation of tweets, we define that this set of words must be located up to a maximum distance of 4 words before or 2 words ahead of the sentiment word to be considered.

The algorithm used for the syntactic sentiment classification is presented in Figure 1 and has its foundations on the work in [96]. With the goal of providing the best accurate sentiment score, we extracted four different values, combining the textual and non textual sentiment approaches and assigning different weights to each one. Table 3.4 presents the percentage of each component in the 4 different scores.

The first score discards the emojis/emoticons sentiment analysis and focus solely on the second component. If the tweet does not contain sentiment words, it is not considered. The second and third score is an weighted average on the syntactic and

```
$sentiment \leftarrow 0$;
$c \leftarrow 0.8$;
for $w \in tweet$ do
    if $w \in SentimentDictionary$ then
        $s_{val} \leftarrow SentimentDictionary[w]$;
        // Get the two previous words and the four words after
        $cluster \leftarrow getWordsBefore(2) + w + getWordsAfter(4)$; $neg_{count} \leftarrow 0$;
         $amp_{count} \leftarrow 0$; $deamp_{count} \leftarrow 0$;
        for $c \in cluster$ do
            if $c \in NegatorList$ then
                $neg_{count} \leftarrow negcount + 1$
            end
            if $c \in AmplifierList$ then
                $amp_{count} \leftarrow amp_{count} + 1$;
            end
            if $c \in Deamplifier$ then
                $deamp_{count} \leftarrow deamp_{count} + 1$;
            end
        end
        $neg_{val} \leftarrow neg_{count} \bmod 2$;
        $amp_{val} \leftarrow neg_{val} * amp_{count}$;
        $deamp_{val} \leftarrow (-neg_{val}) * amp_{count} + deamp_{count}$;
        $D \leftarrow max(deamp_{val}, -1)$;
        $c_{sent} \leftarrow (1 + c * (amp_{val} - D)) * s_{val} * neg_{count}$;
        $sentiment \leftarrow sentiment + c_{sent}$;
    end
end
$N \leftarrow length(tweet)$;
$sentiment \leftarrow sentiment/sqrt(N)$;
// Constrain sentiment between [-1,1]
$sentiment \leftarrow ((1 - (1/(1 + exp(sentiment)))) * 2) - 1$
```

Algorithm 1: Sentiment Analysis text procedure on each tweet

Table 3.4: Sentiment components weights in each score

Score	**Textual Sentiment Weight**	**Emoticon/Emoji Sentiment Weight**
1	100%	0%
2	70%	30%
3	30%	70%
4	0%	100%

emoji/emoticons sentiment analysis where the second gives more relevance to syntactic sentiment analysis and the third to non textual sentiment analysis. The fourth score only considers tweets with emojis and/or emoticons and sentiment analysis is calculated entirely on that component. Tweets without emojis or emoticons are discarded. Then, for each score we calculate the respective average from all tweets and convert it to a Positive/Negative scale. Finally, in this first phase, we assign the four scores to the term (that was used to extract the *corpus*). However, in the next chapter we will select one of the four to be included in the lexicons. An overview of our final sentiment analysis component (with one of the four scores already selected) is presented in Figure 3.3.

Although we are extracting terms from already defined news categories (and therefore probably are correctly assigned to that particular domain) and there are some studies in the Twitter Topic Tracking that show evidence that our sentiment classification approach may be successful, it is important to evaluate the different components from our system. In addition, as mentioned previously, we must also select one of four outputted scores from our system to be included in the sentiment dictionaries.

Therefore, a careful evaluation on term extraction and term classification as well as the tweet sentiment analysis method (which has high influence on each term final sentiment) is needed.

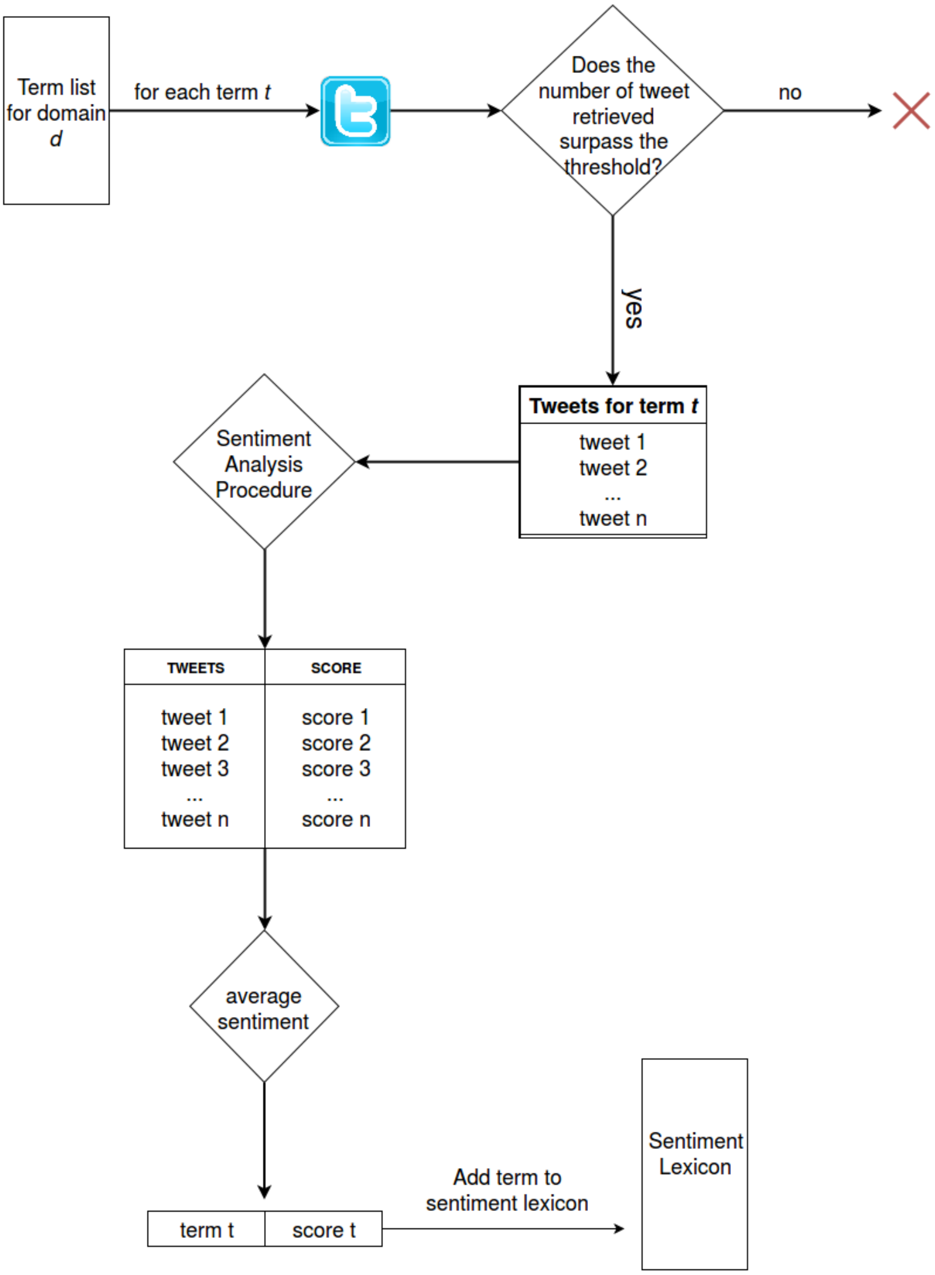


Figure 3.3: Term Sentiment Analysis Workflow

Chapter 4

Evaluation of the System

In the previous chapter, we explained our system workflow. This system intends to expand the current sentiment lexicons by adding terms with polarity based on current public opinion. Although several studies use topic tracking to identify the current sentiment on Twitter on some terms, only a few compare the sentiment obtained with real world surveys. However, in sentiment lexicon expansion studies, evaluation using manual labelling on the added terms is frequent and provides evidence on the accuracy of the sentiment lexicon expansion method. Since we are proposing a sentiment lexicon expansion system, we will rely on this kind of evaluation to assess the accuracy of our lexicon. However, unlike other proposals, our system suffers from the impact that time can have on the terms. Therefore, the evaluation must be handled with special attention due to this factor.

There are four important evaluations that we must consider to ensure the effectiveness of our system. First, we evaluate whether our term extraction component is accurate on the retrieval of domain and time specific terms. Then, due to the importance of the tweet sentiment analysis method on the outcome of each term polarity value, it is necessary to assess its performance in several different domain datasets. In addition, this evaluation will also be used to select one of the four sentiment scores returned by our system. The final two evaluation stages are the term sentiment classification performance (when compared to a manually labelled ground truth) and study the sentiment dynamics on a subset of cases with the goal of providing evidence on their accurate sentiment variation.

4.1 Term Extraction Evaluation

To evaluate the term extraction procedure, we conduct an experimental survey to determine if the terms extracted were up to date and belonged to the domain from where they were retrieved. A web application was built for that purpose. The survey was conducted during the time period of two days (16 and 17 of March 2016). 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. The possible responses were "Yes", "No" and "I don't know" in case the user was unfamiliar with the term.

The survey was shared among social networks and university students. We do not restrict the number of terms that each student could evaluate being only limited to the full extension of the term list extracted. In addition, the terms are extracted from a global term list and assigned to each individual user in a consecutive way. Therefore, with this approach we will have approximately the same number of evaluations by term.

A total of 1414 entries were classified by 57 different users consisting mostly of university students. We discarded all results whose response was "I don't know" which correspond to approximately 5.5% of all evaluations. Furthermore, we only considered terms who had at least 3 evaluations and we consider our ground truth the majority of the evaluations.

Our results show an accuracy of 90.9% on the fitness of the domain and time. In 4.2% of the terms, consensus among evaluators was not achieved and in 4.9% our term extraction feature failed to correct assess the domain or time of the term. These results provide strong empirical evidence for our term selection method.

4.2 Tweets Sentiment Analysis

The second evaluation concerns the sentiment analysis of tweets. Since this step has a great impact on the final term sentiment we find indispensable to assess it's performance.

This evaluation will serve two purposes: 1) evaluate if the sentiment approach of our system is accurate in different datasets containing tweets regarding different domains and 2) since we are returning 4 different sentiment scores for each entry of our corpus and our hypothesis is that the average sentiment corresponds to the term sentiment, it is important to determine which one is the more accurate in assessing the correct polarity of tweets.

With those purposes in mind, we selected tweet datasets already labelled with sentiment by manual annotators. Our main source was Crowdflower's "Data for Everyone" library. Crowdflower[1] is a platform where is possible to submit jobs or tasks to be manually done by workers in exchange of a small fee. Their "Data for Everyone" library contains the results of some of those jobs in an open access form. We found some jobs whose goal included the sentiment classification of tweets. We obtained several datasets which focus on different subjects (such as technology, politics and entertainment) since we will use the same system to classify tweets regarding different domains.

Five datasets of tweets, classified with sentiment by human annotators were selected. A brief explanation of each dataset follows (for more details please refer to [24]):

- **GOP Debate (GOP):** contains over ten thousand tweets about the GOP debate in Ohio. Workers classified the sentiment of each tweet as Positive, Neutral or Negative.

- **Self Driving Cars (SDC):** includes approximately 7000 tweets. 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.

- **Airline Twitter Sentiment (USAIR):** dataset with around 16000 tweets about major US airlines. Contributors were asked to assign a Positive, Neutral or Negative sentiment to each tweet.

- **Coachella (COACH):** dataset with 3847 tweets with reactions to the Coachella festival 2015 lineup announcement. Workers classified the sentiment of each tweet as Positive, Neutral or Negative

[1] https://www.crowdflower.com/

- **Apple Computers (APPLE):** 4000 tweets containing references to the Apple company. Sentiment classification was done with a Positive, Neutral and Negative scale.

Our system evaluates tweets in a two class (Positive/Negative) fashion. Therefore, to evaluate this component, we focus solely on tweets that are classified by annotators with positive or negative sentiment and discard neutral classification.

It is important to refer that this process is used to evaluate the polarity (positive/negative) accuracy in our sentiment analysis component. But, most of the tweet sentiment systems evaluate tweets in 3 classes: negative, neutral and positive, where the neutral class is when no sentiment cues are find in the tweet (whether they are opinion words or non textual features). However, our main focus are the tweets who have a positive or negative opinion on a certain term since our sentiment lexicons will only be composed of terms with those two classes. Therefore, in our system we discard the tweets without sentiment cues and the same is done in this evaluation.

This is an unusual process in the evaluation of sentiment systems since we discard possible positive or negative tweets. However, since the final purpose of this component is the evaluation of the term and not the tweet itself, our hypothesis is that the removal of possible tweets with sentiment that are not captured by our system, will not have a huge impact on the term classification, if the sample is large enough.

At this stage, due to the absence of emoticons and emojis in the majority of the tweets contained in the datasets, we discarded the score which assign a 100% weight to this non textual features.

We assess the performance of this component using 4 different metrics: precision, recall, F1-score and accuracy. We will now briefly explain these concepts. However, in order to understand the measures used, we must first introduce the concepts of true positive, true negative, false positive and false negative. We must stress that the positive and negative on these concepts is not related to our sentiment labels. To avoid possible misinterpretations we will use a different problem.

Consider the problem of classifying a tweet as subjective or not.

- **True Positive (TP)**: When a tweet is classified as subjective by the system and is in fact subjective

- **False Positive (FP)**: When a tweet is classified as subjective by the system but is in fact not subjective.
- **True Negative (TN)**: When a tweet is classified as not subjective by the system and is in fact not subjective
- **False Negative (FN)**: When a tweet is classified as not subjective by the system but is in fact subjective

In the evaluation of classification tasks, the classification which the system is trying to match is normally made by human annotators and it's designated as the "ground truth". With these concepts defined, we can now expose the formula for each one of the metrics.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

Precision deals with how many of the classifications returned by the system are correct (i.e. how many of the subjective tweets returned by the system were in fact subjective) whether recall handles the number of correct classifications that the system returns (i.e. how many subjective classifications does the system returns). The F-measure is an harmonic average of both precision and recall. Finally, accuracy is the percentage of correct classified cases that the system returns.

Analysing the results, although the discrepancy on the scores was minimal (mostly because the quantity of tweets with emoticons was reduced) we opt to use the score with 70% analysis on the text and 30% on the emoticons/emojis since it performed, on average, slightly better. Table 4.1 shows the results of our system in the correct polarity classification of the refer datasets using the selected score. The remaining tables can be consulted in the Appendix 8.1.

Table 4.1: Results in terms of precision (Prec.), recall (Rec.), F1-score (F1), and accuracy (Acc.)

Dataset	Prec. (%)		Rec.(%)		F1(%)		Acc.(%)
	Pos.	**Neg.**	**Pos.**	**Neg.**	**Pos.**	**Neg.**	**Pos.+Neg.**
GOP	32.69	90.88	78.28	57.32	46.12	70.30	61.26
SD-Cars	82.59	53.99	76.70	62.84	79.54	58.08	71.40
US Airlines	39.84	97.62	95.27	57.43	56.18	72.32	71.85
Coachella	85.32	41.61	73.35	60.07	78.89	49.16	70.20
Apple	56.00	93.90	86.96	74.60	68.13	83.14	77.31

Results show that the sentiment analysis component of our system reached satisfactory results in datasets containing tweets from different domains. Accuracy reaches the lowest value in the GOP dataset. Similar conclusion was reached in [111] where the authors assess the low performance on some web extracted datasets due to political and controversial topics. These results are in agreement with other two class sentiment analysis dictionary based approaches compared in [95] and provide a good support for the reliability on tweet classification of our system. At this point the score that assigns a 70% weight for sentiment text evaluation and the remaining 30% for emojis and emoticons was selected and the other scores were discarded from our system.

4.3 System Evaluation

Finally, we evaluate the performance of our system as a whole, assessing its accuracy in the evaluation of the sentiment of terms extracted from Twitter and if it reflects the current sentiment.

To assess this hypothesis, we resort to Crowdflower to conduct a sentiment survey since this is the most common validation of sentiment lexicon expansion or sentiment lexicon creation in the literature . As it was previous mentioned, Crowdflower allows to submit jobs to human annotators (workers) in exchange for a small fee. Our job had the following specifications. We selected workers from the United States and the United Kingdom due to the origin of our news sources. Therefore we presume this would avoid unfamiliarity with the terms. We also limited our job to workers with level 3 performance. This level is assigned to accounts who have already answered more than 100 test questions and achieved a very high accuracy [22]. Test questions are quality control questions which are already labelled with the answer. They are inserted during the job to exclude workers that have low performance. In our job, we did not select any test question because 1) it would increase the cost of the job and 2) by selecting only level 3 workers, we are already filtering our crowd-sourcing sample with high quality annotators. The question asked 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. We provided a likert scale for the answer that ranged from 1 (very negative) to 5 (very positive). The labels very negative and very positive were mapped to 1 and 5 respectively, to provide additional information to the workers. 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 random (and easier) choice.

As far as our data is concerned, we used terms retrieved from 01/04/2016 till 03/04/2016. The job on Crowdflower began on 04/04/2016 at approximately 3:15 pm and took 30 hours to complete. We submitted 101 term/domain pairs in equal number for the same domain from the daily extractions. Each term was evaluated by 7 different workers where each worker could evaluate a maximum of 10 terms.

The median of the 7 evaluations was used as the ground truth for each term. For example, if six workers assign the term a sentiment value of 2 and only a worker assigns a 5, the average would result in a unrealistic 3. Therefore, using the median, we still consider all workers opinions and the final sentiment, which returns a more realistic value (in this case 2).

The next step was to convert the final values to a Positive/Negative scale. Values above 3 were considered Positive while values below were assigned with a Negative label. Neutral values were once again discarded since our system assigns only a positive/negative scale.

However, the neutral terms had a significant high presence in the sample evaluated (around 40%). We suppose that there are two main reasons for this value.

The first is that the unfamiliarity of the term could lead workers to assign a neutral value. In fact, before the experiment, we deliberate on the addition of a "I don't know" option when we were designing the job. However, our conclusion was that lazy workers would select this option to quickly answer the question.

The second is that, in fact, the term has a neutral value and therefore, the addition of a neutral classification or a subjectivity detection feature (in the term extraction procedure) should be accounted in future work.

Curiously, the remaining terms returned were balanced since 47% of data is labelled as positive and the rest as negative. Using this as ground truth, we compared the results from our system (using the selected score), 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 that is more frequent). The results are presented in Table 4.2.

Table 4.2: 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% with our selected score. 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, failing solely in "Batman vs 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 (and therefore ignoring the time factor) 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 narrow scope could result in more neutral classifications due to unfamiliarity of the term to the workers.

4.3.1 Our System vs DatumBox

In addition to comparing our system with the random and majority baseline, we also compare it with a machine learning platform called Datumbox[2]. This framework includes an API capable of, given a document of any length, provide a sentiment score in Negative/Neutral/Positive scale. Through experimentation, we are able to presume that the API adapts according to time. For example, when looking for words like "Paris" and "ISIS" at the time of the terrorist attacks, the returned polarity in both cases was "Negative". However, we could not confirm this since we found no information regarding it and we obtained no answer from the system creator.

Once again, we were forced to discard the neutral evaluations since our system limits the evaluation to a positive/negative scale. This results in a even smaller number of terms for evaluation since the neutral were excluded from the Crowdflower and the Datumbox framework in order to do a direct comparison between the ground truth and both systems. The results (evaluated with the same metrics as above) are provided in Table 4.3 and show that our system also surpasses the DatumBox framework.

Table 4.3: Comparison of results of our system (SR) against the results from DatumBox framework (Datum)

	Prec.(%)		**Rec.(%)**		**F1(%)**		**Acc.(%)**
	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos+Neg
SR	**70.8**	**89.5**	**89.5**	**70.8**	**79.1**	**79.1**	**79.1**
Datum	66.7	61.8	31.6	87.5	42.9	72.4	62.8

4.4 Analysis on sentiment changes through time

Not only is important to assess the sentiment of the terms provided by our dictionary but also to detect if the sentiment changes through time with respect to those terms. As it was mentioned before, our approach relies on the principle that public opinion (or sentiment respecting the terms) changes according to the news.

Based on the premise described above, one way to validate the sentiment changes on the terms is to determine if, when a term appears on the news, there is an interest on

[2] `http://www.datumbox.com/`

assessing information on that term in order to make valid decisions on its sentiment. In other words, to determine if 1) there is an increase on a term popularity when news regarding the term are released and 2) if that popularity is, in anyway, associated with changes on the sentiment of that term.

For that purpose we analysed the daily sentiment dictionaries retrieved from March to July 2016 to inspect terms that were constantly present. Then, in order to assess if there was some public interest regarding those specific terms, Google Trends[3] was used.

Google Trends allows users to retrieve data relatively to the number of Google searches on a certain topic. It also allows to tune the filters to a given region or time interval. The "trending" variable is returned in a percentage like value in which 100% is the maximum number of queries reached in the defined time interval and all other values are scaled based on it.

Therefore, defining the time interval from 11th March , to 26th June of 2016 and using each term as a Google Trends query, it is possible to analyse the variations on the number of searches of that term in this specific time interval.

As an example, we retrieve the term "donald trump" since it was a rather controversial entity and with considerable presence on the news due to the U.S. presidential elections. The trending plot is shown in Figure 4.1.

Finally, adding the sentiment analysis of the term retrieved from our system in each day and, overlaying it in the trend line, results in the plot presented in Figure 4.2. The "Missing Data" markers represent the days where for some reason (like Internet connection loss or errors/bugs on the system) it was not possible to retrieve sentiment dictionaries.

There are some interesting remarks we can observe in the plot. The first is that when there is a trending peak on "donald trump" there is a polarity detection by our system in the same day or in the days close to it. This is more or less expectable since our term extraction component relies on news and when a term becomes trending is normally because something has occurred (which is reported by news sources).

[3] `https://www.google.com/trends/`

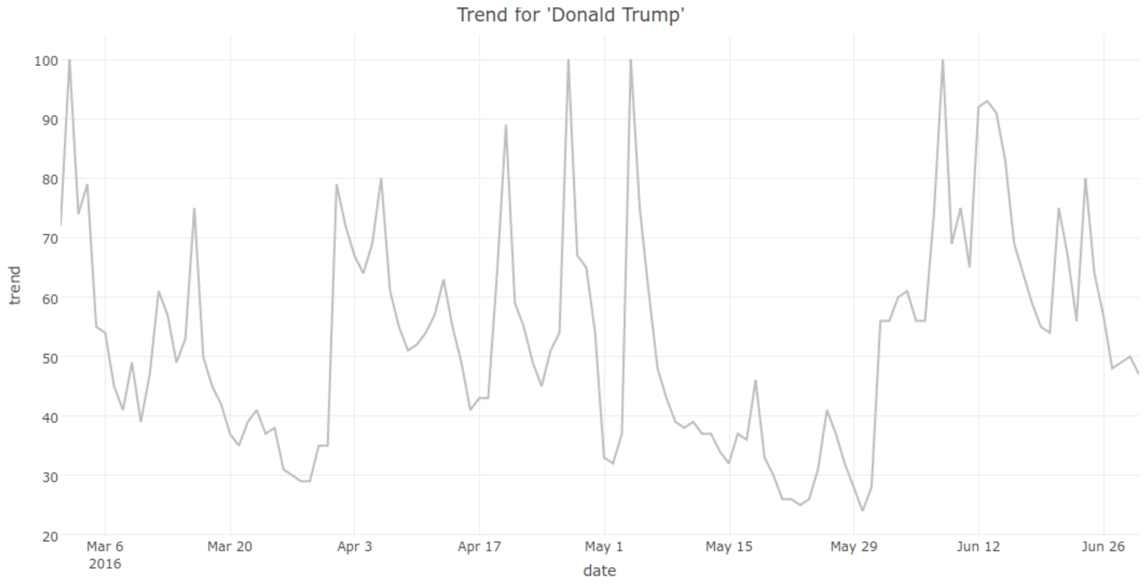


Figure 4.1: Trend for "donald trump" using Google Trends

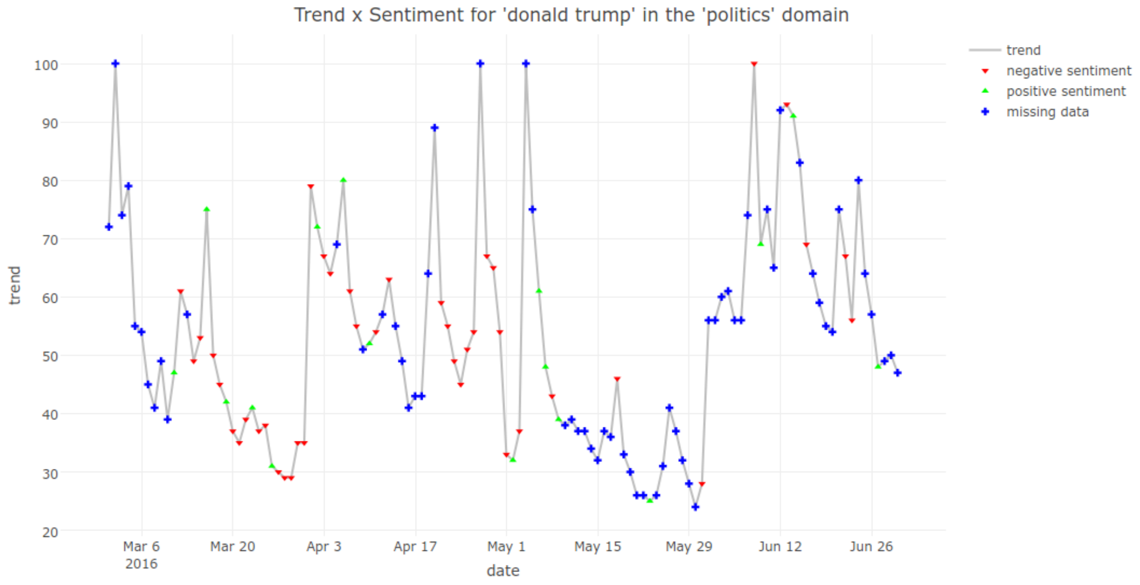
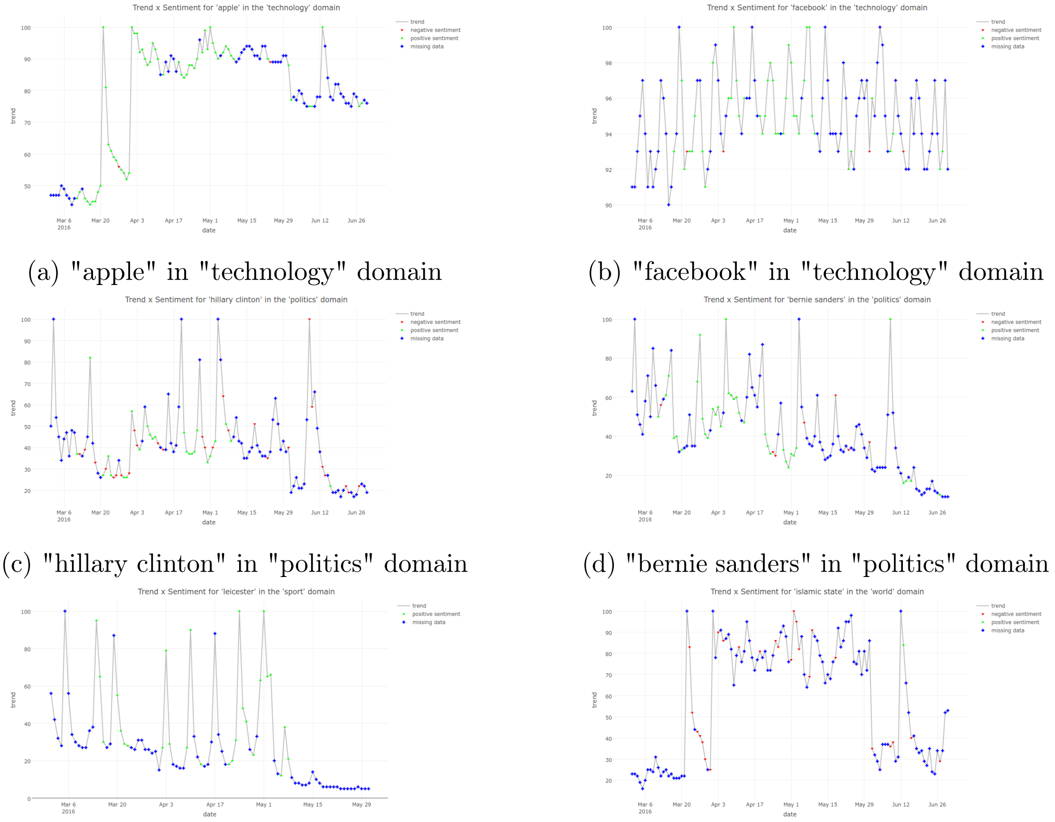


Figure 4.2: Trend and sentiment for the term "donald trump" in "politics" domain

However, this single observation is not conclusive whether a trending peak is only associated with changes on the sentiment polarity. We can examine that in some cases happens (for example the biggest peak between March and April and the second peak between April and May), but is not conclusive that these changes always occur at the



(a) "apple" in "technology" domain

(b) "facebook" in "technology" domain

(c) "hillary clinton" in "politics" domain

(d) "bernie sanders" in "politics" domain

(e) "leicester" in "sports" domain

(f) "islamic state" in "world" domain

Figure 4.3: Plots of trend and sentiment for different terms from different domains

peaks. Similar plots for other terms are shown in Figure 4.3. For a better visualisation of the plots, please refer to Appendix 8.2

There are some observations worth mentioning if we dissect each individual term present in the sample provided in Figure 4.3. Both terms in the technology domain seem to have few polarity changes, being that the positive sentiment is much more frequent than the negative one. Furthermore, the trend variation is very small (10% in "facebook" term) comparing to the other examples provided. Our assumption is that, in terms that are web related services, the common user often "googles" a certain service entering directly into the service website (users often search "facebook" instead of accessing it via URL).

Political terms (such as "gop" and "elections") and entities (such as the ones provided in Figure 4.3c and 4.3d) have a more volatile sentiment. In addition trend values in this domain seem to fluctuate more. The terms "hillary clinton" and "bernie sanders" vary in 80% (between 20% and 100%) while "donald trump" minimum trend value is 60%.

Lastly, we present two cases where the sentiment polarity seems to remain unchanged through the interval of time considered. In one hand, Figure 4.3e presents the trend and sentiment of the term "leicester", which in the sports domain, was the very unlikely Premier League champion. Furthermore, the last maximum peak corresponds to the day they became Premier League champions. It is plausible that, being an underdog team in a path to accomplish a great achievement, the public opinion would remain positive through time. In the other hand we have the "islamic state" term which due to the association with terrorist attacks has a negative sentiment across time. The peaks are frequently associated with the presence of the term in the news media and so, when no missing data occurs, Figure 4.3f displays sentiment retrievals in, or very close to them.

4.5 Result Analysis and Discussion

In this section we provide an evaluation of both components of our system as well as a full system evaluation and a posterior analysis on sentiment dynamics. As it was mention in Chapter 2, there are two types of automatic or semi-automatic lexicon expansion methods: theasaurus and corpus based. However, (and although we consider our approach to fit the corpus based category) our system cannot be compared to any of these methods. This is because traditional corpus based methods focus solely in one corpus and retrieve the terms of that corpus instead of generating a corpus for each extracted term of those categories. Furthermore, such approaches focus in retrieving opinion words classified majorly as adjectives and nouns. Consequently, any term comparisons with other state of the art methods is hard to achieve.

With this lack of obvious competitors we turned our focus on determining if our system was accurate in both of its components. First, we assess if the extracted terms fitted the domain and the time they were extracted. This was evaluated by workers and achieved an accuracy of 90.1%. An analysis was conducted to understand which extracted terms did not fit the domain where they were placed or the time when they were extracted. A sample of these terms can be found in Table 4.4.

Table 4.4: Sample of unfit terms

Domains					
sport	**entertainment**	**politics**	**business**	**health**	**technology**
wales	world	race	need	times	next
captain	king	cruz	bangladesh	court grants	
history	white house	calls	february		
eddie jones	david		need		
leaves					
...					

The number of terms that do not fit appear mainly in the sports domain. In fact, in the Table 4.4 are presented all terms that do not fit the domain except the sports domain list (which contains more terms than the presented ones). As we can also observe, there are some terms which were wrongly classified by the POS tagger. For example, the term "leaves" in the sports domains or "calls" in politics should have been filtered in our term extraction component. Furthermore, there are some terms (specially unigram terms) that do not represent any particular sentiment without a context. As an example we have "february" in the business domain, "times" in health or "next" in technology. There are also some bigram terms like "court grants" which do not represent any particular context.

There are also some clearly wrong domain classifications. The term "white house" should fit more the politic domain than the entertainment one. However, since it appeared in news from the entertainment domain[4], the term is included.

Table 4.4 also includes some examples where the term fits the domains but by unfamiliarity of the users, it ended up as an unfit. This is the case of "eddie jones" which refers to two different persons: a rugby coach (and former player) and an actor. The familiarity of the worker relative to one, but not to the other may ended up as a miss classification term. In addition, this raises another problem in the sentiment component of our system. Although unlikely, there is a possibility of terms which refer to two different persons or entities (like "eddie jones") occurring in the same

[4] `http://www.npr.org/sections/therecord/2016/03/15/470515719/hamilton-freestyles-at-the-white-house-mic-drop`

time interval. When retrieving the tweets for posterior sentiment analysis, we must be aware to which domain each tweet is referring to.

Furthermore, even if they are referring to the same entity, it can also be in a domain specific context. For example, "Arnold Schwarzenegger" may be associated with the entertainment domain as an actor but also to political domain as the former governor of California. Public opinion may be different when considering each one of the domains. This stresses the importance to distinguish the domain in duplicated terms.

As for the sentiment component of our system, we are aware that it lacks comparison with other systems. In fact, the majority of the systems for sentiment classification include the neutral class on their classification. An inclusion of such class can be useful in our sentiment system analysis component in order to directly compare it against other systems and also because of the high presence of neutral terms.However, since we were interested on the classification of positive and negative terms, we develop our sentiment component accordingly.

Regarding the system itself, it is important to highlight that a direct comparison with other lexicon expansion methods mentioned in Chapter 2 is not possible. Dictionary-based systems focus on expanding the seed words provided with synonyms and antonyms via thesaurus whether corpus based approaches tend to find domain specific words in a text corpus. Both approaches focus on verbs and adjectives and in time independent words. Therefore, we compared our system against a random baseline and a majority baseline to, at least, prove that our system yield better results than arbitrary choice or choosing the class with more elements. We also matched our system with an existing sentiment framework that, although seeming to be aware of the sentiment variation of terms through time, it is probably not adjusted for the particular task of assessing public opinion sentiment on terms. This also suggests that "traditional" methods for sentiment analysis are unfit for this task. Nevertheless, our system provides a better performance in all metrics evaluated comparing to this sentiment framework.

Finally, we analysed the changes of sentiment through time of a small sample of terms and ascertain if it was somehow associated with peaks on search engines trend values provided by Google Trends. This experiment was a bit inconclusive since in some cases sentiment changed in a trending peak (which is plausible since the increment of searches could be associated to an event that consequently changes public sentiment) and sometimes did not (which is also conceivable if we consider the example of "leicester" provided in Figure 4.3e). In addition, results like the ones in Figures 4.3e and 4.3f are good examples of the performance on sentiment detection of our system since these two terms are somewhat expectable to have the same polarity across the sampled time interval of our experience. Same goes with the results shown in Figures 4.3c and 4.3d which are terms whose sentiment presumably varies due to political campaign and that are also captured by our system.

All the evaluations carried out in this chapter provide solid evidence that our system is already capable of automatically provide correct sentiment on terms which are domain specific, and also detect variations of that sentiment along time. However, some improvements were discussed in order to further increase the performance of the system. Domain disambiguation when duplicated terms occur in different domains and a 3 class (Negative/Neutral/Positive) sentiment classification for tweet analysis seem to be the more important and so, its implementation becomes crucial before assessing the performance of the extracted dictionaries in state of the art sentiment analysis approaches.

Chapter 5

System Improvements

In the previous chapter, we have collected strong empirical evidence that Twitter can be a good and up to date source for public opinion on relevant terms. We test this hypothesis by retrieving terms from several news sources in different domains and send them as search queries to Twitter. Then, we assess the polarity of the tweets regarding each term, and test (through a Crowdflower job) if the average value is in fact the sentiment of the term.

Although the results achieved are good indicators of the sentiment at the present time, the system could use some improvements. First, there are terms that appear in several domains which may have different sentiment on each one. For example, "apple" in a technology domain has a different meaning than the same term in a health domain (therefore, it is plausible that they also may have different sentiment polarities). So, when detecting terms that appear in two or more domains, it would be important before detecting the sentiment of the tweets, to classify them with the correspondent domain in order to improve the quality of the dictionaries.

On the other hand, the evaluation described in the previous chapter has shown that although terms are relevant at the present time and fit the corresponding domains, they are not always associated with a positive/negative class. In fact, the classification of neutral terms by the workers has reached 40% of the sample. Therefore, these results showed that a 3 class sentiment evaluation is more suitable for our problem.

As a consequence, the inclusion of a method to detect tweets domain (when the same word is present in different dictionaries) and the improvement of our sentiment analysis component method to a Positive/Neutral/Negative classification, are probably the most important features that can be added to the system in order to increase its effectiveness.

5.1 Domain Disambiguation Method

Recalling the workflow of our system (Section 3.3). Summarising, the terms are retrieved from the RSS feeds corresponding to a specific domain and then are sent to Twitter as a search query, where a sample of tweets is retrieved. Then, the sentiment analysis component classifies each tweet and the average of the scores corresponds to the final term sentiment.

Therefore, in order to detect duplicated terms in different domains, we modify our system to aggregate all pairs (word, domain), before we proceed to the sentiment analysis component. Only the "world" domain is excluded from this procedure since the news from this class are very often included in other categories. Therefore, it is not suitable to disambiguate the terms in this domain because we are unaware if it was already referring to news included in other categories.

The duplicated terms and correspondent domains are extracted and passed down to our Domain Disambiguation component, whereas the remaining ones follow the original workflow (mentioned in Chapter 3).

A duplicated term is converted to a Twitter query following the same conditions as in the original workflow. However, when the sample of tweets is retrieved, each tweet is evaluated and classified with one of the domains where the term occurs. For example if "rio" appears in both "sports" and "entertainment" domains, for each tweet retrieved, the system will try to "place it" in one of them. If it fails to assign one of the domains, the tweet is discarded from the sample.

The domain classification procedure is a simple method that uses the lexicons provided by the Oxford Learners's Topic Dictionaries [83] (that were previously used to remove general domain specific words from the terms extracted from news). For each tweet we measured the domain "fitness" by the frequency of words in the tweet that occur in each domain lexicon. A scheme of the procedure can be seen in 5.1

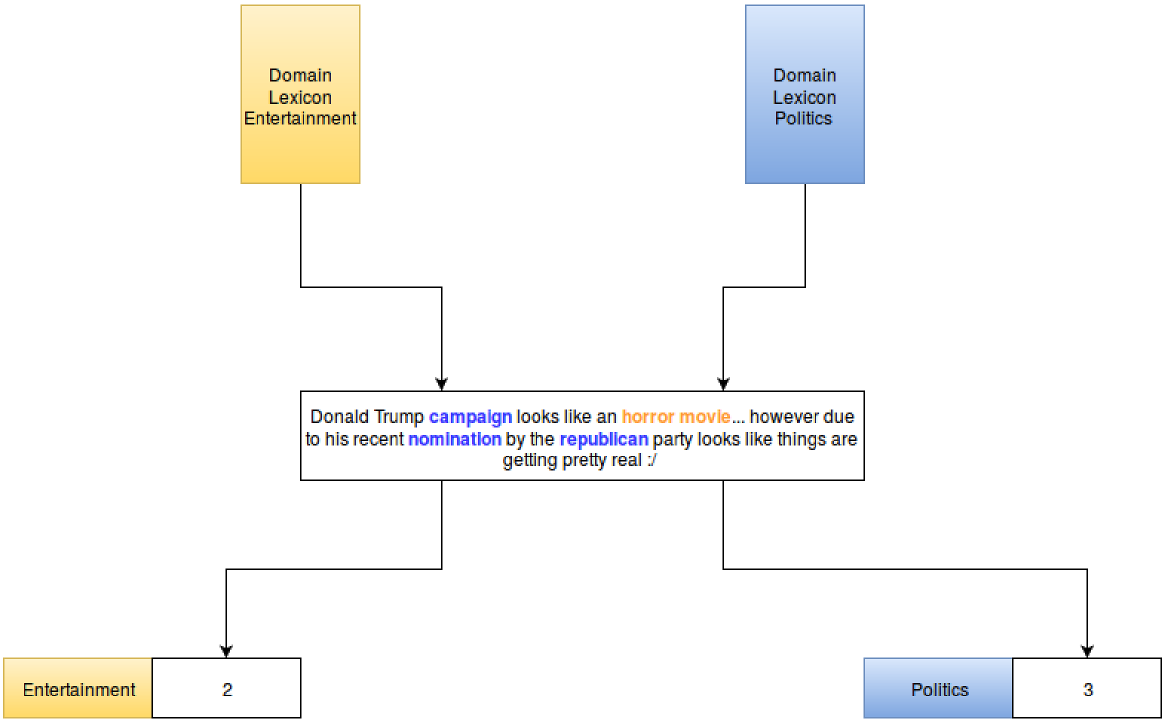

Domain Lexicon Entertainment
Domain Lexicon Politics
Donald Trump campaign looks like an horror movie... however due to his recent nomination by the republican party looks like things are getting pretty real :/
Entertainment
2
Politics
3

Figure 5.1: Tweet Domain Disambiguation Procedure

One of the problems with this approach relies on the way Twitter Search API works. When a specific term with several domains is sent as query, it is necessary to repeat the process until the tweets for each domain match the minimum number of samples required.

The number of retries was defined experimental to 5. This was approximately the maximum number of times that was required for the minimum number of tweets (from different domains) to be extracted. Additionally, in our experiments, duplicated terms never surpassed three different domains. Therefore and to avoid a long processing time on duplicated terms, five retries appears to be a reasonable number to use. Verification to identify and exclude duplicated tweets was also considered by checking other information like the tweet timestamp and user id.

When testing the implementation, we already had a small example of how domain disambiguation is important for some terms. We retrieved a sample of 100 tweets for the term "england" which was found in two different domains on a test made in August

2016. Those domains were business and sports and whereas the sentiment assigned to the term in the first was negative, in the second it was the opposite. This is in accordance with our expectations since in a business domain the sentiment of "england" could be related with the consequences of "brexit", while in the sports domain can be associated with the Olympic Games victories.

5.2 3-Classes Sentiment Analysis Component

A three classes sentiment system on tweets could improve the accuracy on the terms. Therefore, we adapt the algorithm used on the original sentiment component of our system to detect neutral tweets.

This variation consists in a small threshold interval which determines that if the sentiment score is to close to zero then it is a neutral tweet. This can occur in two different scenarios: when the sum of the clusters results in a value inside the defined interval or when there aren't any sentiment words on the tweet. Implementing these changes makes it possible to compare our original system with other state of the art approaches.

5.2.1 Evaluation of sentiment component against other systems

With the goal of achieving the highest performance on sentiment classification on tweets, a comparison of our sentiment component with several state of the art methods was conducted using the iFeel framework (which aggregates 19 sentiment analysis systems). Giving a tweet and a method, iFeel returns a -1, 0 or 1 corresponding to a negative, neutral and positive sentiment respectively [44].

We used the previous mentioned datasets in Section 4.2 and added a new one, consisting of tweets regarding the New England Patriots "Deflategate". This way, we include a sample of sport related tweets [110].

Since we want to obtain a system that is accurate in all 3 classes on all domains, we selected a balanced sample of 1200 entries in each dataset, equally divided by sentiment classes (400 negative, 400 neutral, and 400 positive). Finally, using the iFeel framework, we compare our system with 19 sentiment analysis methods. Although some have already been mentioned, a brief description on each one follows:

- **AFINN**: Twitter based sentiment lexicon expanded from ANEW [14]. It contains words that are frequently used in this social network such as Internet slang and offencive words [80].
- **Emolex**: Manual created emotion lexicon using crowd-sourcing. The terms were extracted from a combination of The Macquarie Thesaurus [9], General Inquirer [104] and WordNet Affect Lexicon [123]. Although the words were classified with emotion and polarity, only the second was used for this method [75].
- **EmoticonDS**: It is a lexicon created using a corpus based approach. The method consists in the extraction of tweets with only a happy (*":)"* or *":-)"*) or sad (*":("* , *":-("*) emoticon. Then, the assumption is that tweets with a smiling emoticon correspond to positive tweets and with a sad emoticon to negative ones. Finally, the corpus is divided considering the emoticons and the most frequent words in each division are included in the lexicon [46].
- **Emoticon**: Uses a set of negative, neutral and positive emoticons as sentiment lexicon to classify tweets. This is a limited method since all tweets without emoticons are evaluated as neutral [44].
- **Hapiness Index**: Uses words from ANEW that were manually classified with an 1 to 9 happiness scale. To assess the sentiment, this method considers that positivity is achieved when the happiness value for a tweet is between 6 and 9 whereas negativity is between 1 and 4. Tweets with no words associated or with happiness value 5 are considered neutral [44].
- **MPQA (or Opinion Finder)**: It is a machine-learning model to detect subjectivity and consequently, the polarity of a sentence based on sentiment clues [129]. Since each sentence can have more than one sentiment clue, this method considers the sum of them as the final sentiment score.
- **NRC Hashtag**: uses the same concept as EmoticonsDS although, instead of emoticons, the tweet retrieval process is done with emotion hashtags such as

"#angry" and "#happy". The lexicon evaluates each word with six different emotions and positive and negative sentiment [74].

- **Opinion Lexicon**: Extracts and classifies opinion words from a corpus of reviews to build a lexicon. Uses a thesaurus based approach and a seed lexicon of 30 words as a starting point [53].

- **PANAS-t**: Is an adaptation of the Positive Affect Negative Affect Scale (PANAS) to detect variations on Twitter mood [45]. The lexicon was initially built by classifying each word with one of ten different moods. The adaptation made in IFeel was assigning the words classified as joviality, assurance, surprise and serenity as positive sentiment whereas the negative words are the ones originally classified with fear, sadness, guilt, hostility, shyness, and fatigue. The attentiveness mood was considered as the neutral class . However, the resulting lexicon on this method has only 53 words which is very low for a sentiment lexicon.

- **SANN**: Uses the sentiment lexicon of MPQA along with polarity shifters, negation and amplifiers to build a sentence-level sentiment classifier. It was original used in user comments present in Ted Talks videos [87].

- **Sasa**: It is a supervised method based on a Naive-Bayes approach. It uses the unigram features of each tweet. This method was original used to detect sentiment on tweets in real time during the U.S. 2012 elections. [126]

- **SenticNet**: Assigns sentiment to common sense concepts to achieve a semantic sentiment analysis approach rather than the most common sentence level [19].

- **Sentiment140 Lexicon**: is a corpus based sentiment lexicon extracted from the tweets provided in [43]. It has similarities with the NRC Hashtag method for lexicon extraction and practically equal to the EmoticonDS (only in a different corpus).

- **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 [114, 113, 111] .

- **SentiWordNet**: Is a lexical resource which provides all WordNet entries with a positive, negative or neutral polarity. A short lexicon consisting of seven positive terms and seven negative terms were used. Next, a dictionary based approach was used on WordNet, with a limited reach on each word (meaning that each seed lexicon entry should not expand by synonyms or antonyms more than k times). Finally, all the classified terms are used as training data on a supervised model to assign a score to the remaining ones [7].

- **SoCal**: Uses a sentiment dictionary and features like, negation and amplification words. The authors claim that the dictionaries used are robust through several Mechanical Turk evaluations [105].

- **Stanford Adapter**: Uses a deep learning scheme more concretely a Recursive Neural Tensor Network to determine the sentiment at a sentence level. This method provides a differentiating feature which is the order of the words in the sentence is taken into account for sentiment assessing [103].

- **Umigon Adapter**: is a system designed specifically for tweets sentiment analysis. It is a dictionary based approach that has characteristics like the detection of smileys and onomatopes (p.e. "yeeeeeaaaaaah"), hashtag evaluation (p.e. detecting negative sentiment in #notverygood) and decomposition of the tweet in n-grams (to be able to distinguish "good" from "not good") [64]

- **Vader**: Directed for micro-blogging sentiment analysis, Vader uses sentiment lexicons of words, smileys and Internet acronyms and slang, validated by human annotators. Furthermore, it also evaluates the impact of punctuation and uppercase words using Mechanical Turk. All this is combined in a rule based system with polarity shifters and trigram analysis (for negation detection and amplification words). [55].

It is important to point out that some of the methods are solely based on the use of dictionaries and thus it is necessary to specify how the classification of the text will be done. Taking this into account, for the lexicon only approaches (AFINN, Emolex, EmoticonDS, NRC Hashtag, Emoticons, Opinion Lexicon, Panas-t, Sentiment 140, and SentiWordNet), iFeel uses Vader rule based system to push forward the performance of these lexicons [95].

The results obtained in terms of accuracy for the 6 domains are graphically shown in Figure 5.2. The blue bar represents our sentiment component (SC) and each histogram shows the accuracy of all systems in the specified domain.

As we see in Figure 5.2 our system is able to surpass (in general) some methods, such as Emoticon, Panas-t and Sann. The first two were more or less expected due to their limitations on the lexical resources. Furthermore, we can observe that Emoticon and Panas-t reach an accuracy close to 33% due to the neutral class. In other words, Emoticons and Panas-t classify all entries as neutral due to their reduced lexicons and, because class are balanced, they reach one third of the total accuracy. For a more clear analysis, another graph using the average F1-score of the classes in each dataset is presented in 5.3 . Since that the classes are balanced, the average F1-score for each system in each domain is calculated using the following formula except when the individual F1-score values cannot be calculated. In that case the average F1-score is 0.

$$F1_{Average} = \frac{F1_{Positive} + F1_{Neutral} + F1_{Negative}}{3}$$

Looking at Figure 5.3 it is clear that systems like Emoticon and Sann are unable to achieve results in some datasets. The absence of emoticons on the text causes the first system to be mistrust, whereas the limited lexical resources makes the second unreliable. For example, looking at the confusion matrix of Emoticon in the GOP dataset, the system failed to classify any tweet as negative. Therefore, a precision value and consequently the F1-score could not be calculated. Consequently, we assign a 0 value.

We can observe that our SC method does not achieve the best accuracy result in any of the datasets. In addition, it is surpassed in all datasets by several methods such as AFINN, SentiStrength Adapter and Umigon Adapter. So the use of any of these methods, in a 3 class sentiment classification, will be more accurate than the current method implemented in our SC. Furthermore, in terms of average F1-score the conclusion is similar considering that several systems outperform our method.

The same conclusion can be reached when analysing the average F1-score for each individual class, of each system in all domains, presented in Figure 5.4. Although it has relatively good results when comparing the positive and negative classes with other

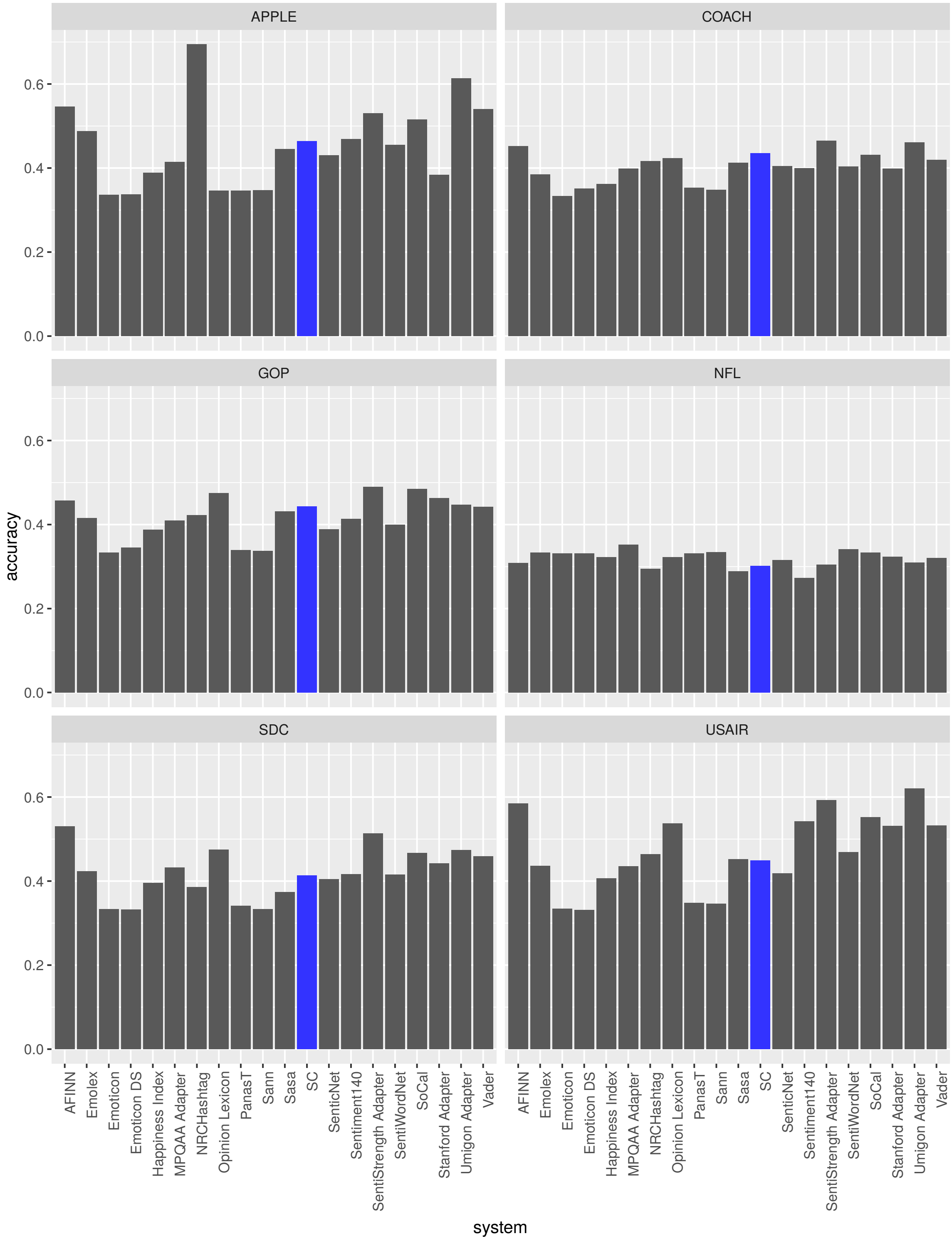


Figure 5.2: Accuracy comparison of our SC against other Sentiment Analysis systems by domain

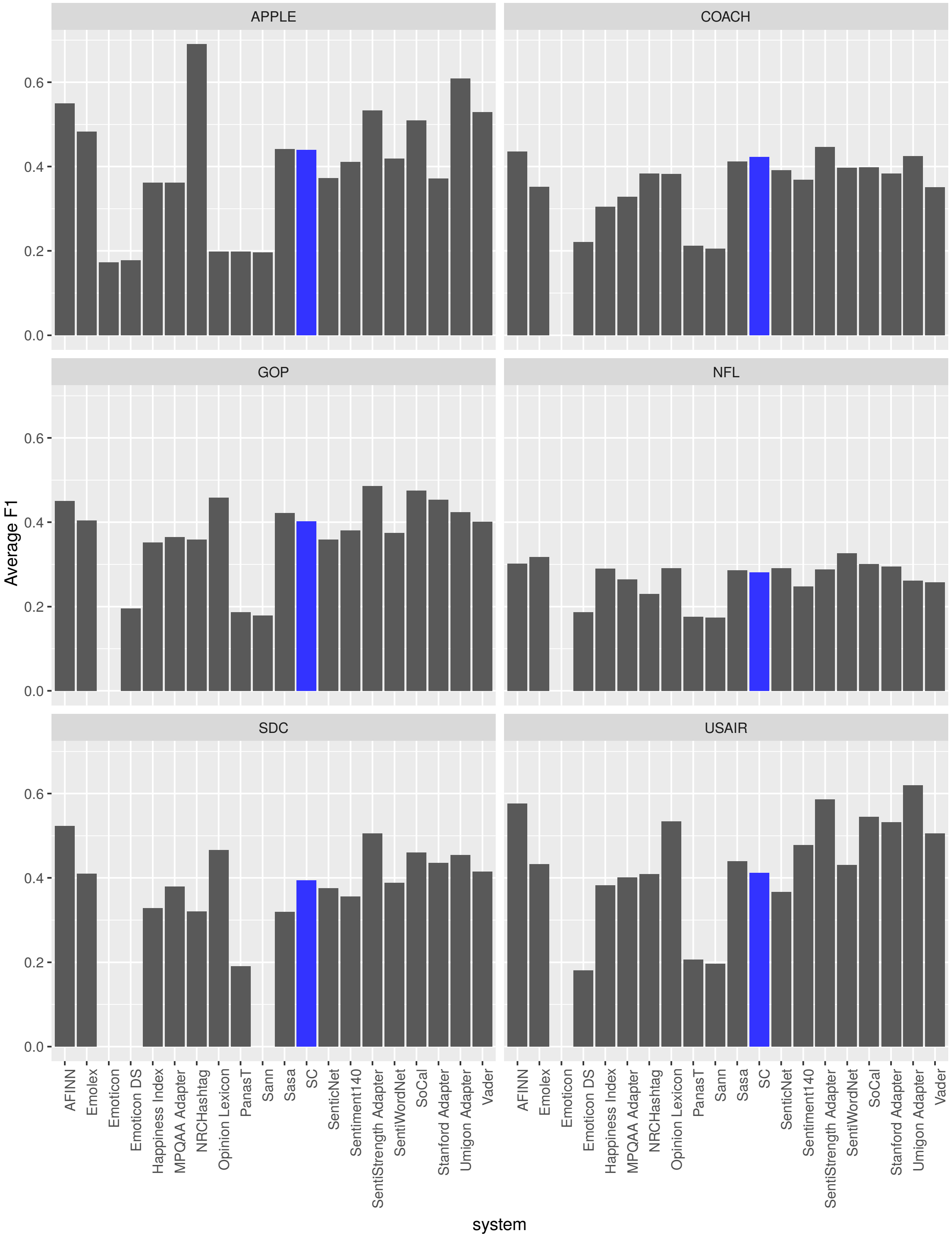


Figure 5.3: Average F1-score comparison of our SC against other Sentiment Analysis systems by domain

systems, the neutral class has low performance (it is in the bottom 3 systems). This is understandable since the system was originally directed to classify sentiment in a Positive/Negative scale. Therefore despite the good results in this type of classification, the current implementation becomes unreliable when we include the Neutral class.

This conclusion becomes clearer when we visualise and compare the accuracy in each individual class (concatenating all datasets). Figure 5.5 illustrates that comparison. Neutral class in our sentiment component has, in terms of accuracy, significantly poor results when comparing to the other systems. Although, some caution is needed when looking at these results. Panas-t, Sann, Emoticon and Emoticon DS have high accuracy on the neutral class but very low in the other two. Once again this is due to the limitation on the methods that classify almost every tweet as neutral.

5.2.2 Ensemble System Description

As it was mentioned before, the current approach doesn't achieve the best performance in a Negative/Neutral/Positive classification when compared to other sentiment systems, since it was originally built to perform a two class sentiment classification. Furthermore, according to our analysis, none of the systems stands out since different systems perform better in different domains. In consequence, to try to improve inter-domain performance, an ensemble system which takes into account each individual sentiment analysis procedure was implemented.

The ensemble sentiment classifier uses a decision making procedure where the score returned by each of individual methods is used as a vote for the assessment of the final sentiment. When a tie occurs, the rules for the final score assessment are the following:

- When there is a tie between positive and negative classes, the neutral sentiment is returned
- When there is a tie between the neutral class and other class, the other class is returned
- Since there are 19 sentiment systems, a 3-way tie is not possible. However, assuming different setups in terms of ensemble composition are possible, we define that in this case, the neutral value is returned

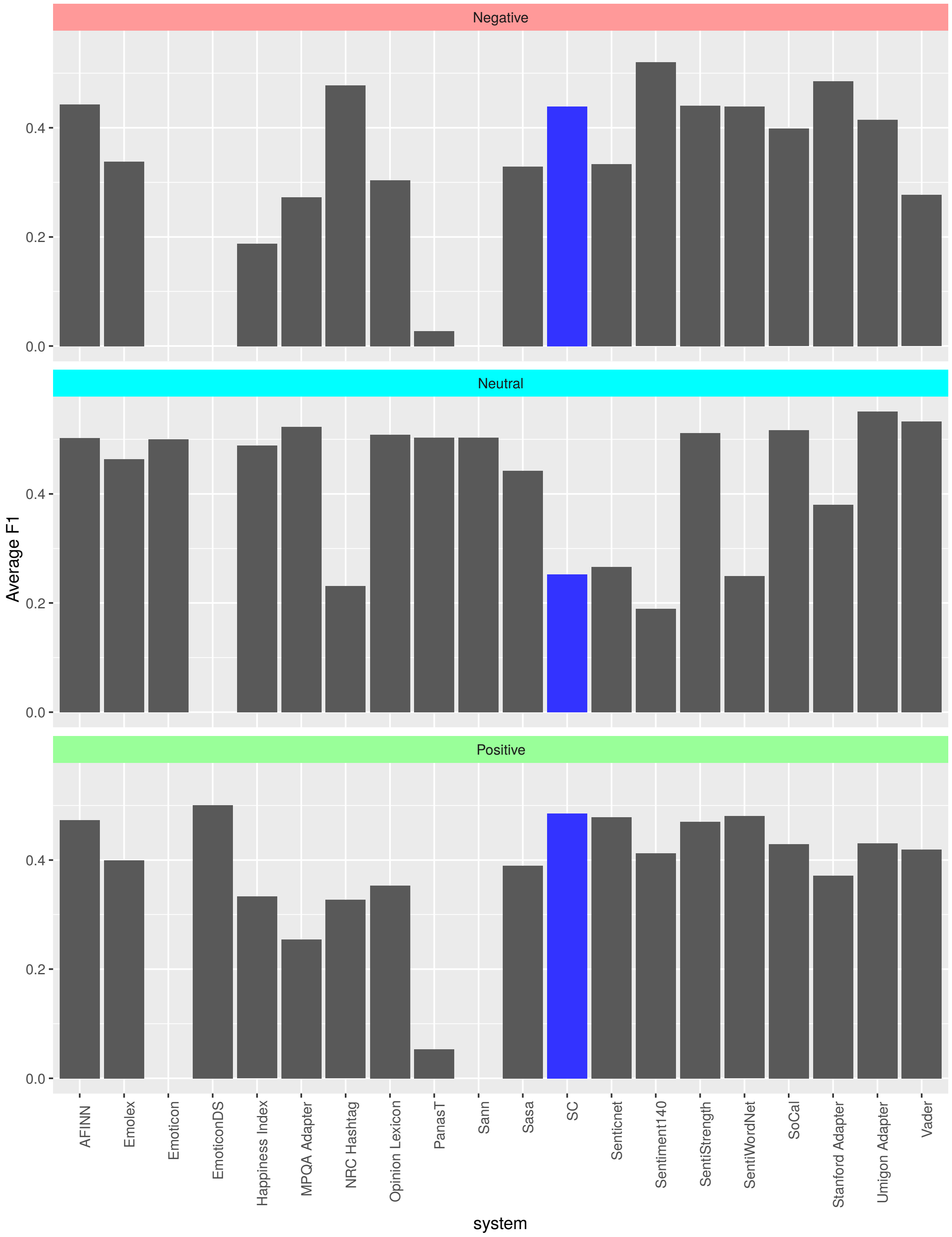


Figure 5.4: Sentiment Analysis Systems Comparison across different classes using F1-score

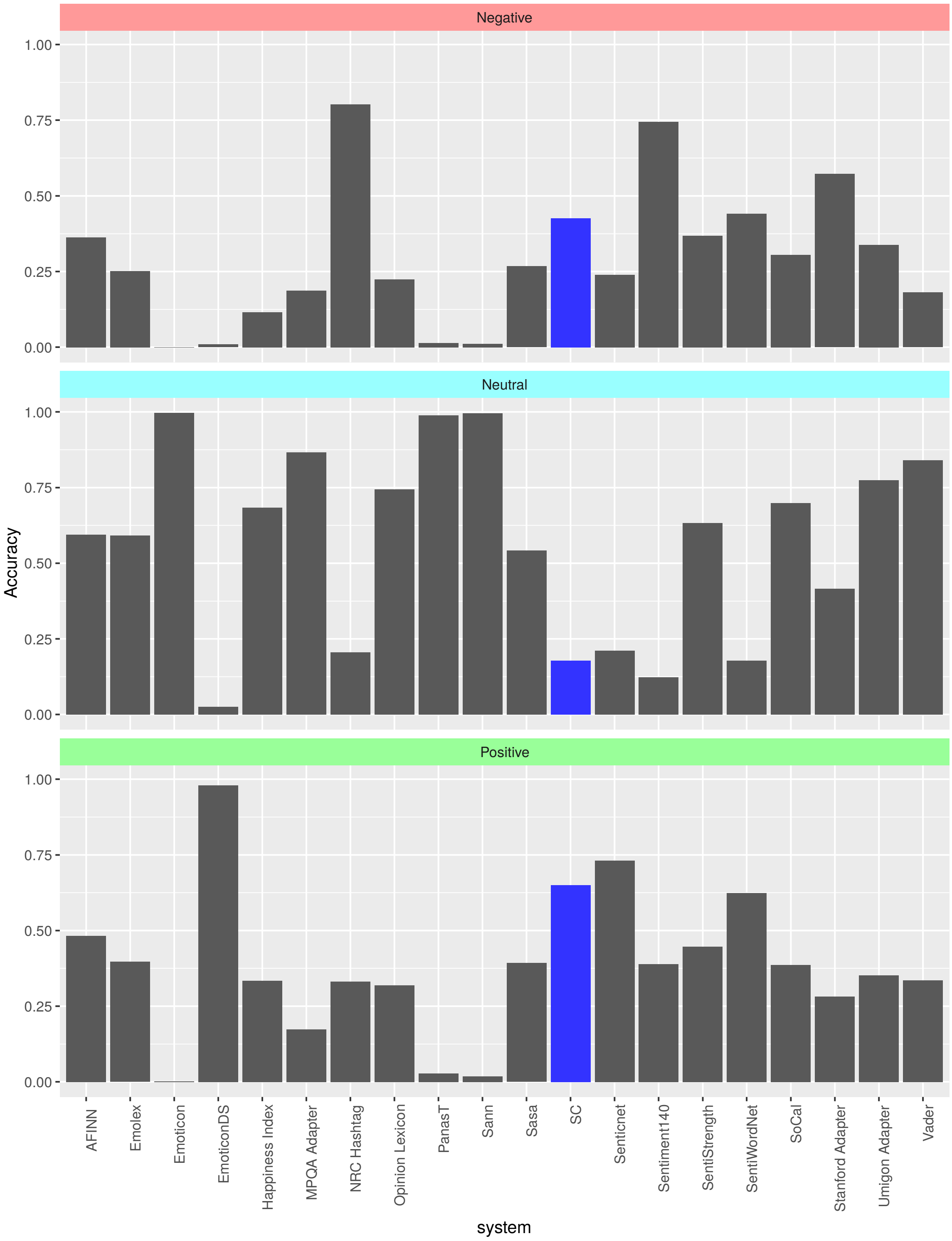


Figure 5.5: Sentiment Analysis Systems Comparison across different classes using Accuracy

An example of the behaviour of the ensemble system can be seen in Figure 5.6. Since our system is currently implemented in the R language [91] and the current version of IFeel is developed in Java [4], adaptations were necessary to incorporate IFeel in our workflow. In this context, we have developed an R package that provides an easier integration. We think that this is an important (although secondary) contribution of this thesis given the lack of sentiment tools in the R language. Additional information on the package is provided in the Appendix 8.3.

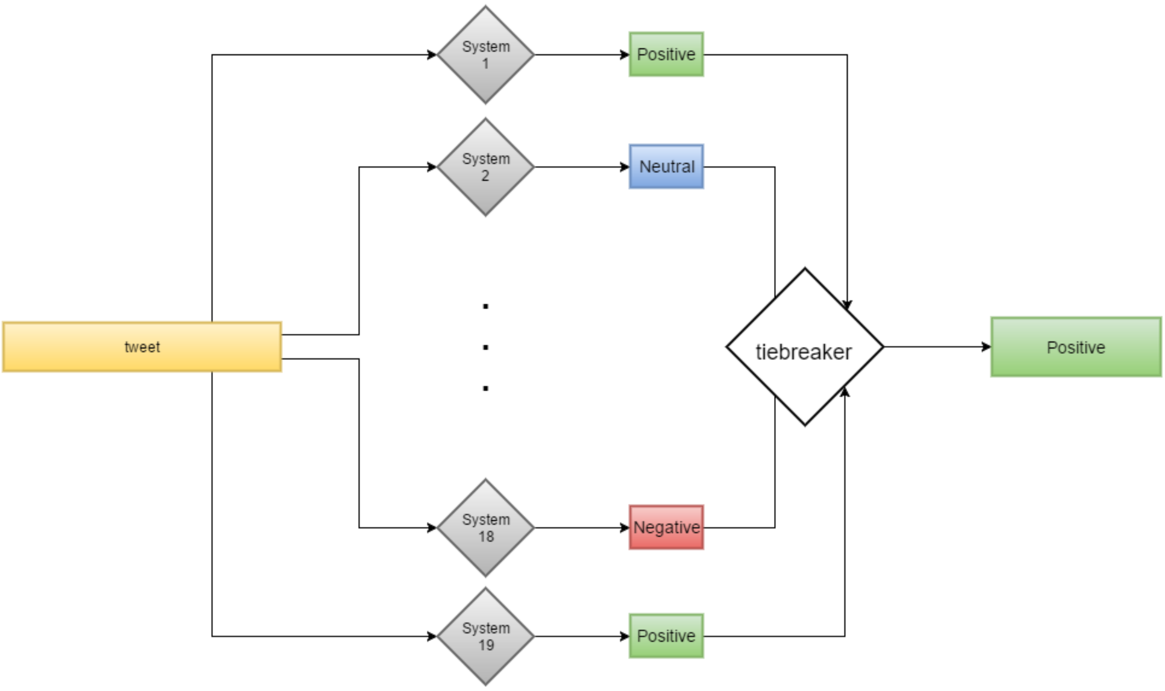


Figure 5.6: Ensemble system example

5.2.3 Ensemble System Evaluation

In order to check if the ensemble approach outperforms our current SC as well as other state of the art methods, we evaluate this method on the previously used tweet datasets. We have considered two different configurations of the ensemble approach. The first includes all methods on the IFeelR package. In the second configuration we exclude the ones using limited lexical resources (Emoticons and Panas-T). We assess our ensemble results in terms of accuracy and F1-score in different classes and different domains. The results obtained regarding the 19 (ENS19) and 17 (ENS17) ensemble variants are provided in below. First, we compare our ensembles with the top 3 more accurate systems in each domain and in the aggregation of all datasets. The results are presented in Table 5.1.

When compared with each stand alone system, the ENS19 has the best accuracy in the GOP dataset and it is on the top 3 in SDC, APPLE, USAIR and COACH datasets. Only in the NFL data it is below that mark. As for the ENS17 it is in the top 3 most accurate systems in all datasets excluding the COACH dataset (although the difference is of 0.1%). This ensemble does not achieve the best score in any of the datasets. However, if we look at the accuracy across all domains (in other words the average accuracy in all datasets) the ENS17 and ENS19 outperform the best individual systems.

Table 5.1: Comparison of ensemble method against the more accurate individual systems in each domain

Top Individual Systems	**Dataset Accuracy**						
(on each dataset)	**GOP**	**SDC**	**APPLE**	**USAIR**	**COACH**	**NFL**	**Average**
1st System	49.0	**53.0**	**69.5**	**62.0**	**46.5**	35.2	48.8
2nd System	48.1	51.3	61.3	59.3	46.1	34.1	48.3
3rd System	47.5	47.5	54.6	58.4	45.3	33.4	48.0
Ensemble Systems							
ENS17	51.0	51.6	60.4	60.0	45.2	**45.2**	**52.2**
ENS19	**51.6**	52.5	60.5	61.7	45.8	31.2	50.6

A similar analysis can be done separating the accuracy in Negative, Neutral and Positive classification. With this purpose, we will use the top 3 systems that are the more accurate in all domains (i.e. the individual systems that were selected for the "Average" column in 5.1). These systems are AFINN, SentiStrength and Umigon. We must reinforce that this type of analysis on the average is only possible due to the balanced datasets. The results regarding class accuracy can be examine in Table 5.2.

As we can observe, regarding classes, our ensemble systems are always in the top 3, when comparing with the most accurate individual methods. Furthermore, ENS19 and ENS17 achieve the highest accuracy value on the negative and positive class, respectively. Since the dataset are balanced in number of entries and number of elements in each class, it's no wonder that the all classes accuracy value are the same as the average ones in all datasets. Since accuracy can sometimes be misleading [109],

Table 5.2: Comparison of ensemble method against the most accurate individual systems.

Best Overall Individual Systems (using Accuracy as metric)	**Class Accuracy (%)**			
	Negative	**Neutral**	**Positive**	**Average**
Umigon	33.8	**77.3**	35.1	48.8
SentiStrength	36.8	63.2	44.7	48.3
AFINN	36.3	59.3	48.3	48.0
Ensemble Systems				
ENS17	35.8	72.3	**48.6**	**52.2**
ENS19	**38.25**	67.3	46.0	50.6

we perform the same analysis using the average F1-score in each domain. Therefore, selecting the top 3 systems according to the average F1-score and assessing the same metric with our method, results in the values presented in Table 5.3

Table 5.3: Comparison of ensemble method against the top individual systems (according to F1-metric) in each domain

Top Individual Systems (on each dataset)	**Dataset F1-score (%)**						
	GOP	**SDC**	**APPLE**	**USAIR**	**COACH**	**NFL**	**Average**
1st System	48.6	**52.3**	**69.1**	61.9	44.6	32.7	47.8
2nd System	47.5	50.6	60.9	58.6	43.5	31.8	47.2
3rd System	45.8	46.6	55.0	57.6	42.3	30.1	47.6
Ensemble Systems							
ENS17	50.3	50.8	60.4	59.4	42.2	**42.2**	**51.5**
ENS19	**51.2**	51.9	60.6	61.3	43.1	29.4	50.0

Once again it is clear that each of the ensemble systems perform well enough to be in the top 3 systems using F1-score in each dataset. Therefore, it is no surprise that, when considering all datasets, both ENS17 and ENS19 achieve an average F1-score superior than each of the individual systems.

Finally we take a closer look on the performance of the class classification using the concatenation of all datasets and the F1-score metric. Results are provided in Table 5.4. It is easily noticeable that both ensembles outperform the individual systems. The ENS17 in particular has the best performance in each class using the F1-score metric. In addition, it also has the best average F1-score value when concatenating all datasets. Furthermore, as accuracy is concerned, is only surpassed in 2 percent by

Table 5.4: Comparison of ensemble method against the top individual systems (according to F1-metric) in each class

Best Overall Individual Systems (using F1 measure as metric)	**Class F1-score (%)**			
	Negative	**Neutral**	**Positive**	**Average**
SentiStrength	44.0	51.2	48.2	47.8
AFINN	44.3	50.2	48.3	47.6
Umigon	41.8	54.7	45.2	47.2
Ensemble Systems				
ENS17	**46.6**	**55.6**	**52.2**	**51.5**
ENS19	46.3	53.6	50.1	50.0

ENS19. Therefore, based on these results, we decide to select ENS17 as the new SC method for our system.

5.2.4 New Sentiment Component Overview

Result on the previous subsection have showed that the ENS17 is the most suitable to integrate our system's sentiment component. However there is a trade off between the sentiment classification accuracy and the time to analyse a single tweet. In fact, using ENS17, a fragment of text takes 140 times more to be analyse than the original sentiment component of the system. So it is important to question if the improvements on classification, using ENS17, overcomes the time it takes to classify a sample of tweets. To reach a decision we must take into account that there is a limitation on the number of tweets that the Twitter Search API provides. If that number is reached then it is necessary to wait 15 minutes to retrieve a new sample. However, with the ENS17, that limit is never reached due to the workflow of our system. Since we retrieved the tweet corpus and start the sentiment analysis procedure before another extraction begins, the API limit is refreshed in each term extracted due to the time it take to analyse a sample. In addition, (although not present in our current implementation) this task can be improved with parallel computation since it can be easily divided in sub-tasks. For example a master-slaves paradigm [99] can be used to improve the efficiency of the ensemble, by dividing the sample of tweets by the slaves or assigning a sentiment method to each slave. Therefore, in our opinion, the time cost of the task can be easily reduced if this implementation is considered in the future.

The ENS17 does not completely substitute our SC. Since a set of the individual methods used in our ensemble already consider emoticons, we decided to removed them of our posterior analysis with the goal of not repeating, and therefore increasing the importance of this non textual feature. However, emojis are not consider in any of the ENS17 systems. Therefore the score for each individual tweet (and based on the evaluation made in chapter 3) is determine by the following formula:

$$\text{score}_t = \text{score}_{ENS17} * 0.75 + \text{score}_{emojis} * 0.25$$

where the score of emojis is simple calculated by simple average (sum of the sentiment of emojis divided by the number of occurrences).

We also adapt the calculus for each individual term sentiment. We defined in Chapter 3 that the previous term score was the result of averaging on the sentiment of tweets. This score was good for a two class sentiment analysis since the probability of a term having a neutral value was significantly low. However, since now we are evaluating in three classes, the same score calculation cannot be applied. Therefore, in our new SC we calculate a different score for each class using the following formula:

$$\text{score}_c = \frac{\text{number of tweets classified as c}}{\text{total number of tweets}}$$

where c is the respective the sentiment class. The class with the highest score is returned as the sentiment class of the term.

Is our opinion that the new features implemented in our system (based on the evaluation made previously and described in this chapter) will improve the classification on the sentiment of the extracted terms. We do not feel that a new evaluation on the terms is strictly necessary since we demonstrate that: 1) regarding the sentiment component, the tweet sentiment classification has surpass the previous method (therefore is our assumption that can only improve the term sentiment classification) and 2) domain disambiguation may help to separate eventual duplicated terms (instead of not classifying them depending on the domain). Finally, it is time to assess if the dictionaries retrieved can improve the performance on sentiment analysis methods.

Chapter 6

Preliminary Lexicons Evaluation

The final stage of this work is to determine if the dictionaries built with the proposed system (which was described in the previous chapters) can improve the sentiment analysis task in short informal texts. To answer our research question, we used a dataset that contains posts and comments from Facebook and tweets from September 7th to September 14th, evaluated with sentiment on Crowdflower. This dataset was extracted for the REMINDS project whose main goal is to automatically identify journalistic relevance in social network posts [93]. The dictionaries from our system were retrieved from September 4th to September 6th. This way, we guarantee that the tweets and Facebook posts used to create the dictionaries were not included in the dataset where we performed the evaluation.

The dataset combines 3 types of short informal texts: Facebook posts, Facebook comments and tweets. The Facebook posts and comments were retrieved from the top most popular pages in different categories from the United States, according to the LikeAlyzer tool [65], and whose names we present in Table 6.1. For each post on the defined time interval, we extracted a maximum of 20 comments (ordered by the Facebook relevance metric). From that extraction, a sample of 1000 comments and 3995 posts were sent to Crowdflower for evaluation.

Regarding the tweets extraction, relevant topics (which appeared on recent news) were used on the Search API due to the REMINDS project requisites. Consequently, the results obtained may be skewed due to the conflicts of the keywords with the terms of our lexicons. In other words, some terms used as query were evaluated with

Table 6.1: Pages used for Facebook posts and comments

Page Name	
1. Young Entrepreneur	2. U.S. News and World Report
3. Tech Viral	4. USA TODAY
5. NaturalNews.com	6. Game Informer
7. The University of Texas at Austin	8. CNN
9. Business Insider	10. The Huffington Post
11. Fortune Magazine	12. NBC News
13. TED	14. The Wall Street Journal
15. Fox News	16. New York Post
17. The New York Times	18. Los Angeles Times
19. Mashable	20. Daily Wire
21. Washington Post	22. SparkPeople.com
23. MTV News	24. Seeker Daily
25. The Scientist	26. Tampa Bay Times
27. One Green Planet	28. Harvard University
29. Washington Post Opinions, Outlook and PostEverything	

sentiment in our lexicons. Consequently to avoid biased results, we exclude those from the dictionaries. The keywords used were the following:

- terrorism
- refugees
- elections
- paralympic
- champions league
- emmys
- wall street

For each keyword 714 tweets were extracted forming a total of 4998 tweets. Concatenating this data with the one extracted from Facebook, we have a final dataset with 9993 entries of short informal texts for evaluation.

The experiment on Crowdflower was conducted differently from the previous one, since we only had one evaluation per entry. This can lead to a weaker "ground truth" but to an higher number of texts evaluated. The decisions on taking this approach was due to the requisites of project REMINDS. The sentiment question asked to the workers was *"The sentiment expressed in this text is:"* To answer, a likert scale ranging from 1 to 5 and labelled from "very negative" to "very positive" was presented. In addition, we also included a follow up question: *"Choose (from the provided text) the word that best supports your previous answer"*. Our goal was to force the worker to take a more careful decision, justifying it. Furthermore, by having the set of terms that were used to decide on the sentiment of the text, we can have a "preview" if our dictionaries will be useful for the sentiment classification task on this particular dataset. Specifically, after the experiment was over, we analysed how many of the workers had answered the second question with terms that were included in the lexicon retrieved from our system. In the total of 9993 entries we only verify 415 in that condition, which corresponds to approximately 4% of the sample.

This early analysis provides evidence on the limitations of the dataset for evaluating our method. In fact, opinion words are frequently present in the justification answer for the sentiment classification. Therefore, the traditional sentiment lexicons should provide good results in the majority of the dataset.

When concatenating the lexicons from AFINN, Opinion Lexicon, NRC Hashtag, MPQA, Vader and SentiWord we conclude that 4948 have a word of those lexicon as justification. Since our goal is to assess that not all tweets that lack opinion words are neutral, such high presence of opinion words as justification of sentiment classification may negatively affect our analysis. .

6.1 Factual and Non-Factual Text

The first approach in our evaluation analysis was to add the outputted lexicon from our system to other sentiment methods in the entire REMINDS dataset.

In Section 5.2 when comparing with our ensemble sentiment system, we determine the best overall individual methods on the tweets datasets tested (as Table 5.4 and 5.2 show). Therefore, we modify and create 7 different IFeelR packages (one for each domain) where we added our lexicons to those 3 specific sentiment methods.

To decide on which lexicon to use in each dataset entry, we need to fit the text in one of the domains previously defined (world, sports, entertainment, politics, business, technology and health). In tweets, we could easily make the match between the keyword used and the domain. However, with Facebook posts and comments, the task is not so easy, since we retrieve them from the most popular pages according to LikeAlyzer. Consequently, we do not have each entry categorised with a domain. Therefore, to standardise our domain selection in this experiment, we apply the disambiguation process that we refer in Section 5.1 and added the words that occur in our sentiment lexicons to each domain, in order to improve accuracy. For the entries that no domain was found, we assigned the "world" value. Finally, we scale the sentiment classification on Crowdflower to Negative,Neutral or Positive values to match AFINN, SentiStrength and Umigon score representation.

In Table 6.2 we can see the variations on accuracy and F1-score from the addition of our sentiment lexicons to the previous mentioned methods. A positive value states an improvement with the addition of our sentiment lexicons, whereas a negative value shows a decrease. The results reveal no major changes between the addition of our lexicons to the default ones. In fact, on some cases, the results were worse when using these new lexicons.

Table 6.2: Variation between the Sentiment Systems with and without the Expanded Lexicons

Sentiment System	**Accuracy %**	**Average F1%**
AFINN	-0.81	+0.01
SentiStrength	+0.19	-0.61
Umigon	-1.26	+0.89

These results are the confirmation that our lexicons cannot be used as regular sentiment dictionaries due to the specificity of the problem we are trying to solve. As a matter of fact, using these concatenated with the more traditional approaches in datasets containing all types of short informal texts will probably lead to worst results since we are, amongst other terms, classifying entities. Consequently, all texts which are simply facts (e.g. "Hillary was in Ohio") will be classified, by the majority of the systems containing the expanded lexicons, with the polarity associated with the entity. In addition, it is important to notice that neutral classification texts can be factual texts (p.e. "Real Madrid plays tomorrow against Barcelona") or texts which point out to a neutral opinion (p.e. "The movie has good actors but bad plot"). By analysing manually a sample of entries and due to the main goal of the extraction of the REMINDS dataset (i.e. find newsworthy information), we found evidence that the majority of neutral cases fit the non-factual category, consequently justifying the poor results achieved.

6.2 Non-Factual Texts

One of the major problems that consequently lead to the outcomes shown in Table 6.2 was the fact that a large percentage of the sample was composed by factual texts (i.e. texts that express facts and not opinions). We do believe that our lexicons should only be applied after a process of subjectivity detection to distinguish between facts and opinion texts. For example, looking for cues in the text that express opinion ("in my opinion", "i think"...) or a machine learning approach like the one presented in [129]. However, this detection is not on the scope of this work and therefore, we excluded the factual texts by removing the neutral classified entries.

The results of our second experiment are presented in Table 6.3.

Table 6.3: Variation between the Sentiment Systems with and without the Expanded Lexicons in non factual text

Sentiment System	**Accuracy %**	**Average F1%**
AFINN	+1.12	+0.48
SentiStrength	+1.36	+1.43
Umigon	+3.14	+2.63

On this subset of our data, which in our opinion better represents the problem we are trying to tackle, the addition of the lexicons outputted by our system improved in both accuracy and F1-score. Umigon is the system that benefits the most on the addition of these lexicons and AFINN the less. The average accuracy improvement is around 1.87% whereas the F-measure is 1.51%.

Although it is not a major difference between both sentiment dictionary approaches (traditional and traditional + expanded) it is a steady improvement since it consistent across all 3 systems analysed.

We can go further in our analysis of non factual texts and restrict our dataset to the entries whose response to the question *"Choose (from the provided text) the word that best supports your previous answer"* was included in our expanded sentiment lexicon. The filtered dataset contains 215 entries and results of these specific cases can be consulted in Table 6.4.

Table 6.4: Variation between the Sentiment Systems with and without the Expanded Lexicons in non factual text with sentiment justification word in the Expanded Lexicons

Sentiment System	**Accuracy %**	**Average F1%**
AFINN	+2.23	+2.31
SentiStrength	+23.13	+9.81
Umigon	+24.11	+12.55

Although we are forcing the word for the sentiment justification to be present in our dictionary (and therefore imposing the condition that it will be used for the text sentiment evaluation), this analysis intends to show that, in specific cases of short informal opinion texts where the argument to infer the sentiment is not on traditional lexicons, using our system do expand the dictionaries can result in an reasonable improvement. In fact, SentiStrength and Umigon have an accuracy boost superior to 20% whereas the F1-score increases 9.81% and 12.55% respectively. This demonstrates that not only is important to consider our system sentiment dictionaries but also that our term sentiment analysis (despite the modifications that were introduced in Chapter 5) is still capable of accurately classify the terms since, if that was not the case, the results on this specific dataset would not improve regarding the traditional sentiment lexicons.

6.3 Results Overview and Discussion

On the previous sections we evaluated the performance of our expanded lexicons. The first attempt was to directly add them to AFINN, SentiStrength and Umigon methods and evaluate on the entire REMINDS dataset. This lead to a conclusion that was more or less expected: the expanded lexicons cannot be used as normal sentiment dictionaries. The main justification is that all the factual texts which included terms from our lexicons would be subject to a Positive/Negative classification. For example texts regarding facts that involved entities like "Donald Trump", "Hillary Clinton" and "Facebook" would all be evaluated based on the current public opinion of those terms. This analysis combined with the fact that the percentage of tweets classified as neutral in our dataset is 49%, provides plausible explanation for the results presented in Table 6.2.

Having this in consideration, we reinforce that our lexicon might still be useful but in a more complex sentiment system workflow. First we need to make sure that the text under analysis is subjective and only then apply the necessary combined lexicons. To simulate that environment, and in the absence of a subjectivity detection tool, we excluded the factual texts from our dataset by selecting the entries only with a positive or negative polarity. We are aware that this is not the best method to select opinion texts. However, we must consider that we are using a dataset whose main purpose was to extract tweets and Facebook newsworthy posts. In addition, the Facebook pages (presented in Table 6.1) where the comments and posts were extracted are strongly associated with news sources. Consequently, it is fair to assume that the vast majority of the posts retrieved will be facts and not opinions. Furthermore, since the Twitter Search API was used, (and as it was previously mentioned on Table 3.3, it looks for relevance) the probability of tweets or retweets regarding news is higher (even more when no URL filter like the one in our system was applied).

When we remove the non-factual posts from the complete REMINDS dataset, the accuracy and F1-score improved in all selected sentiment methods with the addition of our lexicons. This answers our second research question. In fact, the addition of domain and time specific lexicons can help in sentiment analysis classification by generally improving the accuracy and average F1-score in positive and negative classification. Although the differences are not substantial, we do believe this is a small advance towards a better performance on the sentiment classification task. However, we are conscious of the limitation of our analysis and aware that the importance of these lexicons is dependent on a good subjectivity detection tool.

Chapter 7

Conclusion

7.1 Review/Synthesis

In this work we assess the viability of using public opinion sentiment lexicons to improve current sentiment analysis methods. Consequently, we began by assessing the capability of Twitter as a good source to assess the sentiment of relevant and time-domain dependent terms. Although prior state of art has provided evidence that this social network performed good on tracking the sentiment of a specific topic or event, we assess a broader approach by having sets of relevant terms in specific domains and whose presence was not constant. Therefore, to answer this question, we develop a system for automatically extracting the more relevant terms from seven different domains and classify their sentiment in a positive/negative scale. With that purpose, our system retrieved the more frequent terms from news headlines using RSS feeds from several news sources and then queried Twitter with the same terms, assessing the polarity using the average sentiment classification obtained from a tweets sample.

We proceed to evaluate three parts of our system which we considered the most important for the creation of domain and time specific sentiment lexicons: term extraction, tweet sentiment analysis and the sentiment on each extracted term. Experiments shown that the proposed term extraction component is rather effective, achieving a 90% accuracy on the fitness of time and domain. In addition our sentiment classifier (which is crucial for the final sentiment of terms) also produced satisfactory results in detecting the polarity of tweets from several different domains. Tests on classified Twitter datasets achieved an overall accuracy ranging from 61.26% (GOP debate dataset) to 77.31% (Apple dataset). Finally, experimental term classification results using Crowdflower lead to an overall accuracy of 79.67% with positive terms achieving better F1-score (82.86%) than the negative ones (75.00%).

However, there was one major limitation that was revealed by the Crowdflower experiment. A high number of terms were classified as neutral by the workers, which lead to an adaptation of our system from a Positive/Negative term classification to a Positive/Neutral/Negative one. For that purpose, we built an ensemble tweet sentiment classification system (entitled ENS17) that provided superior performance compared with 19 state of the art classification methods, in different domain datasets. We also added a procedure to disambiguate the domains on tweets for the cases of duplicated terms (p.e. "apple" can be included in both technology and health categories).

The final part of the work was to test the lexicons generated by our system. With that goal in mind we used the REMINDS dataset which included tweets, Facebook posts and comments. Our analysis lead to the conclusions that state of the art sentiment systems can benefit from our lexicons for a better classification, although dependent on a good subjectivity detection method. Experiments on opinion texts with a positive or negative sentiment lead to an increase on both accuracy and average F1-score with the addition of the sentiment lexicons provided by our system. Although the increase on both these metrics are not tremendous, they are stable since they improve the three sentiment methods tested.

7.2 List of Contributions

During the course of this work, there were three main contributions that enrich the current state of the art on sentiment analysis.

- A system for extract and classify domain and time dependent terms was developed. A paper with the description and evaluation of the system was submitted and accepted in the 8th International Conference on Knowledge Discovery and Information Retrieval (KDIR 2016). The article is included in Appendix 8.4.1.

- The development of an R packages for sentiment analysis. This package calls the sentiment methods implemented in IFeel framework and allows the user to define an ensemble sentiment system (with user-defined sentiment methods) based on majority voting. It also provides a function that displays a confidence score one each sentiment class based on the same methods. The IFeelR package documentation is included in Appendix 8.3 and the package, due to licence restrictions, can be made available by request.

- An evaluation on the importance of time and domain dependent sentiment lexicons on the sentiment evaluation of short informal texts. A full paper with these results was submitted to Social Network and Media Analysis (SONOMA) track on the 32nd ACM Symposium on Applied Computing. Please refer to Appendix 8.4.2 for the full paper.

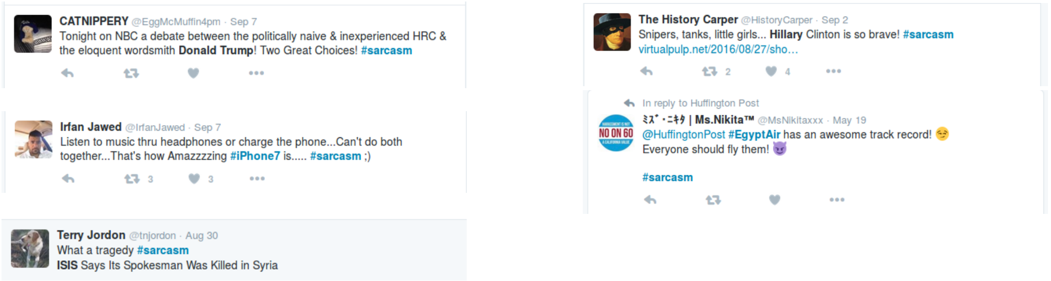
7.3 Future Work

There are several paths we can follow in future work. One of the limitations of our term extraction method is related to the accuracy of the adopted NLP classifier that can lead to some noisy unigram terms. Future research will try to improve it by exploring more filters for a fine grain selection of unigrams in different domains. Possible filters may include the use of different NLP classifiers to determine the part of speech tags and name entity recognition techniques to relate terms that are referring to the same entity (e.g. "Obama" and "POTUS"). We also plan to uncover the connections 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.

With the addition of an ensemble sentiment system, the time to extract and classify a set of terms has largely increased. Therefore, we intend to reduce it by using parallel computing techniques. Consequently, we can easily extend the sample of tweets extracted, increasing the precision on the sentiment of each term.

Based on the results achieved in the previous chapter, automatically classifying texts as facts or opinions can also be a path to follow. By integrating this step, we could build a workflow and evaluate the performance of our lexicons without any previous text filtering.

We also intend to apply these lexicons to the specific problem of detecting irony and sarcasm in tweets. Our assumption is that our system can facilitate the classification of sarcasm by detecting discrepancies between the sentiment of traditional opinion words sentiment and the terms extracted. We suppose this will be more visible in entities that have a clear negative/positive public opinion. For example, considering the tweets in Figure 7.1 we, as humans, can clearly see the presence of sarcasm due to our knowledge of the public opinion regarding the entities presented. Therefore, detecting opposite polarities between the sentiment of the entity (in our sentiment lexicon) and other opinion words, can be seen as clue for sarcasm detection. However, this observation alone is not enough and other studied features should also be applied.



CATNIPPERY @EggMcMuffin4pm · Sep 7
Tonight on NBC a debate between the politically naive & inexperienced HRC & the eloquent wordsmith **Donald Trump**! Two Great Choices! #sarcasm

Irfan Jawed @IrfanJawed · Sep 7
Listen to music thru headphones or charge the phone...Can't do both together...That's how Amazzzzing #iPhone7 is..... #sarcasm ;)

Terry Jordon @tnjordon · Aug 30
What a tragedy #sarcasm
ISIS Says Its Spokesman Was Killed in Syria

The History Carper @HistoryCarper · Sep 2
Snipers, tanks, little girls... **Hillary** Clinton is so brave! #sarcasm
virtualpulp.net/2016/08/27/sho...

In reply to Huffington Post
ﾐｽﾞ・ﾆｷﾀ | Ms.Nikita™ @MsNikitaxxx · May 19
@HuffingtonPost #EgyptAir has an awesome track record! 😏
Everyone should fly them! 😈

#sarcasm

Figure 7.1: Sarcastic tweets (retrieved using the hashtag #sarcasm)

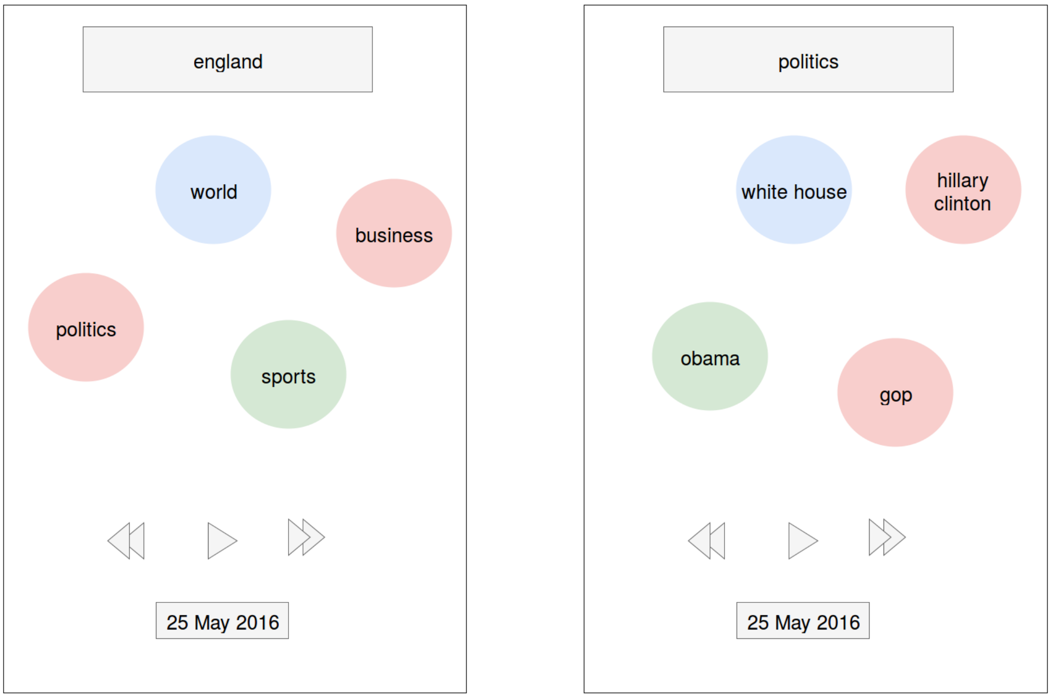
In addition, we expect to use the system, as well as the data extracted by it through the course of this thesis, in other applications besides sentiment lexicons expansion. In fact, since our system can retrieve the sentiment regarding entities, events and other terms that are relevant in the news, we can make a more detailed study on the approach described in Section 4.4 and try to uncover the relations about the news of a certain term, its sentiment changes and its "trending" variation.

We can also use this data to build a data mining visualisation system with features like:

- tracing plots based on sentiment changes of the terms through time and their trending factor (as it was already visualised in section 4.4)
- visualise inter-domain terms and represent their associated sentiment for each domain through time (for example like Figure 7.2a shows)

- visualise the terms for each domain and their sentiment variation through time (as Figure 7.2b shows)

This can be particularly useful for data scientists. By interacting with the data, they can explore and expose possible relations between terms (for example relations between the united states presidential candidates).



(a) (b)

Figure 7.2: Prototype of a visualization tool

Other future approaches (out of the scope of sentiment lexicon expansion) with the current system and data extracted can explore the possibility of predicting the public sentiment of an entity, given a news related to it. For example, given the sentiment values of the term "ISIS" as well as the corresponding news through time as features, we can try to assess if the headline "ISIS launches new offensive in Deir Ezzor" will translate into a positive or negative public opinion. Once again, the data alone may not be enough and other procedures like sentiment analysis on the news and natural language processing could be required.

Summarising, there are several improvements that still can be made to the current system, in the term extraction procedure and tweet sentiment analysis. We do also think that the system and data it retrieves can have several applications beyond sentiment lexicons, whether they are related to visual data mining or supervised approaches for sentiment prediction and sarcasm detection. Consequently, there are several choices we can take to extend this work and we look forward to investigate them in a near future.

Bibliography

[1] Ahmed Abbasi, Ammar Hassan, and Milan Dhar. Benchmarking twitter sentiment analysis tools. In *LREC*, 2014.

[2] Amazon. Amazon mechanical turk. https://www.mturk.com/mturk/welcome, 2016. Acessed: 2016-08-21.

[3] Sihem Amer-Yahia, Samreen Anjum, Amira Ghenai, Aysha Siddique, Sofiane Abbar, Sam Madden, Adam Marcus, and Mohammed El-Haddad. 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, 2012.

[4] Ken Arnold, James Gosling, and David Holmes. *The Java Programming Language*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 3rd edition, 2000.

[5] Farzindar Atefeh and Wael Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164, 2015.

[6] L. Augustyniak, T. Kajdanowicz, P. Szymański, W. Tuligłowicz, P. Kazienko, R. Alhajj, and B. Szymanski. Simpler is better? lexicon-based ensemble sentiment classification beats supervised methods. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 924–929, Aug 2014.

[7] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors,

Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[8] Adam Bermingham and Alan F. Smeaton. On using twitter to monitor political sentiment and predict election results, 2011.

[9] J.R.L. Bernard. *The Macquarie thesaurus.* Herron Publications West End, Qld, new budget ed. edition, 1987.

[10] Albert Bifet, Geoffrey Holmes, Bernhard Pfahringer, and Ricard Gavaldà. Detecting sentiment change in twitter streaming data. In *Proceedings of the Second Workshop on Applications of Pattern Analysis, WAPA 2011, Castro Urdiales, Spain, October 19-21, 2011*, pages 5–11, 2011.

[11] André Blais, Elisabeth Gidengil, and Neil Nevitte. Do Polls Influence the Vote? The University of Michigan Press, 2001.

[12] RSS Board. Rss 2.0 specification. http://www.rssboard.org/rss-specification, 2016. Acessed: 2016-05-23.

[13] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[14] M. M. Bradley and P. J. Lang. Affective norms for English words (ANEW): Stimuli, instruction manual, and affective ratings. Technical report, Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, 1999.

[15] Felipe Bravo-Marquez, Eibe Frank, and Bernhard Pfahringer. 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, 2015. ACM.

[16] Felipe Bravo-Marquez, Eibe Frank, and Bernhard Pfahringer. 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, 2015.

[17] Felipe Bravo-Marquez, Daniel Gayo-Avello, Marcelo Mendoza, and Barbara Poblete. Opinion dynamics of elections in twitter. In *Eighth Latin American Web Congress, LA-WEB 2012, Cartagena de Indias, Colombia, October 25-27, 2012*, pages 32–39, 2012.

[18] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[19] Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. 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, 2014.

[20] Lu Chen, Wenbo Wang 0002, Meenakshi Nagarajan, Shaojun Wang, and Amit P. Sheth. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, editors, *ICWSM*. The AAAI Press, 2012.

[21] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[22] Crowdflower. Introducing contributor performance levels. http://crowdflowercommunity.tumblr.com/post/80598014542/introducing-contributor-performance-levels, 2014. Acessed: 2016-04-10.

[23] CrowdFlower. Crowdflower: Make your data useful. https://www.crowdflower.com/, 2016. Acessed: 2016-08-21.

[24] Crowdflower. Data for everyone. http://www.crowdflower.com/data-for-everyone/, 2016. Acessed: 2016-04-10.

[25] Geng Cui, Hon-Kwong Lui, and Xiaoning Guo. The Effect of Online Consumer Reviews on New Product Sales. *International Journal of Electronic Commerce*, 17(1):39–58, 2012.

[26] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 241–249, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[27] Nicholas A. Diakopoulos and David A. Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1195–1198, New York, NY, USA, 2010. ACM.

[28] Xiaowen Ding, Bing Liu, and Philip S. Yu. A holistic lexicon-based approach to opinion mining, 2008.

[29] Daniel Dor. On newspaper headlines as relevance optimizers. *Journal of Pragmatics*, 35(5):695 – 721, 2003.

[30] Weifu Du, Songbo Tan, Xueqi Cheng, and Xiaochun Yun. 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, 2010. ACM.

[31] Mohamed Elarnaoty, Samir AbdelRahman, and Aly Fahmy. A machine learning approach for opinion holder extraction in arabic language. *CoRR*, abs/1206.1011, 2012.

[32] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, pages 417–422, 2006.

[33] Feedly. 60,000 pro subscribers and what to expect next. https://blog.feedly.com/60000-pro-subscribers-and-what-to-expect-next/, 2015. Acessed: 2016-05-23.

[34] Feedly. feedly: your work newsfeed. https://play.google.com/store/apps/details ?id=com.devhd.feedly, 2015. Acessed: 2016-05-23.

[35] Christiane Fellbaum, editor. *WordNet: an electronic lexical database.* MIT Press, 1998.

[36] Song Feng, Ritwik Bose, and Yejin Choi. 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, 2011. Association for Computational Linguistics.

[37] Alice Foster. Terror attacks timeline: From paris and brussels terror to most recent attacks in europe. http://www.express.co.uk/news/world/693421/Terror-attacks-timeline-France-Brussels-Europe-ISIS-killings-Germany-dates-terrorism, 2016. Acessed: 2016-08-21.

[38] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In Lorenza Saitta, editor, *Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996)*, pages 148–156. Morgan Kaufmann, 1996.

[39] Daniel Gayo-Avello, Panagiotis Takis Metaxas, and Eni Mustafaraj. Limits of electoral predictions using twitter. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press, 2011.

[40] M. Ghiassi, J. Skinner, and D. Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Syst. Appl.*, 40(16):6266–6282, November 2013.

[41] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *In Proceedings of the Twenty-eight International Conference on Machine Learning, ICML*, 2011.

[42] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6, 2009.

[43] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12, 2009.

[44] Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha. Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks*, pages 27–38. ACM, 2013.

[45] Pollyanna Gonçalves, Fabrício Benevenuto, and Meeyoung Cha. Panas-t: A pychometric scale for measuring sentiments on twitter. *CoRR*, abs/1308.1857, 2013.

[46] Aniko Hannak, Eric Anderson, Lisa Feldman Barrett, Sune Lehmann, Alan Mislove, and Mirek Riedewald. Tweetin ' in the rain: Exploring societal-scale effects of weather on mood.

[47] Vasileios Hatzivassiloglou and Kathleen R. McKeown. 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, 1997. Association for Computational Linguistics.

[48] Devon Haynie. The u.s. and u.k. are the world's most influential countries, survey finds. www.usnews.com/news/best-countries/best-international-influence, 2015. Acessed: 2016-05-23.

[49] Alex Hern. Don't know the difference between emoji and emoticons? let me explain. https://www.theguardian.com/technology/2015/feb/06/difference-between-emoji-and-emoticons-explained, 2016. Acessed: 2016-05-09.

[50] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, Aug 1998.

[51] Alexander Hogenboom, Danella Bal, Flavius Frasincar, Malissa Bal, Franciska De Jong, and Uzay Kaymak. Exploiting emoticons in polarity classification of text. *J. Web Eng.*, 14(1-2):22–40, March 2015.

[52] Alexander Hogenboom, Daniella Bal, Flavius Frasincar, Malissa Bal, Franciska de Jong, and Uzay Kaymak. Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 703–710, New York, NY, USA, 2013. ACM.

[53] Minqing Hu and Bing Liu. 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, 2004. ACM.

[54] Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, pages 755–760. AAAI Press, 2004.

[55] Clayton J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eytan Adar, Paul Resnick, Munmun De

Choudhury, Bernie Hogan, and Alice H. Oh, editors, *ICWSM*. The AAAI Press, 2014.

[56] Xiang Ji, Soon Ae Chun, and James Geller. Monitoring public health concerns using twitter sentiment classifications. In *IEEE International Conference on Healthcare Informatics, ICHI 2013, 9-11 September, 2013, Philadelphia, PA, USA*, pages 335–344, 2013.

[57] Fei Jiang, Yiqun Liu, Huanbo Luan, Min Zhang, and Shaoping Ma. *Social Media Processing: Third National Conference, SMP 2014, Beijing, China, November 1-2, 2014. Proceedings*, chapter Microblog Sentiment Analysis with Emoticon Space Model, pages 76–87. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[58] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 151–160, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[59] Yohan Jo and Alice H. Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 815–824, New York, NY, USA, 2011. ACM.

[60] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[61] Soo-Min Kim and Eduard Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, pages 1–8, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[62] Michel Krieger and David Ahn. Tweetmotif: exploratory search and topic summarization for twitter. In *In Proc. of AAAI Conference on Weblogs and Social*, 2010.

[63] Huey Yee Lee, Epulze Sdn Bhd, Block C, Ehsan Malaysia, and Hemnaath Renganathan. Chinese sentiment analysis using maximum entropy, 2011.

[64] Clement Levallois. Umigon: sentiment analysis for tweets based on lexicons and heuristics. *Proceedings of the International Workshop on Semantic Evaluation, SemEval*, 13, 2013.

[65] LikeAlyzer. Likealyzer: Analyze and monitor your facebook pages. http://likealyzer.com/, 2016. Acessed: 2016-09-21.

[66] Kuan-Cheng Lin, Shih-Hung Wu, Liang-Pu Chen, Tsun Ku, and Gwo-Dong Chen. Mining the user clusters on facebook fan pages based on topic and sentiment analysis. In *Proceedings of the 15th IEEE International Conference on Information Reuse and Integration, IRI 2014, Redwood City, CA, USA, August 13-15, 2014*, pages 627–632, 2014.

[67] Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. Emoticon smoothed language models for twitter sentiment analysis. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, pages 1678–1684. AAAI Press, 2012.

[68] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press.

[69] Guy McDowell. Is rss dead? a look at the numbers. http://www.makeuseof.com/tag/rss-dead-look-numbers/, 2015. Acessed: 2016-05-23.

[70] Arun Meena and T. V. Prabhakar. Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. In *Proceedings of the 29th European Conference on IR Research*, ECIR'07, pages 573–580, Berlin, Heidelberg, 2007. Springer-Verlag.

[71] Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. "blissfully happy" or "ready to fight": Varying interpretations of emojis. In *Proceedings of the The 10th International AAAI Conference on Web and Social Media (ICWSM-16)*. AAAI Press, 2016.

[72] Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. Open domain targeted sentiment. In *EMNLP*, pages 1643–1654. ACL, 2013.

[73] Saif Mohammad, Cody Dunne, and Bonnie Dorr. 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, 2009. Association for Computational Linguistics.

[74] Saif M. Mohammad. #emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 246–255, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[75] Saif M. Mohammad and Peter D. Turney. 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, 2010. Association for Computational Linguistics.

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

[77] Ahmed Morsy, Sara A.and Rafea. *Natural Language Processing and Information Systems: 17th International Conference on Applications of Natural Language to Information Systems, NLDB 2012, Groningen, The Netherlands, June 26-28, 2012. Proceedings*, chapter Improving Document-Level Sentiment Classification Using Contextual Valence Shifters, pages 253–258. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[78] Le T. Nguyen, Pang Wu, William Chan, Wei Peng, and Ying Zhang. 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, 2012. ACM.

[79] F. A. Nielsen. Afinn, March 2011.

[80] Finn Årup Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903, 2011.

[81] Kamal Nigam. Using maximum entropy for text classification. In *In IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, 1999.

[82] Petra Kralj Novak, Jasmina Smailovic, Borut Sluban, and Igor Mozetic. Sentiment of emojis. *CoRR*, abs/1509.07761, 2015.

[83] Oxford. Oxford Learner's Dictionaries topic dictionaries. http://www.oxfordlearnersdictionaries.com/topic/, 2016. Acessed: 2016-07-03.

[84] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, October 2010.

[85] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January 2008.

[86] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[87] Nikolaos Pappas and Andrei Popescu-Belis. Sentiment analysis of user comments for one-class collaborative filtering over ted talks. In *36th ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 2013.

[88] Livia Polanyi and Annie Zaenen. *Computing Attitude and Affect in Text: Theory and Applications*, chapter Contextual Valence Shifters, pages 1–10. Springer Netherlands, Dordrecht, 2006.

[89] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Comput. Linguist.*, 37(1):9–27, March 2011.

[90] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[91] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[92] Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, pages 43–48, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[93] REMINDS. Reminds: Relevance mining detection system. http://projectreminds.com/index.php/en/, 2016. Acessed: 2016-09-21.

[94] R. Remus, U. Quasthoff, and G. Heyer. Sentiws – a publicly available german-language resource for sentiment analysis. In *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*, pages 1168–1171, 2010.

[95] Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29, 2016.

[96] Tyler W. Rinker. *qdap: Quantitative Discourse Analysis Package*. University at Buffalo/SUNY, Buffalo, New York, 2013. 2.2.4.

[97] P. M. Roget. *Roget's Thesaurus of English words and phrases*. Available from Project Gutemberg, Illinois Benedectine College, Lisle IL (USA), 1852.

[98] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2 edition, 2003.

[99] Sartaj Sahni and George Vairaktarakis. The master-slave paradigm in parallel computer and industrial settings. *Journal of Global Optimization*, 9(3):357–377, 1996.

[100] Shankar Shetty, Rajendra Jadi, Sabya Shaikh, Chandan Mattikalli, and Uma Mudenagudi. Classification of facebook news feeds and sentiment analysis. In *2014 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2014, Delhi, India, September 24-27, 2014*, pages 18–23, 2014.

[101] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. How useful are your comments?: Analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 891–900, New York, NY, USA, 2010. ACM.

[102] Mário J. Silva, Paula Carvalho, Luís Sarmento, Eugénio Oliveira, and Pedro Magalhães. The design of OPTIMISM, an opinion mining system for portuguese politics. In *New Trends in Artificial Intelligence: Proceedings of EPIA 2009 - Fourteenth Portuguese Conference on Artificial Intelligence*. Universidade de Aveiro, October 2009.

[103] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank.

[104] Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA, 1966.

[105] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307, June 2011.

[106] Maite Taboada and Jack Grieve. Analyzing appraisal automatically. In *In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pages 158–161, 2004.

[107] Songbo Tan and Jin Zhang. An empirical study of sentiment analysis for chinese documents. *Expert Systems with Applications*, 34(4):2622 – 2629, 2008.

[108] Duyu Tang, Furu Wei, Bing Qin, Ming Zhou, and Ting Liu. 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, 2014.

[109] Justin Tenuto. Classification accuracy is not enough: More performance measures you can use. http://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/, 2014. Acessed: 2016-08-20.

[110] Justin Tenuto. #deflategate was just a chance for us to make some really bad jokes. https://www.crowdflower.com/deflategate-sentiment/, 2015. Acessed: 2016-08-12.

[111] Mike Thelwall. Heart and soul: Sentiment strength detection in the social web with sentistrength 1, 2013.

[112] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in twitter events. *J. Am. Soc. Inf. Sci. Technol.*, 62(2):406–418, February 2011.

[113] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol.*, 63(1):163–173, January 2012.

[114] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558, December 2010.

[115] Tun Thura Thet, Jin-Cheon Na, and Christopher S.G. Khoo. Aspect-based sentiment analysis of movie reviews on discussion boards. *J. Inf. Sci.*, 36(6):823–848, December 2010.

[116] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[117] Mikalai Tsytsarau and Themis Palpanas. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478–514, 2012.

[118] A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178–185, 2010.

[119] Peter D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual*

Meeting on Association for Computational Linguistics, ACL '02, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[120] Twitter. Stream api documentation. https://dev.twitter.com/rest/public. Acessed: 2015-10-19.

[121] Twitter. Twitter Company about. https://about.twitter.com/company, 2015. Acessed: 2015-10-19.

[122] Twitter. Twitter Company rest. https://dev.twitter.com/rest/public, 2015. Acessed: 2015-10-19.

[123] Ro Valitutti. Wordnet-affect: an affective extension of wordnet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086, 2004.

[124] Ulli Waltinger. Germanpolarityclues: A lexical resource for german sentiment analysis. In *LREC*, 2010.

[125] Y. Wan and Q. Gao. An ensemble sentiment classification system of twitter data for airline services analysis. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1318–1325, Nov 2015.

[126] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 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, 2012. Association for Computational Linguistics.

[127] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1031–1040, New York, NY, USA, 2011. ACM.

[128] Matthew Whitehead and Larry Yaeger. *Sentiment Mining Using Ensemble Classification Models*, pages 509–514. Springer Netherlands, Dordrecht, 2010.

[129] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, pages 34–35, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[130] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[131] Qiang Ye, Rob Law, and Bin Gu. The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28(1):180–182, 2009.

[132] Qiang Ye, Ziqiong Zhang, and Rob Law. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36(3, Part 2):6527 – 6535, 2009.

[133] Yasuhisa Yoshida, Tsutomu Hirao, Tomoharu Iwata, Masaaki Nagata, and Yuji Matsumoto. Transfer learning for multiple-domain sentiment analysis - identifying domain dependent/independent word polarity. In Wolfram Burgard and Dan Roth, editors, *AAAI*. AAAI Press, 2011.

[134] Lei Zhang and Bing Liu. 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, 2011. Association for Computational Linguistics.

Chapter 8

Appendix

8.1 Appendix I: Chapter 5 Extended Results

Table 8.1: Results in terms of precision (Prec.), recall (Rec.), F1-score (F1), and accuracy (Acc.) of score with 100% text evaluation and 0% emoticon/emoji evaluation

Dataset	Prec. (%)		Rec.(%)		F1(%)		Acc.(%)
	{Pos.}	{Neg.}	{Pos.}	{Neg.}	{Pos.}	{Neg.}	{Pos.+Neg.}
GOP	32.82	90.87	45.22	32.93	38.04	48.35	61.84
SD-Cars	82.53	53.82	76.24	63.19	79.27	58.13	72.65
US Airlines	39.43	97.44	94.93	56.97	55.71	71.90	65.62
Coachella	85.36	42.27	74.04	59.93	79.30	49.58	70.65
Apple	55.44	93.85	86.75	74.36	67.65	82.98	77.69

Table 8.2: Results in terms of precision (Prec.), recall (Rec.), F1-score (F1), and accuracy (Acc.) of score with 30% text evaluation and 70% emoticon/emoji evaluation

Dataset	Prec. (%)		Rec.(%)		F1(%)		Acc.(%)
	{Pos.}	{Neg.}	{Pos.}	{Neg.}	{Pos.}	{Neg.}	{Pos.+Neg.}
GOP	32.85	90.85	78.21	57.50	46.26	70.42	61.84
SD-Cars	82.59	53.99	76.70	62.84	79.54	58.08	72.50
US Airlines	39.85	97.67	95.36	57.42	56.21	72.31	66.08
Coachella	85.32	41.75	73.35	60.20	78.89	49.30	70.19
Apple	44.0	89.53	49.47	74.70	57.88	81.44	73.54

8.2 Appendix II: Trend x Sentiment Plots

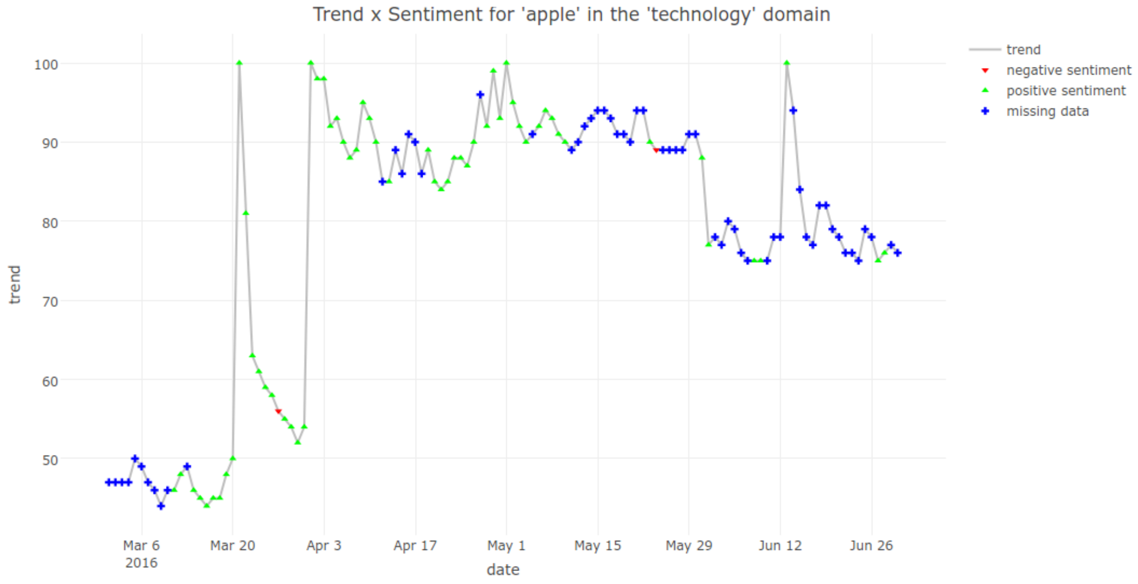


Figure 8.1: "apple" in "technology" domain

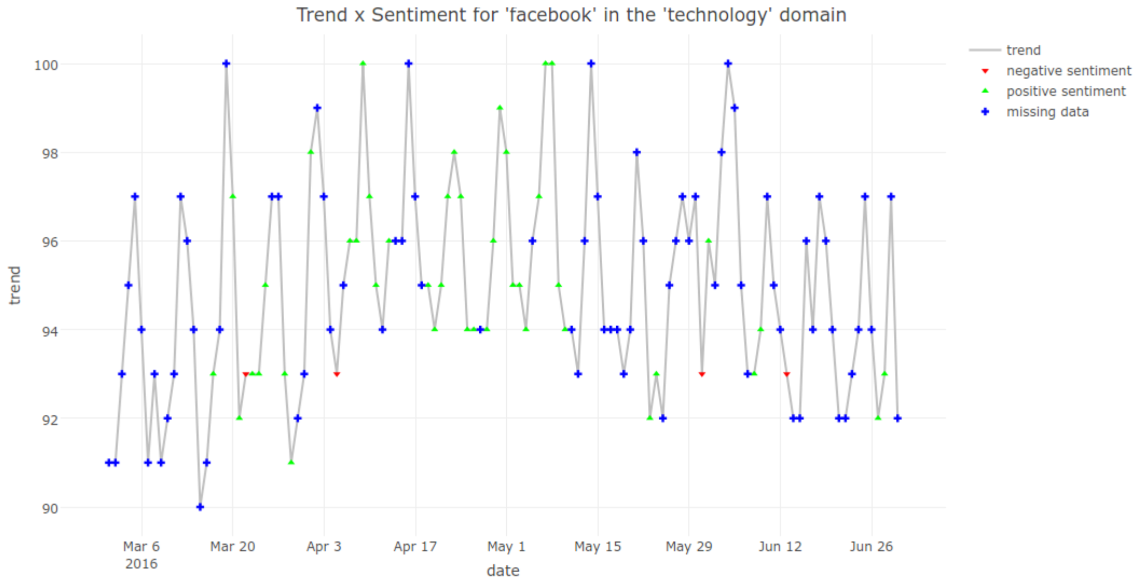


Figure 8.2: "facebook" in "technology" domain

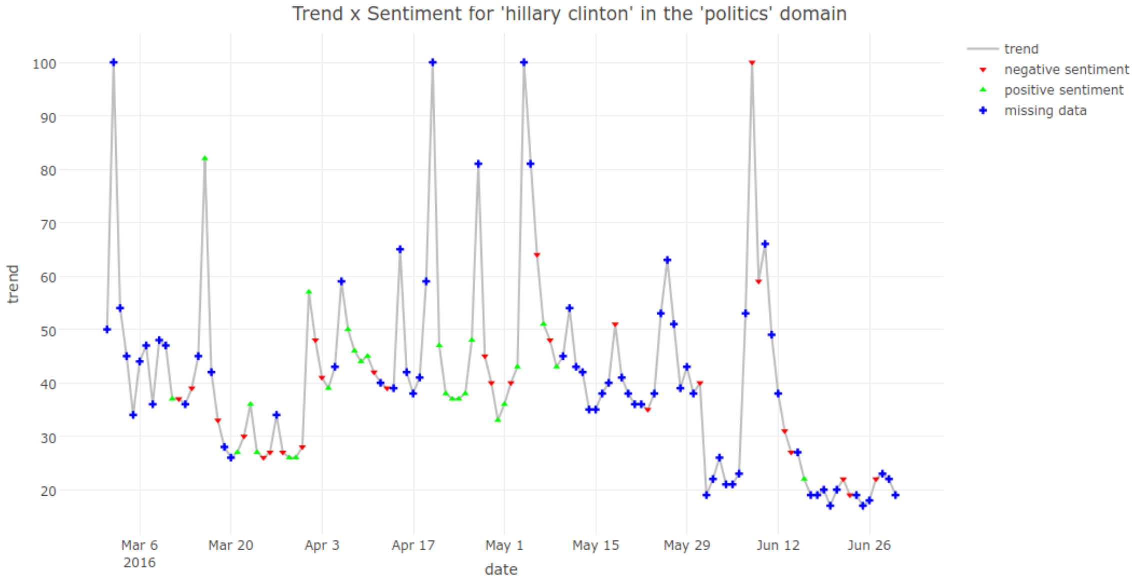


Figure 8.3: "hillary clinton" in "politics" domain

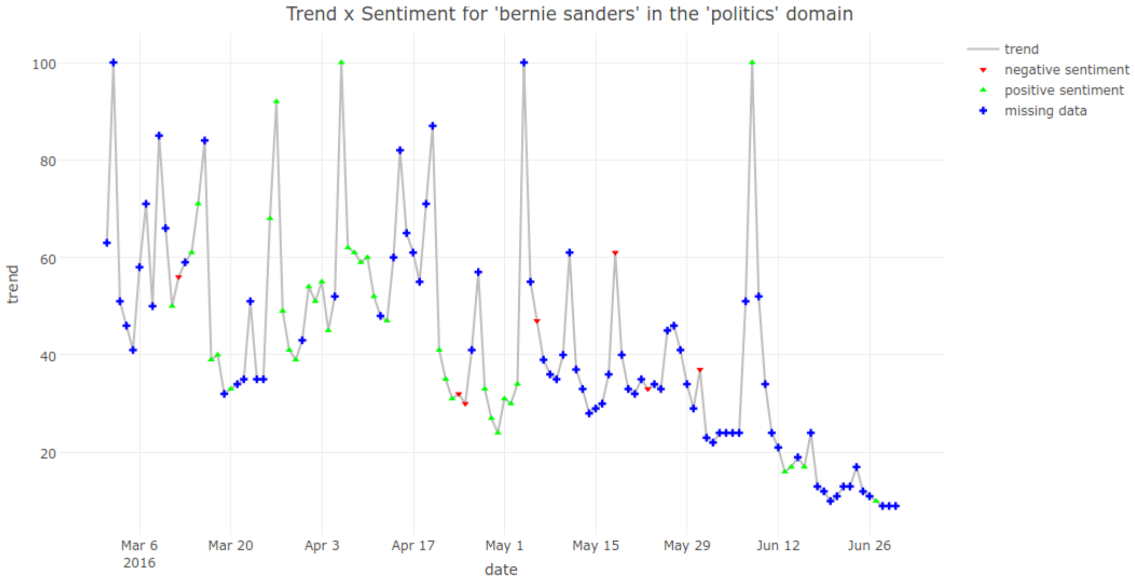


Figure 8.4: "bernie sanders" in "politics" domain

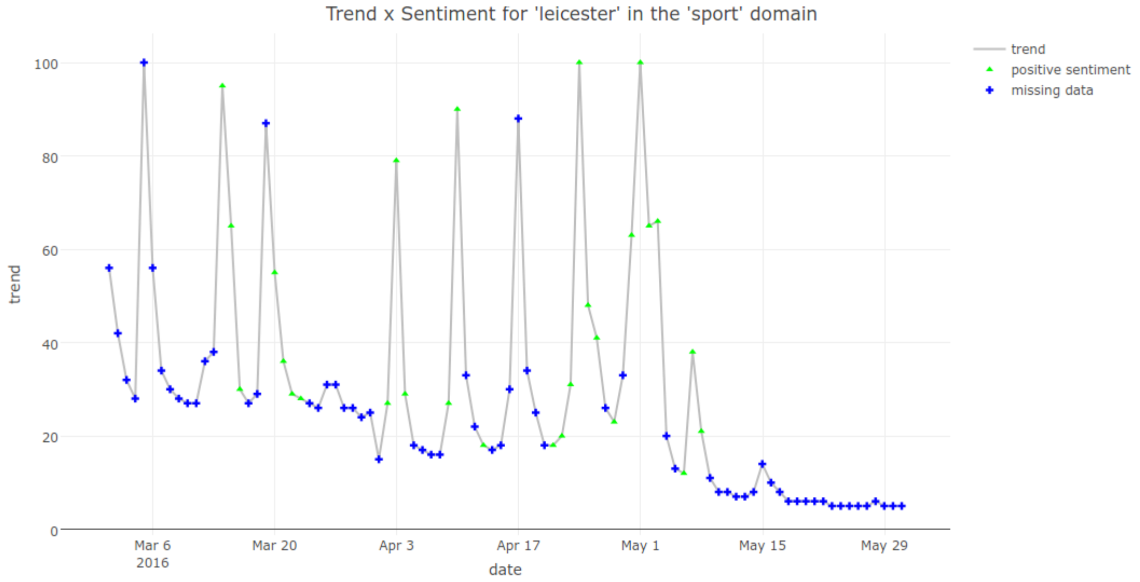


Figure 8.5: "leicester" in "sports" domain

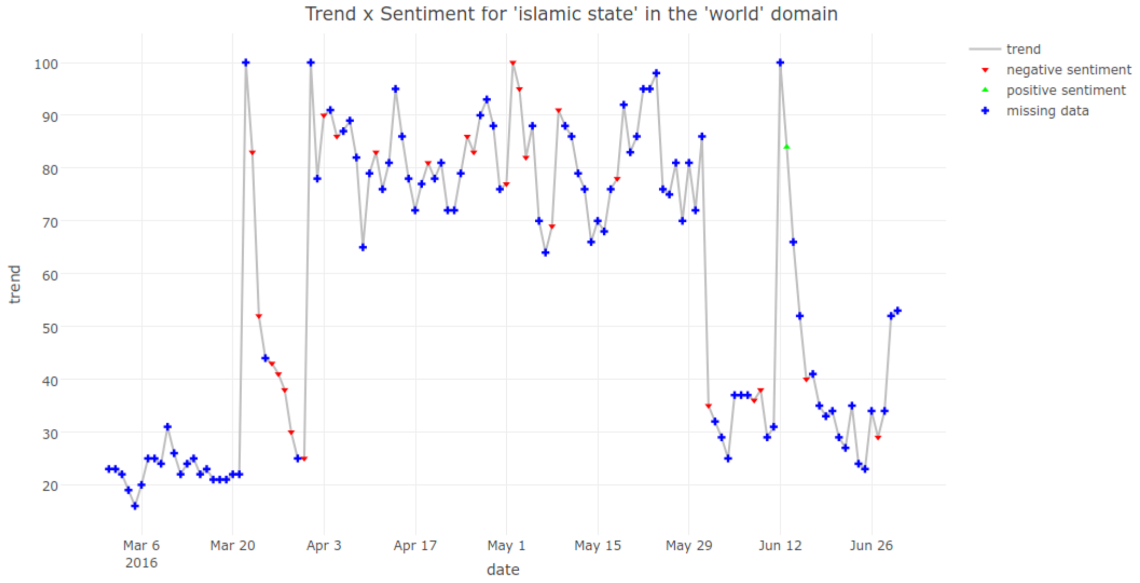


Figure 8.6: "islamic state" in "world" domain

8.3 Appendix III: Documentation IFeelR Package

Package 'ifeelR'

August 8, 2016

Type Package

Version 0.0.3

Date 2016-06-15

Title iFeel R Package

Author Nuno Guimaraes [aut, cre]

Maintainer Nuno Guimaraes <nunorguimaraes@inesc.pt>

Depends R (>= 3.1.0), rJava

Suggests qdap

SystemRequirements Java JDK 1.2 or higher (for JRI/REngine JDK 1.4 or
higher), GNU make

Description Package with several state of the art methods for performing sentiment analysis in text.

License GPL (>= 2)

URL https://www.r-project.org, http://www.another.url

BugReports https://pkgname.bugtracker.url

R topics documented:

`getConfidenceSentiment`

Get Confidence Sentiment Function

Description

The `getConfidenceSentiment` performs a sentiment analysis on the text provided using all methods or the selected methods pass by the "methods" argument. The result will be a length 3 vector with the sentiment confidence for each one of the classes (Positive/Neutral/Negative) based on the sentiment provided by each one of the methods.

Usage

```
getConfidenceSentiment(text,verbose=F)
```

Arguments

`text`	The text which should be analyze.
`verbose`	Whether or not to print information about the sentiment voting process.
`methods`	The list of methods to be used. It can be a vector with the id of the methods or 'A' for all

Examples

```
## Not run:
getConfidenceSentiment("I love the new Star Wars movie", methods=c(1,2,3,4,17))
getConfidenceSentiment("I hate you so much!!!!", verbose=T, methods="A")

## End(Not run)
```

`getEnsembleCastSentiment`

Get Ensemble Cast Sentiment Function

Description

The `getEnsembleCastSentiment` performs a sentiment analysis on the text provided using all methods or the list of methods provided in the "methods" argument. The final sentiment value is reached by the majority of votes where each vote is the sentiment returned by each method. If there is a tie between positive and negative, the neutral class is returned. If there is a tie between the neutral class and positive/negative class, the second one is returned. In case there is a 3 class tie, the neutral class is returned. The possible result values are 1 (Positive Sentiment), 0 (Neutral Sentiment) or -1 (Negative Sentiment).

Usage

```
getEnsembleCastSentiment(text,verbose=F)
```

Arguments

`text` The text which should be analyze.

`verbose` Whether or not to print information about the sentiment voting process.

`methods` The list of methods to be used. It can be a vector with the id of the methods or 'A' for all

Examples

```
## Not run:
getEnsembleCastSentiment("I like you so much!!!", methods="A")
getEnsembleCastSentiment("That was a bad investment.", verbose=T, methods=c(1,4,5,10,19,17))


## End(Not run)
```

`getSentiment` *Get Sentiment Function*

Description

The `getSentiment` performs a sentiment analysis on the text provided using the method indicated by the `methodID`. The possible result values are 1 (Positive Sentiment), 0 (Neutral Sentiment) or -1 (Negative Sentiment).

Usage

```
getSentiment(text,methodID)
```

Arguments

`text` The text which should be analyze.

`methodID` The id of the method that should be used. Please see `printMethods to see the available methods.`

Examples

```
## Not run:
getSentiment("I love the new Star Wars movie", 13)
getSentiment("I hate you so much!!!!", 7)

## End(Not run)
```

`printMethods` *Print methods*

Description

Print the available methods for sentiment analysis.

Usage

```
printMethods()
```

Examples

```
## Not run:
printMethods()

## End(Not run)
```

Index

8.4 Appendix IV: Submitted Papers

8.4.1 Lexicon Expansion System for Domain and Time Oriented Sentiment Analysis

This work was submitted and accepted in the 8th International Conference on Knowledge Discovery and Information Retrieval.

Lexicon Expansion System for Domain and Time Oriented Sentiment Analysis

Nuno Ricardo Guimaraes[1], Luis Torgo[2] and Alvaro Figueira[1]
[1]*CRACS & INESC TEC, University of Porto, Rua do Campo Alegre 1021/1055, Porto, Portugal*
[2]*LIADD & INESC TEC, University of Porto, Rua do Campo Alegre, 1021/1055, Porto, Portugal*
nuno.r.guimaraes@inesctec.pt, {*ltorgo, arf*}*@dcc.fc.up.pt*

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 major 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 as Table 1 shows. On the other hand, analysing directly headlines (or the full news article) may not provide a accurate sentiment on the terms it mentions.

Table 1: Top Trends on Twitter (2016-03-08)

#2DaysTilDangerousWoman
#73YMasPorTuFelicidadViciconte
#AKPninKadnaBak
Aldo Ferrer
#ALDUBTheCompromise
ALYCIA WANTED TO COME BACK
#AnittaRepresentaAMulherBR
#BenDeme
Bertrand
#Bibi

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 using 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).

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

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).

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

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.

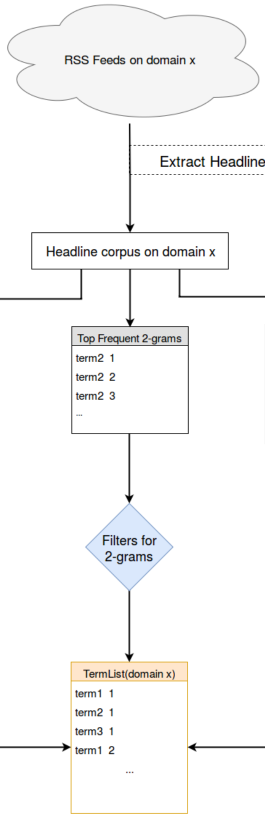


Figure 1: Terms Extraction Workflow

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 num-

```
$sentiment \leftarrow 0$;
$c \leftarrow 0.8$;
for $w \in tweet$ do
    if $w \in SentimentDictionary$ then
        $s_{val} \leftarrow SentimentDictionary[w]$;
        // Get the two previous words
           and the four words after
        $cluster \leftarrow getWordsBefore(2) + w +$
         $getWordsAfter(4)$; $neg_{count} \leftarrow 0$;
         $amp_{count} \leftarrow 0$; $deamp_{count} \leftarrow 0$;
        for $c \in cluster$ do
            if $c \in NegatorList$ then
                $neg_{count} \leftarrow negcount + 1$
            end
            if $c \in AmplifierList$ then
                $amp_{count} \leftarrow amp_{count} + 1$;
            end
            if $c \in Deamplifier$ then
                $deamp_{count} \leftarrow deamp_{count} + 1$;
            end
        end
        $neg_{val} \leftarrow neg_{count}$ mod 2;
        $amp_{val} \leftarrow neg_{val} * amp_{count}$;
        $deamp_{val} \leftarrow$
          $(-neg_{val}) * amp_{count} + deamp_{count}$;
        $D \leftarrow max(deamp_{val}, -1)$;
        $c_{sent} \leftarrow$
          $(1 + c * (amp_{val} - D)) * s_{val} * neg_{count}$;
        $sentiment \leftarrow sentiment + c_{sent}$;
    end
end
$N \leftarrow length(tweet)$;
$sentiment \leftarrow sentiment/sqrt(N)$;
// Constrain sentiment between [-1,1]
$sentiment \leftarrow$
  $((1 - (1/(1 + exp(sentiment)))) * 2) - 1$
```

Algorithm 1: Sentiment Analysis procedure on each tweet

ber 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).

An overview of our term extraction workflow is represented in Fig. 1. Table 2 shows an example of the resulting terms for each domain.

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

Table 2: Sample of terms for each domain retrieved in 03-04-2016

Domains						
World	**Entertainment**	**Technology**	**Sports**	**Business**	**Health**	**Politics**
azerbaijan	tonight	microsoft	villanova	sales	valeant	sanders
migrant	ronnie corbett	google	thompson	money	cdc	cruz
syrian	zaha hadid	ipad pro	west indies	steel crisis	abortion pill	massive recession
tsunami	batman v superman	april fools day	bahrain grand prix	virgin america	nuclear waste	donald trump
islamic state	guns n roses	tesla model 3	ncaa title game	minimum wage	zika virus	state department

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 previously mentioned and the algorithm 1 (based on the work in (Rinker, 2013)). 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 average by summing up the sentiment score of each tweet from the term *corpus* and dividing it by the number of tweets in that *corpus*. An overview of our sentiment analysis component can be seen in Figure 2.

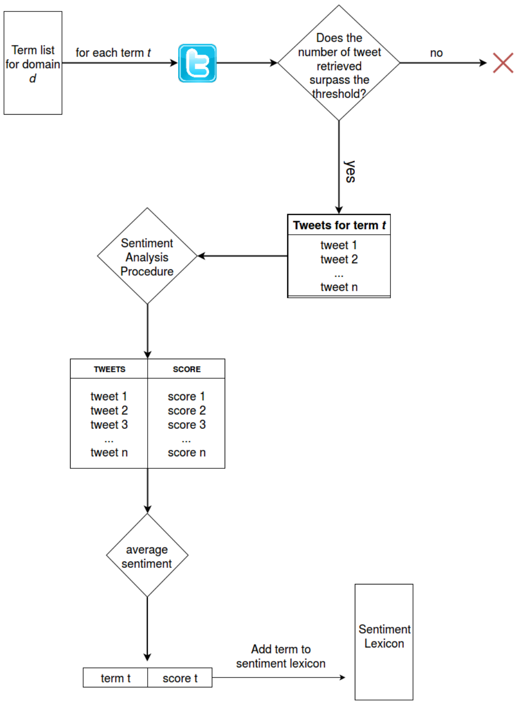


Figure 2: Term Sentiment Analysis Workflow

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 3.

Table 3: 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, 2013) 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., 2010; Thelwall et al., 2012; Thelwall, 2013).

The results can be seen in Table 4.

Table 4: 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**

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. 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 5.

Table 5: 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 Annual 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.

Rinker, T. W. (2013). *qdap: Quantitative Discourse Analysis Package*. University at Buffalo/SUNY, Buffalo, New York. 2.2.4.

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. (2013). Heart and soul: Sentiment strength detection in the social web with sentistrength 1.

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.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558.

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.

8.4.2 The Importance of Public Opinion on Sentiment Analysis in Subjective Texts

This work was submitted to the Social Network and Media Analysis (SONOMA) track on the 32nd ACM Symposium on Applied Computing.

The Importance of Public Opinion on Sentiment Analysis in Subjective Texts

ABSTRACT

Sentiment lexicons are an essential component on most state of the art sentiment analysis methods. However, the terms included are usually restricted to verbs and adjectives since they are normally used abroad different domains, with similar meanings. This can lead to a problem on the classification of short informal texts since sometimes public opinion on these terms is crucial to determine the correct polarity. Therefore, to complement the traditional sentiment dictionaries we present a system for lexicon expansion who extracts the most relevant terms and assesses their positive or negative score through Twitter. Preliminary results on a labelled dataset shows that our complementary lexicons increase the performance of three state of the art sentiment dictionaries, therefore proving the effectiveness of our approach.

CCS Concepts

•**Information systems** → **Sentiment analysis;** *Information extraction;*

Keywords

lexicon expansion, sentiment analysis, social network applications

1. INTRODUCTION

A sentiment analysis task aims to, given a fragment of text, classify it with a score associated with a positive, neutral or negative value. Early research has focussed on user reviews on online sites. However, the massive growth of social networks has provided a different source for sentiment analysis. This boom on the area was caused mostly because the way users share their opinion through short comments or text, in several different domains. In addition, the large quantity of data available and the quickness of its extraction has recently promoted the area to a ”hot topic” research subject.

One of the key factors for a precise and correct sentiment analysis classification are the sentiment lexicons or sentiment dictionaries. This consist in list of words, mainly adjectives and verbs, with an associated sentiment value (p.e. ”beautiful: +2” and ”bad: -1”). A basic example of a sentiment analysis procedure looks for all the words in the text that are within the dictionary. The sum of the values of those entries corresponds to the final sentiment score of the analysed text.

However, due to the way that social network users express their opinion's and to the reduced length of their texts, lexicons composed only of verbs and adjectives are generally not enough. For example ”I don't know what to think about Brussels...” at the specific time of terrorist attacks to the city [22] expresses a negative sentiment. In the same way, the tweet ”I KNEW IT! LEICESTER!!” refers to a positive sentiment at the time this club won in the English Premier League [14]. In both cases, state of the art sentiment lexicons would not correct classify them, since there are no adjectives or verbs included. However, human knowledge on the public opinion of some terms can lead to easier assessment on the sentiment of these texts.

Therefore, our research hypothesis states: ”Can public opinion generated lexicons improve the sentiment classification of sentiment analysis methods on short informal texts?”. With that goal in mind, we propose a system that automatically extracts and classifies domain and time specific terms with sentiment based on public opinion. This system can 'return' dictionaries, on a daily basis, to complement the more traditional sentiment lexicons.

2. RELATED WORK

There have been several approaches to the creation and/or expansion of sentiment dictionaries. We can divide the subject in three main categories: manual labelled, thesaurus based, and corpus based approaches.

Manual labelled sentiment dictionaries rely on human annotators to assess the score on each entry. The author in [19] selected a set of words from several affective word lists (like ANEW [5]), added slang and obscene terms and manually labeled with a sentiment score ranging from -5 to 5. Another work [13] takes a similar approach. The authors

SAC'17, April 3-7, 2017, Marrakesh, Morocco

Copyright 2017 ACM 978-1-4503-4486-9/17/04...$15.00

http://dx.doi.org/xx.xxxx/xxxxxxx.xxxxxxx

create a manual label sentiment dictionary by inspecting already well established lexicons and adding acronyms and slang words. Then, recurring to Amazon Mechanical Turk, they assess each word sentiment using 10 independent workers and a careful quality control on the data extracted. Finally they combine this lexicon with a rule based system that takes into consideration negations (”not good”), degree modifiers (”very good”), punctuation and capitalization to outperform seven state of the art sentiment lexicons. Other approach [18] classifies the words with emotion and polarity. The terms were extracted from a combination of Macquarie Thesaurus [1], General Inquirer [25] and WordNet Affect Lexicon [31].

Corpus based approaches relies in a small set of previously defined sentiment words (titled ”seed lexicon”) to create or extend sentiment lexicons. For example, the work in [10] is a Twitter corpus based lexicon. The creation process consists in the extraction of tweets with only a happy (:) or :-)) or sad (:(, :-( ) emoticon. Then, the assumption is that tweets with a smiling emoticon correspond to positive tweets and with a sad emoticon to negative ones. Finally, the corpus is divided considering the emoticons and the most frequent words in each one are included in the lexicon with a positive or negative value. A similar approach is presented in [17]. However, instead of emoticons, the tweet retrieval process is done with emotion hashtags such as ”#angry” and ”#happy”. The lexicon evaluates each word with six different emotions and positive and negative sentiment.

Finally, thesaurus based approaches use word resources like WordNet [8] for expanding an initial defined seed lexicon. WordNet is a lexical database that includes nouns, verbs and adjectives grouped by synonyms sets. In adjectives there is also a connection between antonyms. This is particularly useful for expanding sentiment lexicons. As an example, SentiWordNet uses this feature to expand a seed lexicon by assigning the same polarity to synonyms the opposite to the antonyms. Several other studies [15, 12] use WordNet to expand sentiment or create sentiment lexicons, making it one of the most used resources for the creation or expansion of dictionaries.

Nevertheless, none of this approaches considers the use of relevant (domain and time dependent) terms, which may be crucial for the correct polarity assessment on short informal texts – ultimately, constitutes the motivation for this work.

3. DATA EXTRACTION

In this section, we describe how the lexicons were extracted. First we select 6 of the most common news domains: ”world”, ”entertainment”, ”politics”, ”sports”, ”health”, ”technology” and ”business”. Next, using different news sources, we crawl the headlines on a daily basis to retrieve the more relevant terms on each domain. Then, we use those terms as queries to extract a corpus of tweets for each one of them. Finally, using sentiment analysis procedures on tweets, we assess the public opinion of each term. Therefore we have a daily construction of dictionaries for domain and time specific entries.

3.1 Terms Extraction

To extract the more relevant terms in each domain we create a corpus of headlines from several different news sources. The number of sources in each domain ranges between 9 and 14. We limited our news sources research to the English language and whose origin countries are the United States or included in the United Kingdom. In fact, a survey puts the US and UK as the two most influential countries in the world according to several different factors [11]. Therefore, we argue that international media coverage is bigger in this countries and consequently, public opinion data should also be vast and easier to acquire using terms from this geographical sources. The sources used were: CNN, BBC, The Economist, The Wall Street Journal, ABC News, CBS News, The Washington Post, NBC, The Guardian, Reuters, Yahoo News, Sky News, Daily Mail, The New York Times, Financial Times, Forbes and MedicineNet.

For each domain corpus, we remove punctuation and impose lower case. Then we build three lists by extracting, all unigrams, bigrams and trigrams in order of frequency. Through experimentation we realise that, most of the times, terms above trigrams were unique (in other words, they only occur in one headline) so we discard them .

Next, we perform a series of text filtering. We exclude both verbs and adjectives from the lists using OpenNLP Part of Speech Tagger [3]. In addition, we also exclude terms that are within the domain and are not subject to public opinion (e.g. ”soccer” in sports or ”film” in entertainment) recurring to the word lists provided in [20]. Furthermore, we also exclude possible sentiment words using AFINN lexicon [19]. This way, some subjective adjectives or verbs that could pass the OpenNLP classifier are left out. Finally we removed words that were duplicated in plural form (”syrian”/”syrians”), and lemmatized when in a presence of an apostrophe (”Clinton”/”Clinton’s”).

The POS-Tagger filter is only applied to unigrams whereas the other filters are used in all term lists, since terms with two or more words already imply a certain context. The last filter is applied to all lists at the time when the sample of tweets for each term is extracted. Through experimentation, we defined a threshold of 33% on the minimum sample of tweets to be retrieved. Consequently, terms below that minimum are excluded. Since we are searching for a exact match on the queries, incomplete or irrelevant terms are unlikely to reach the minimum number of tweets.

3.2 Public Opinion Assessment

The second component of our system is to determine the public opinion of the extracted terms using Twitter. For that purpose, we extract 100 tweets regarding each term. This number was achieved by experimental procedures which took into account the restrictions imposed by the Twitter API [30] as well as the time to classify each tweet with sentiment.

We also impose some restrictions on the tweets extracted for each term. Since we want to keep the sentiment updated, we only retrieve tweets posted in the same day as the term extraction procedure. In addition, we use the parameters provided by the Twitter REST API [29] to retrieve the most recent tweets. Furthermore, in order to avoid extracting posts by news sources (since we want to analyse the sentiment exposed by common users and not by news media) we do not extract tweets that contain an external link. This is due to the fact that the majority of Twitter accounts that belong to the news industry refer to their web page in each news post (so the user can read the full article).

As soon as the tweet term corpus is built, we applied some cleaning procedures to it and begin the sentiment analysis in each tweet. For that purpose, we built an ensemble system (ENS17) which takes into account a selection of sentiment analysis methods to improve the inter-domain performance. The methods/sentiment dictionaries used were the following: AFINN [19], Emolex [18], EmoticonDS [10], HappinessIndex [5], MPQA [33], NRC Hashtag [17], Opinion Lexicon [12], SANN [21], Sasa [32], SenticNet [6], Sentiment140 [9], SentiStrength [28], SentiWordNet [4], SoCal [26], Standford Adapter [24], Umigon Adpater [2] and Vader [13].

To facilitate the building of the ensemble we used the iFeel framework, which allows the selection of specific sentiment analysis methods [23]. The decision making procedure on the score returned is done by majority voting. When a tie occurs, the rules are the following:

- When there is a tie between positive and negative classes, the neutral sentiment is returned
- When there is a tie between the neutral class and other class, the other class is returned
- Since there are 17 sentiment systems, a 3-way tie is not possible. However, in the events of future changes, we defined that in this case, the neutral value is returned

An example of the behaviour of the ensemble system can be seen in Fig.1. It is important to point out that some of the methods used are only dictionaries and thus it is necessary to specify how the classification of the text will be done. Taking this into account, for the lexicon only approaches (AFINN, Emolex, EmoticonDS, NRC Hashtag, Opinion Lexicon, Sentiment 140, and SentiWordNet), iFeel uses Vader rule based system to leverage the performance of this lexicons [23].

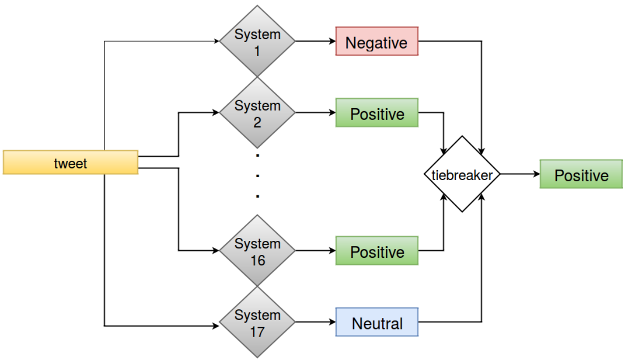


Figure 1: Ensemble system example

3.3 Ensemble System Evaluation

Since an accurate tweet sentiment analysis is essential for the results of our system, in this section we compare our approach against the individual state of the art methods which comprise our ensemble system, on a sample of different domain tweet datasets. The datasets used were extracted from Crowdflower Data for Everyone Library [7] and included set of tweets referring to: the 2016 GOP debate (GOP), Google self-driving cars (SDC), coachella line-up announcment (COACH), united states airlines (USAIR) and the Deflategate scandal (NFL). We assess our ensemble results in terms of accuracy and average F1-score in different classes and different domains. We select an equal number of entries in each dataset. In addition we also balanced classes. Therefore, each dataset has 1200 tweets (400 positive, 400 neutral and 400 negative). First, we compare our ensemble system with the top 3 more accurate systems in each domain and using the aggregation of all datasets. The results are presented in Table 1. When compared with each stand alone system, the ENS17 ensemble is in the top three most accurate systems in almost all datasets (it fails in the COACH dataset, although the difference is 0.1%).

This ensemble does not achieve the best score in any of the datasets. However, if we look at the accuracy across all domains (in other words the average accuracy in all datasets) the ENS17 outperform the best individual systems.

Table 1: Comparison of ensemble method against the more accurate individual systems in each domain

Top Individual Systems (on each dataset)	Dataset Accuracy						
	GOP	SDC	APPLE	USAIR	COACH	NFL	Total
1st System	49.0	**53.0**	**69.5**	**62.0**	**46.5**	35.2	48.8
2nd System	48.1	51.3	61.3	59.3	46.1	34.1	48.3
3rd System	47.5	47.5	54.6	58.4	45.3	33.4	48.0
Ensemble Systems							
ENS17	51.0	51.6	60.4	60.0	45.2	**45.2**	**52.2**

A similar analysis can be done separating the accuracy in Negative, Neutral and Positive classification. With this purpose we will use the top 3 systems that are the more accurate in all domains (i.e. the individual systems that were selected for the "Total" column in 1). These systems are AFINN, SentiStrength and Umigon. The table regarding class accuracy can be examine in Table 2.

Table 2: Comparison of ensemble method against the most accurate individual systems.

Best Overall Individual Systems (using Accuracy as metric)	Class Accuracy (%)			
	Negative	Neutral	Positive	Total
Umigon	33.8	**77.3**	35.1	48.8
SentiStrength	36.8	63.2	44.7	48.3
AFINN	36.3	59.3	48.3	48.0
Ensemble Systems				
ENS17	35.8	72.3	**48.6**	**52.2**

As we can observe, regarding classes, ENS17 is always in the top 3 systems when comparing with the most accurate individual systems. Furthermore, it achieves the highest accuracy value on the negative and positive class, respectively. Since the dataset are balanced in number of entries and number of elements in each class, it's no wonder that the all classes accuracy value are the same as the total accuracy values in all datasets. Since accuracy values can sometimes be misleading [27], we perform the same analysis using the average F1-score in each domain. Therefore, selecting the top 3 systems according to the average F1-score and assessing the same metric with our method results in the values presented in table 3

Table 3: Comparison of ensemble method against the top individual systems (according to F1-metric) in each domain

Top Individual Systems	**Dataset F1-score (%)**						
(on each dataset)	**GOP**	**SDC**	**APPLE**	**USAIR**	**COACH**	**NFL**	**Total**
1st System	48.6	**52.3**	**69.1**	61.9	44.6	32.7	47.8
2nd System	47.5	50.6	60.9	58.6	43.5	31.8	47.2
3rd System	45.8	46.6	55.0	57.6	42.3	30.1	47.6
Ensemble Systems							
ENS17	50.3	50.8	60.4	59.4	42.2	**42.2**	**51.5**

Once again it is clear that each of the ensemble systems perform well enough to be in the top 3 systems using F1-score in each dataset. Therefore, it is no surprise that, when considering all datasets, ENS17 achieves an average F1-score superior than each of the individual systems.

Table 4: Comparison of ensemble method against the top individual systems (according to F1-metric) in each class

Best Overall Individual Systems	**Class F1-score (%)**			
(using F1 measure as metric)	**Negative**	**Neutral**	**Positive**	**Total**
SentiStrength	44.0	51.2	48.2	47.8
AFINN	44.3	50.2	48.3	47.6
Umigon	41.8	54.7	45.2	47.2
Ensemble System				
ENS17	**46.6**	**55.6**	**52.2**	**51.5**

Finally we take a closer look on the performance of the class classification using the concatenation of all datasets and the F1-score metric. Results are provided in table 4. It is easily noticeable that the ensemble system outperform the individual systems, therefore proving the validity of our approach.

4. EVALUATION

The final stage of this work is to determine if the dictionaries built with the described system can improve the sentiment analysis task for short 'informal' texts. To answer our research question, we used a dataset that contains posts and comments from Facebook and tweets from September 7th to September 14th, evaluated with sentiment on Crowdflower. The dictionaries from our system were retrieved in September 5th. This way, we guarantee that the tweets and Facebook posts used to create the dictionaries were not included in the dataset where we performed the evaluation but, are close enough so the public opinion on the terms extracted does not fade or changes substantially.

4.1 Dataset Description

As it was already mentioned, the dataset combines 3 types of short informal texts: Facebook posts, Facebook comments and tweets. The Facebook posts and comments were retrieved from the top most popular pages in different categories from the United States according to the LikeAlyzer tool [16]. For each post on the defined time interval, we extracted up to a maximum of 20 comments (order by the Facebook relevance metric). From that extraction, a sample of 1000 comments and 3995 posts were sent to Crowdflower for evaluation.

Regarding the tweets extraction, relevant topics (which appeared on recent news) were used on the Search API. To retrieve a large number of tweets in different domains, we used the terms we knew it would generate opinion tweets. Therefore, some keywords used as query were evaluated with sentiment in our lexicons. Consequently to avoid biases results, we excluded those terms from the dictionaries. The keywords used as queries were the following: terrorism, refugees, elections, paralympic, champions league, emmys and wall street.

For each keyword 714 tweets were extracted forming a total of 4998 tweets. Concatenating this data with the one extracted from Facebook, we have a final dataset of 9993 entries of short informal texts for evaluation.

This experiment on Crowdflower had one evaluation per entry. This can lead to a weaker ”ground truth” but, also to an higher number of texts evaluated. The decisions on taking this evaluation approach were due to the importance on having consistency on the short informal texts classified in contempt of a strong evaluation methodology or agreement that could easily be manipulated to fit our data. The sentiment question asked to the workers was *”The sentiment expressed in this text is:”* To answer, the worker had a likert scale ranging from 1 to 5 and labelled from ”very negative” to ”very positive”. In addition, we also included a follow up question: *”Choose (from the provided text) the word that best supports your previous answer”*. Our goal was to lead the worker to take a more careful decision and to justify it. Finally, since we want to assess the impact of our complementary lexicon on improving the accuracy on subjectivity texts, we exclude the entries classified as neutral (since they are very likely to be factual) of our dataset. This left us with a dataset containing 5090 entries.

4.2 Evaluation on Non-Factual texts

When comparing with our ensemble sentiment system, we determine the best overall individual methods on the tweets datasets tested (as shown in Table 2 and 4). Therefore, we select AFINN [19], UMIGON [2] and SentiStrength [28] as the sentiment methods to complement with our lexicons.

To decide on which lexicon to use in each entry of the dataset, we need to fit each text in one of the domains previously defined (world, sports, entertainment, politics, business, technology and health). In this experiment, we used the frequency of words on the text that appear on the different dictionaries generated by our system to assess its domain. For the entries where no domain was found, we assigned the ”world” value.

Finally, we scale the sentiment classification on Crowdflower to Negative or Positive values to match our methods scales. The results of our experiment are presented in Table 5.

Table 5: Variation between the Sentiment Systems with and without the Expanded Lexicons in non factual text

Sentiment System	**Accuracy %**	**Average F1%**
AFINN	+1.12	+0.48
SentiStrength	+1.36	+1.43
Umigon	+3.14	+2.63

The addition of the lexicons outputted by our system improved in both accuracy and average F1-measure the tested methods. Umigon is the system that benefits the most on the addition of this lexicons and AFINN the less. The average accuracy improvement is around 1.87% whereas the F-measure is 1.51%.

Although it is not a major difference between both sentiment dictionary approaches (traditional and traditional + expanded) it is a steady improvement since it is consistent across all 3 analysed systems.

We can go further in our analysis of non factual texts and restrict our dataset to the entries whose response to the question *"Choose (from the provided text) the word that best supports your previous answer"* " was included in our expanded sentiment lexicon. The filtered dataset contains 215 entries and results of these specific cases can be consulted in Table 6.

Table 6: Variation between the Sentiment Systems with and without the Expanded Lexicons in non factual text with sentiment justification word in the Expanded Lexicons

Sentiment System	Accuracy %	Average F1%
AFINN	+2.23	+2.31
SentiStrength	+23.13	+9.81
Umigon	+24.11	+12.55

Although we are forcing that the word for the sentiment justification is present in our dictionary (and therefore imposing the condition that it will be used for the text sentiment evaluation), this analysis intends to show that, in specific cases of subjective short informal texts where the argument to assess the sentiment is not on traditional lexicons, using our system can result in a reasonable improvement. In fact, SentiStrength and Umigon have an accuracy boost superior to 20% whereas the F1-measure increases 9.81% and 12.55% respectively. This demonstrates that not only it is important to consider our system sentiment dictionaries but also that our term sentiment analysis is capable of accurately classify the terms.

5. CONCLUSION AND FUTURE WORK

In this work we studied the influence of public opinion for the task of assessing a positive/negative sentiment in subjective short informal texts (like tweets, posts or comments).

We built a framework capable of extracting and assessing the polarity score of the most relevant domain and time dependent terms. This system consisted in an extraction procedure (that relies on news headlines to retrieve relevant terms) and on an ensemble sentiment classifier (combining 17 state of the art sentiment analysis methods) to output 7 different sentiment dictionaries on a daily basis.

Next, we complement three state of the art sentiment lexicons (AFINN, UMIGON, and SentiStrength) with the dictionaries outputted from our system. We selected these three based on their previous performance on tweet datasets. We tested our approach on a sample of tweets, Facebook posts and comments with positive or negative polarity and whose value was manually assigned recurring to Crowdflower's platform.

The results achieved indicate a slight but coherent improvement in all methods. However, when the term for assessing the sentiment is not included in sentiment lexicons, their importance has increased significantly, proving that our approach can increment the performance of sentiment methods in these specific cases.

In future work we plan to further extend our system by adding a geographical component to the dictionaries generated. We intend to use news sources as well as tweet from specific countries for determining the effectiveness of our method in a more narrow scope evaluation. We also aim to extend our lexicons in more domains. This way, we can increase the number of terms and cover a broader area of short informal texts to be analysed.

6. REFERENCES

[1] *The Macquarie thesaurus / [general editor] J.R.L. Bernard.* Herron Publications West End, Qld, new budget ed. edition, 1987.

[2] *Umigon: sentiment analysis for tweets based on lexicons and heuristics.* Zenodo, June 2013.

[3] Apache. Opennlp, 2010.

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

[5] M. M. Bradley and P. J. Lang. Affective norms for English words (ANEW): Stimuli, instruction manual, and affective ratings. Technical report, Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, 1999.

[6] E. Cambria, D. Olsher, and D. Rajagopal. 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, 2014.

[7] Crowdflower. Data for everyone, 2016. Acessed: 2016-04-10.

[8] C. Fellbaum, editor. *WordNet: an electronic lexical database.* MIT Press, 1998.

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

[10] A. Hannak, E. Anderson, L. F. Barrett, S. Lehmann, A. Mislove, and M. Riedewald. Tweetin âĂŹ in the rain: Exploring societal-scale effects of weather on mood.

[11] D. Haynie. The u.s. and u.k. are the worldâĂŹs most influential countries, survey finds. www.usnews.com/news/best-countries/best-international-influence, 2015. Acessed: 2016-05-23.

[12] M. Hu and B. Liu. 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, 2004. ACM.

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

[14] S. James. Leicester city win the premier league title after a fairytale season, 2016.

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

[16] LikeAlyzer. Likealyzer: Analyze and monitor your facebook pages. http://likealyzer.com/, 2016. Acessed: 2016-09-21.

[17] S. M. Mohammad. #emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 246–255, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[18] S. M. Mohammad and P. D. Turney. 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, 2010. Association for Computational Linguistics.

[19] F. Å. Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903, 2011.

[20] Oxford. Oxford Learner's Dictionaries topic dictionaries, 2016. Acessed: 2016-07-03.

[21] N. Pappas and A. Popescu-Belis. Sentiment analysis of user comments for one-class collaborative filtering over ted talks. In *36th ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2013.

[22] C. Phipps. Brussels: Islamic state launches attacks on airport and station âĂŞ as it happened, 2015.

[23] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves, and F. Benevenuto. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29, 2016.

[24] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank.

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

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

[27] J. Tenuto. Classification accuracy is not enough: More performance measures you can use. http://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/, 2014. Acessed: 2016-08-20.

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

[29] Twitter. Rest api documentation. https://dev.twitter.com/rest/public. Acessed: 2015-10-19.

[30] Twitter. Twitter Developers, 2016. Acessed: 2016-03-08.

[31] R. Valitutti. Wordnet-affect: an affective extension of wordnet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086, 2004.

[32] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan. 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, 2012. Association for Computational Linguistics.

[33] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, pages 34–35, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.