Lecture Notes in Social Networks

Mehmet Kaya · Jalal Kawash
Suheil Khoury · Min-Yuh Day *Editors*

Social Network Based Big Data Analysis and Applications

Springer

Lecture Notes in Social Networks

More information about this series at http://www.springer.com/series/8768

Mehmet Kaya • Jalal Kawash • Suheil Khoury
Min-Yuh Day
Editors

Social Network Based Big Data Analysis and Applications

Springer

Editors
Mehmet Kaya
Department of Computer Engineering
Firat University
Elazig, Turkey

Jalal Kawash
Department of Computer Science
University of Calgary
Calgary, AB, Canada

Suheil Khoury
American University of Sharjah
Sharjah, United Arab Emirates

Min-Yuh Day
Tamkang University
Taipei, Taiwan

ISSN 2190-5428 ISSN 2190-5436 (electronic)
Lecture Notes in Social Networks
ISBN 978-3-319-78195-2 ISBN 978-3-319-78196-9 (eBook)
https://doi.org/10.1007/978-3-319-78196-9

Library of Congress Control Number: 2018940255

Preface

Social networks (SN) have brought an unprecedented revolution in how people interact and socialize. SN are used not only as a lifestyle but also in various other domains, including medicine, business, education, politics, and activism. The number of SN amounts to billions of users. At the beginning of 2016, Twitter claimed to have 313 million monthly active users. As of the third quarter of 2017, Facebook had slightly more than 2 billion monthly active users. Online social media (OSM), media produced by SN users, has offered a real and viable alternative to conventional mainstream media. OSM is likely to provide "raw," unedited information, and the details can be overwhelming with the potential of misinformation and disinformation. Yet, OSM is leading to the democratization of knowledge and information. OSM is allowing almost any citizen to become a journalist reporting on specific events of interest. This is resulting in unimaginable amounts of information being shared among huge numbers of OSM participants. For example, Facebook users are generating several billion "likes" and more than 100 million posted pictures in a single day. Twitter users are producing more than 6000 tweets per second. The size of the data generated presents increasing challenges to mine, analyze, utilize, and exploit such content. This book includes eleven contributions that examine several topics related to data analysis and social networks. Applications include sentiment dictionaries, malicious content identification, video recapping, cancer biomarkers, face detection, pattern detection, and cell phone subscription predictions. What follows is a quick summary of each of these chapters.

Nuno Guimarães, Luís Torgo, and Álvaro Figueira complement traditional sentiment dictionaries with a system for lexicon expansion, extracting and classifying domain- and time-specific terms with sentiment based on public opinion. Domain- and time-specific lexicons improve the performance of sentiment analysis methods on short informal texts, such as tweets. The proposed system can generate dictionaries, on a daily basis, to complement the more traditional sentiment lexicons.

Prateek Dewan, Shrey Bagroy, and Ponnurangam Kumaraguru address the issue of identifying malicious content on Facebook, such as publishing untrustworthy information, misleading content, adult and child unsafe content, and scams. The

identified 627 malicious pages revealed through spatial and temporal analysis dominant presence of politically polarized entities engaging in spreading content from untrustworthy domains. Multiple supervised learning algorithms and multiple feature sets are evaluated, and they find that artificial neural networks trained on a fixed sized bag-of-words perform the best in identifying such malicious pages.

Automatic generation of video recaps and summaries is the subject of the chapter by Xavier Bost, Vincent Labatut, Serigne Gueye, and Georges Linarès. They propose narrative smoothing, a method for the extraction of dynamic social networks of video characters. They introduce an algorithm to estimate verbal interactions from a sequence of spoken segments. The data used are a corpus of 109 TV series episodes from three popular TV shows: Breaking Bad, Game of Thrones, and House of Cards.

Gabriela Jurca, Omar Addam, Jon Rokne, and Reda Alhajj study the assessment of candidates for academic positions or for promotion. They employ social network analysis and community detection to measure the influence and diversity of members, within the Department of Computer Science at the University of Calgary. Different measures between various ranks in the department are presented and discussed.

In another chapter, Gabriela Jurca, Omar Addam, Jon Rokne, and Reda Alhajj study biomarkers used to diagnose prostate cancer. They used text mining to provide a tool to examine whether biomarkers are emerging or decreasing in terms of publication popularity. They also provide a tool to examine the increasing or decreasing popularity of gene families with respect to prostate cancer research. Selected biomarkers which have been labeled as emerging in qualitative reviews are then evaluated.

The spread of influence in complex networks is the subject of the chapter by Arun Sathanur, Mahantesh Halappanavar, Yalin Sagduyu, and Yi Shi. They consider the problem of modeling the spread of influence and the identification of influential entities in a complex network with nodal activation, intrinsic or external through neighbors. They approach mining for the influential nodes through influence maximization. One of the findings is how influential content creators can drive engagement on social media platforms.

Yingbo Zhu, Zhenhua Huang, Zhenyu Wang, Linfeng Luo, and Shuang Wu revisit Spiral of Silence in the context of social networks with real information diffusion data. They analyze four information diffusion tree metrics: width, depth, message sentiment, and modularity. Based on Spiral of Silence, polarity prediction of users' review without considering semantic meaning of content is proposed and discovered. Their results indicate that opinions of people in propagation are impacted by the social environment. The Anti-Spiral of Silence is also found to play a significant role in leading rational public opinion and revealing truth in social networks.

Prediction of mobile service subscription types is entertained by Yongjun Liao, Wei Du, Márton Karsai, Carlos Sarraute, Martin Minnoni, and Eric Fleury, specifically the behavioral differences between prepaid and postpaid customers. The findings are used to provide methods that detect the subscription type of customers

by using information about their personal call statistics and their egocentric networks. This allows this classification problem to be treated as a problem of graph labeling, which can be solved by max-flow, min-cut algorithms. The chapter also aims at inferring the subscription type of customers, using node attributes, and a two-ways indirect inference method based on observed hemophiliac structural correlations.

Konstantinos F. Xylogiannopoulos, Panagiotis Karampelas, and Reda Alhajj take on real-time detection of all repeated patterns in a big data stream. A new data structure is introduced: LERP Reduced Suffix Array with a new detection algorithm. This allows the detection of all repeated patterns in a string in a very short time. Specifically, their results show analysis of one million data points and a sliding window of groups of three subsequences of the same size simultaneously with detection in about 300 ms.

Cold start in a dating recommendation service is addressed by Mo Yu, Xiaolong Zhang, Dongwon Lee, and Derek Kreager. They approach this challenge by proposing a novel community-based recommendation framework. Detecting communities to which existing users belong and by matching new users to these communities, the proposed method improves on existing recommendation methods.

The last chapter by Salim Afra and Reda Alhajj studies the performance of face clustering approaches using different feature extraction techniques. Best practices for face recognition of terrorists and criminals are entertained. Performance evaluation for various feature extraction techniques and clustering algorithms using four datasets is also studied.

To conclude this preface, we would like to thank the authors who submitted papers and the reviewers who provided detailed constructive reports which improved the quality of the papers. Various people from Springer deserve great credit for their help and support in all the issues related to publishing this book.

Elazig, Turkey — Mehmet Kaya
Calgary, AB, Canada — Jalal Kawash
Sharjah, United Arab Emirates — Suheil Khoury
Taipei, Taiwan — Min-Yuh Day
November 2017

Contents

Twitter as a Source for Time- and Domain-Dependent Sentiment Lexicons

Nuno Guimarães, Luís Torgo, and Álvaro Figueira

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 because they (1) usually have similar meanings among different domains and (2) are the main indicators of subjectivity in the text. This can lead to a problem in the classification of short informal texts since sometimes the absence of these types of parts of speech does not mean an absence of sentiment. Therefore, our hypothesis states that knowledge of terms regarding certain events and respective sentiment (public opinion) can improve the task of sentiment analysis. Consequently, to complement traditional sentiment dictionaries, we present a system for lexicon expansion that extracts the most relevant terms from news and assesses their positive or negative score through Twitter. Preliminary results on a labelled dataset show that our complementary lexicons increase the performance of three state-of-the-art sentiment systems, therefore proving the effectiveness of our approach.

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

N. Guimarães (✉) · Á. Figueira
CRACS - INESC TEC & University of Porto, Porto, Portugal
e-mail: nuno.r.guimaraes@inesctec.pt; arf@dcc.fc.up.pt

L. Torgo
LIAAD - INESC TEC & University of Porto, Porto, Portugal
e-mail: ltorgo@dcc.fc.up.pt

M. Kaya et al. (eds.), *Social Network Based Big Data Analysis and Applications*,
Lecture Notes in Social Networks, https://doi.org/10.1007/978-3-319-78196-9_1

caused mostly because of the way users share their opinion through short comments or texts, in several different domains. In addition, the large quantity of data available and the quickness of its extraction have recently promoted sentiment analysis to a "hot topic" research subject.

One of the key factors for a precise and correct sentiment analysis classification are sentiment lexicons or sentiment dictionaries. These consist in list of words, mainly adjectives and verbs, with an associated sentiment value (e.g., "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 analyzed text.

Currently, there are several automatic and manually labelled sentiment dictionaries. 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 imply the absence of sentiment. People often tend to use common knowledge to express an opinion without the use of sentiment 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 [30]), but no opinion words to support it. This is because the opinion is expressed using facts regarding Paris 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 [11].

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." The meaning of "Prince" in this particular sentence is specific to the entertainment/music domain (since it is the name of a well-known musician) and not the more general concept (i.e., a member of royalty).

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 organizations. 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.

Therefore, our research hypothesis states: "Can domain and time specific lexicons improve the performance 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 more traditional sentiment lexicons.

2 Related Work

There have been several approaches to the creation and/or expansion of sentiment dictionaries. We can classify them as manually labelled, thesaurus-based, and corpus-based approaches.

Manually labelled sentiment dictionaries rely on human annotators to assess the score on each entry. The author in [27] selected a set of words from several affective word lists (like ANEW [4]), added slang and obscene terms, and manually labelled with a sentiment score ranging from −5 to 5. Another work [19] takes a similar approach. The authors create a manually labeled sentiment dictionary by inspecting already well-established lexicons and adding acronyms and slang words. Then, recurring to Amazon Mechanical Turk [1], they assess each word sentiment using ten 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. Another approach [25] classifies words with emotion (e.g., anger, sadness, and happiness) and polarity (positive/negative). The terms were extracted from a combination of The Macquarie Thesaurus [5], General Inquirer [34], and WordNet Affect Lexicon [43].

Corpus-based approaches rely on the use of a text corpus already labeled (e.g., in a semi-supervised or unsupervised fashion) to create or expand a sentiment lexicon. One of the first works conducted in this area was using a small seed lexicon and conjunctions (such as "and" or "but") to determine the polarity of adjectives [16]. A more recent work [15] presents a methodology to create Twitter corpus-based lexicon. The 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 are included in the lexicon with a positive or negative value. A similar approach is presented in [24]. 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. The authors in [31] use a different approach. Using a small seed lexicon, they are capable to extract sentiment words and features from reviews and expand domain-specific lexicons. The method consists in defining direct and indirect relations between words in sentences (with the help of a POS-tagger) and then, using a set of rules, extract sentiment and feature words. To assign the polarity of the sentiment words extracted, the authors rely on observations such as assign the same polarity for a feature in a review and the same polarity for a sentiment word in a domain-specific corpus.

Finally, thesaurus-based approaches use word resources like WordNet [10] for expanding a small sentiment lexicon (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 and the opposite to the antonyms [8]. Several other studies [18, 20] use WordNet to expand sentiment or create sentiment lexicons, making it one of the most used resources for the creation or expansion of dictionaries.

In other work [26], the authors expand a sentiment lexicon in a two-step procedure. First, they generate the seed lexicon by using eleven affix patterns (e.g., the affix "dis" is used to detect the pair dishonest-honest) and then they use The Macquarie thesaurus to expand the previous defined dictionary. This method is different from the most since it generates and classifies automatically the seed lexicon. The results show that the number of correct entries is far superior to the ones provided by SentiWordNet.

Nevertheless, none of these 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 System Workflow

As it was mentioned before, it is the goal of this work to assess if time- and domain-specific sentiment lexicons can improve the state-of-the-art sentiment analysis methods. In this section, we describe the workflow of the system developed to extract these lexicons automatically.

Our system is divided into two main components: the term extraction and term sentiment evaluation. First we select six of the most common news categories to extract our time- and domain-dependent terms: "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 referring to that term, we assess the public opinion of it. Our hypothesis is that the overall sentiment of the *corpus* extracted using the term as a keyword corresponds to the sentiment of the term.

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 United States and UK as the two most influential countries in the world according to several different factors [17]. 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. 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 lowercase. Then, we build three lists by extracting all unigrams, bigrams, and trigrams in order of frequency. Through experimentation, we realize 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 [2]. 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 by Oxford's Topic Dictionaries [28]. Furthermore, we also exclude possible sentiment words using AFINN lexicon [27]. 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 the presence of an apostrophe ("Clinton"/"Clinton's").

The POS-Tagger filter is only applied to unigrams whereas the other filters are used in all lists, since terms with two or more words already imply a certain context. The last filter is applied to the three 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 an exact match on the queries, incomplete or irrelevant terms are unlikely to reach the minimum number of tweets. The workflow of this component is presented in Fig. 1.

3.1.1 Term Extraction Evaluation

To evaluate the term extraction component, we conduct an experimental survey to determine if the terms were up to date and belonged to the domain from where they were retrieved. The survey was conducted during the time period of 2 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 user could evaluate being only limited to the full extension of the term list extracted. In addition, the terms are extracted from a

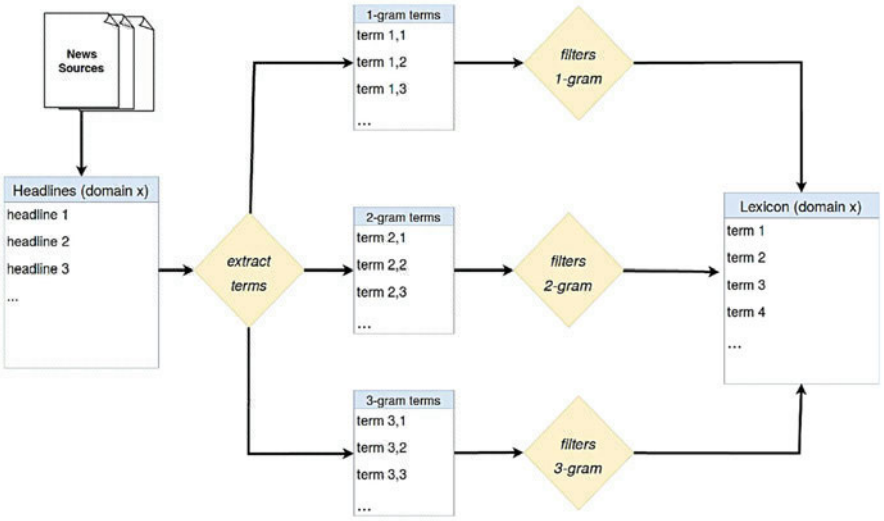

Fig. 1 Term extraction component

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 approximately 60 different users consisting mostly of university students. We discarded all results whose response was "I don't know" which correspond to 5.5% of all evaluations. Furthermore, we only considered terms that had at least three evaluations and we consider our groundtruth 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 correctly assess the domain or time of the term. Although more or less expected (since we are retrieving terms from categories such as news headlines), these results provide strong empirical evidence for our term selection method.

3.2 Term Sentiment Evaluation

The second component of our system determines the sentiment or public opinion of the extracted terms. The majority of corpus-based approaches rely only on Twitter to extract the terms and classify them with sentiment. To the best of our knowledge, this is the first system for lexicon expansion that extracts terms from one source (news headlines) and assesses the sentiment on other (Twitter).

To determine the sentiment of each term extracted, we use the term as a key word in the Twitter API [42]. We then retrieve 100 tweets related to it. This number was achieved by experimental procedures which took into account the restrictions imposed by the Twitter API 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 and in the English language. In addition, we use the parameters provided by the Twitter REST API [41] to retrieve the most recent tweets. Furthermore, in order to avoid extracting posts by news sources (since we want to analyze 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**: Twitter-based sentiment lexicon expanded from ANEW [4]. It contains words that are frequently used in this social network such as Internet slang and offensive words [27].
- **Emolex**: Manually created emotion lexicon using crowdsourcing. The terms were extracted from a combination of The Macquarie Thesaurus [5], General Inquirer [34], and WordNet Affect Lexicon [43]. Although the words were classified with emotion and polarity, only the second was used for this method [25].
- **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 [15].
- **Happiness Index**: Uses words from ANEW that were manually classified with a 1–9 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 [13].
- **MPQA (or Opinion Finder)**: It is a machine-learning model to detect subjectivity and consequently, the polarity of a sentence based on sentiment clues [45]. 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 [24].

- **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 [18].
- **SANN**: Uses the sentiment lexicon of MPQA along with polarity shifters, negation, and amplifiers to build a sentence-level sentiment classifier. It was originally used in user comments present in Ted Talks videos [29].
- **Sasa**: It is a supervised method based on a Naive-Bayes approach. It uses the unigram features of each tweet. This method was originally used to detect sentiment on tweets in real time during the U.S. 2012 election [44].
- **SenticNet**: Assigns sentiment to common sense concepts to achieve a semantic sentiment analysis approach rather than the most common sentence level [6].
- **Sentiment140 Lexicon**: Is a corpus-based sentiment lexicon extracted from the tweets provided in [12]. 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 enhancement. It provides the best results in gold-standard tweet datasets [38–40].
- **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 [3].
- **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 [35].
- **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 [33].
- **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 (e.g., "yeeeeeaaaaaah"), hashtag evaluation (e.g., detecting negative sentiment in #notverygood), and decomposition of the tweet in n-grams (to be able to distinguish "good" from "not good") [21]
- **Vader**: Directed for microblogging 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) [19].

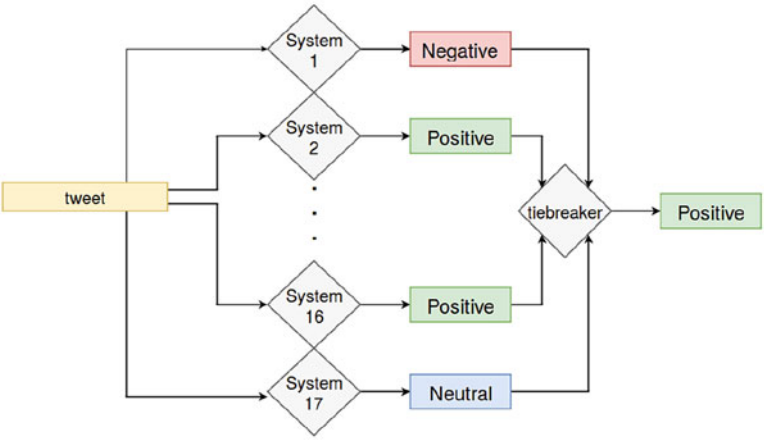

Fig. 2 Ensemble system example

To facilitate the building of the ensemble we used the iFeel framework, which allows the selection of specific sentiment analysis methods [23, 32]. 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, assuming different setups in terms of ensemble composition are possible, we define that in this case, the neutral value is returned.

An example of the behaviour of the ensemble system can be seen in Fig. 2.

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, Opinion Lexicon, Sentiment 140, and SentiWordNet), iFeel uses Vader rule-based system to push forward the performance of these lexicons [32].

In previous work [14], we conclude that term classification using three classes (Negative/Neutral/Positive) was needed. Therefore, to determine the sentiment of the term based on the *corpus* of tweets extracted we used the following formula:

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

where c is the respective sentiment class (Negative, Neutral, or Positive).

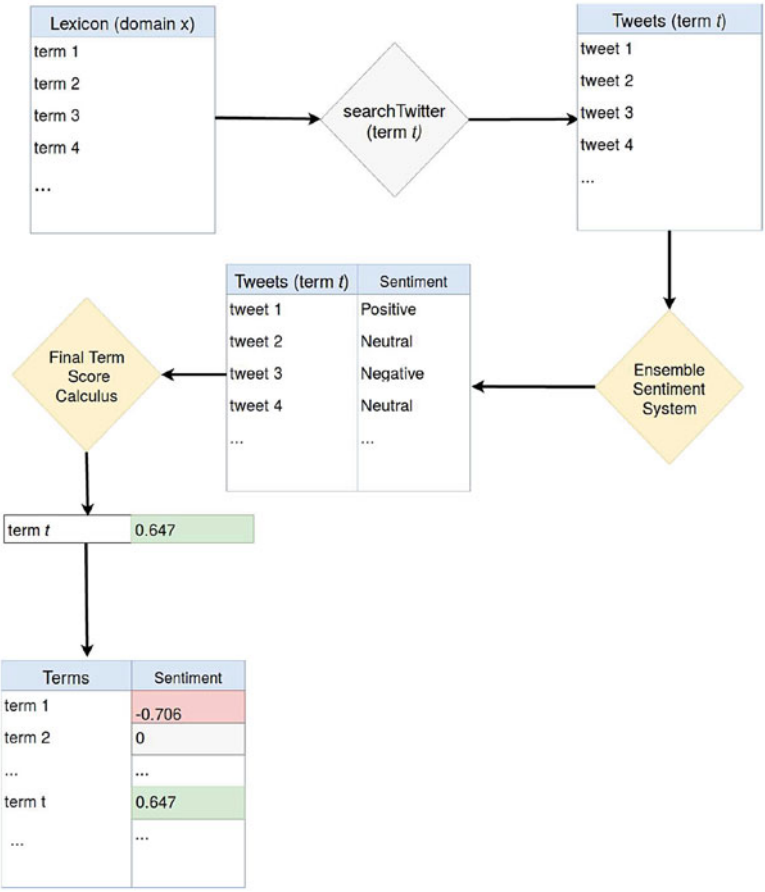

Fig. 3 Term sentiment evaluation

This formula gives us the confidence of the sentiment in each one of the classes. Consequently, the sentiment of the term is assessed using the class with the maximum score value. If the neutral class has the maximum score, we ignore the confidence value and assign a 0 score to that term. Otherwise, we use the confidence value of the class returned by the formula (multiplying it by -1 in the cases where the maximum score belongs to the negative class). This way, the scores for the created lexicons will range from $[-1, 1]$.

Figure 3 represents the workflow of the term sentiment evaluation component.

3.2.1 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 a set of tweets referring to the 2016 GOP debate (GOP), Google self-driving cars (SDC), Coachella line-up announcement (COACH), United States airlines (USAIR), and the Deflategate scandal (NFL). Since we want a good performance across all domains and classes, we select an equal number of entries in each dataset and balance the three classes available. Therefore, each of the previous mentioned datasets has 1200 tweets (400 positive, 400 neutral, and 400 negative). We assess our ensemble results in terms of accuracy and average $F1$-score in different classes and different domains. First, we compare our ensemble system with the top three more accurate systems in each domain and using the aggregation of all datasets, represented by the "Average" column. The results are presented in Table 1. When compared with each stand-alone system, the ENS17 ensemble is in the top three most accurate 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 with the exception of the NFL and GOP. We assume that this is due to the large numbers of jokes in the tweets included [37] and political irony, which may make difficult the classification task in the individual systems. However, if we look at the accuracy across all domains (in other words, the average accuracy in all datasets), the ENS17 outperforms the best individual systems. We stress out that the "Average" column represents the values for the top three individual systems that perform better in the concatenation of all datasets, and not the average of the remaining columns.

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

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

	Dataset accuracy						
	GOP	SDC	APPLE	USAIR	COACH	NFL	Average
Top individual systems (on each dataset)							
First system	49.0	**53.0**	**69.5**	**62.0**	**46.5**	35.2	48.8
Second system	48.1	51.3	61.3	59.3	46.1	34.1	48.3
Third system	47.5	47.5	54.6	58.4	45.3	33.4	48.0
Ensemble system							
ENS17	**51.0**	51.6	60.4	60.0	45.2	**45.2**	**52.2**

The results are presented in Table 1 and the best score for each system in highlighted in bold

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

	Class accuracy (%)			
	Negative	Neutral	Positive	Total
Best overall individual systems (using accuracy as metric)				
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 system				
ENS17	35.8	72.3	**48.6**	**52.2**

The results are presented in Table 2 and the best score for each system in highlighted in bold

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

	Dataset $F1$-score (%)						
	GOP	SDC	APPLE	USAIR	COACH	NFL	Average
Top individual systems (on each dataset)							
First system	48.6	**52.3**	**69.1**	**61.9**	**44.6**	32.7	47.8
Second system	47.5	50.6	60.9	58.6	43.5	31.8	47.2
Third system	45.8	46.6	55.0	57.6	42.3	30.1	47.6
Ensemble system							
ENS17	**50.3**	50.8	60.4	59.4	42.2	**42.2**	**51.5**

The results are presented in Table 3 and the best score for each system in highlighted in bold

As we can observe, regarding classes, ENS17 is always in the top three systems when comparing with the most accurate individual systems. Furthermore, it achieves the highest accuracy value on the positive class. Since the datasets are balanced in the number of entries and elements in each class, it's no wonder that the all-classes accuracy values are the same as the total accuracy values in all datasets. Since accuracy values can sometimes be misleading [36], we perform the same analysis using the average $F1$-score in each domain. Therefore, selecting the top three systems according to the average $F1$-score and assessing the same metric with our method results in the values presented in Table 3.

Once again it is clear that our ensemble system performs well enough to be in the top three systems using $F1$-score in each dataset. Therefore, it is no surprise that, when considering all datasets, ENS17 achieves an average $F1$-score superior to 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 4. It is easily noticeable that the ensemble system outperforms the individual systems, therefore proving the validity of our approach.

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

	Class $F1$-score (%)			
	Negative	Neutral	Positive	Average
Best overall individual systems (using F1 measure as metric)				
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**

The results are presented in Table 4 and the best score for each system in highlighted in bold

4 Lexicons 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 7 to September 14 2016, evaluated with sentiment on CrowdFlower. The dictionaries from our system were retrieved on 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 change substantially.

4.1 Dataset Description

As it was already mentioned, the dataset combines three 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 [22]. For each post on the defined time interval, we extracted up to a maximum of 20 comments (order by the Facebook "ranked" metric [9]). 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 queries were evaluated with sentiment in our lexicons. Consequently to avoid biased results, we excluded those terms from the dictionaries. The key words 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.

The survey on CrowdFlower consisted in two sentiment questions. The first was "The sentiment expressed in this text is:" To answer, the workers had a Likert scale ranging from 1 to 5 and labeled from "very negative" to "very positive." The second was a follow-up question that stated: "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 Nonfactual Texts

We select AFINN [27], UMIGON [21], and SentiStrength [40] as the sentiment methods to complement with our lexicons since (1) they are well-known systems in the state of the art of sentiment analysis and (2) in the tests previously mentioned they were the systems that individually performed better in terms of accuracy and average $F1$-score.

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 other words, we need to assign a domain for each entry of the dataset. In this experiment, we used the frequency of words on the text that appear on Oxford's Topic Dictionaries [28] combined with the 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.

The addition of the lexicons outputted by our system improved the tested methods in both accuracy and average $F1$-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 F-measure is 1.51%.

We can conclude that, although 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 three analyzed systems.

Table 5 Variation between the sentiment systems with and without the expanded lexicons

Sentiment system	Accuracy %	Average $F1$%
AFINN	+1.12	+0.48
SentiStrength	+1.36	+1.43
Umigon	+3.14	+2.63

Table 6 Variation between the sentiment systems with and without the expanded lexicons 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

The main reason why our approach does not improve on a greater scale the results from traditional sentiment analysis lexicons is due to the specificity of the problem we are trying to solve. In fact, if we go further in our analysis 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, we can really tackle the problem we are trying to solve. The filtered dataset contains 215 entries and results of these specific cases can be consulted in Table 6.

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 their $F1$-score increases 9.81% and 12.55%, respectively. This demonstrates that it is important not only to consider our system sentiment dictionaries but also that our term sentiment analysis is capable of accurately classifying the terms. Therefore, the results show that the addition of our lexicons improves the performance of state-of-the-art sentiment systems. Furthermore, they make a major difference when we analyze subjective texts whose sentiment is not determined by opinion words (such as verbs and adjectives) included in traditional sentiment lexicons.

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 tweet sentiment classifier (combining 17 state-of-the-art sentiment analysis methods to analyze tweets regarding the terms). The final output is seven different sentiment dictionaries that are retrieved on a daily basis.

Next, we complement three state-of-the-art sentiment systems (AFINN, UMIGON, and SentiStrength) with the dictionaries outputted from our method. 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 platform.

The results achieved indicate a coherent improvement in all methods. In addition, when the term for assessing the sentiment is not included in sentiment dictionaries, the importance of our domain- and time-specific lexicons increases significantly, proving that our approach can increment the performance of sentiment methods in these specific cases. These results allow us to conclude that, although our lexicons try to solve a specific problem, they are effective on that task and do not compromise the performance of traditional sentiment analysis methods.

It is important to notice that the lexicons generated by this framework do not replace traditional sentiment lexicons. The goal is to complement them, by having domain and time sentiment terms attached to state-of-the-art methods. This approach goes against what is normally proposed in the area, since the majority of works have focus in building sentiment lexicons and compare them with state-of-the-art methods.

Therefore, we do believe that our framework can keep up and complement even the more recent proposals on sentiment methods. However, our lexicons integration is simpler in rule-based approaches, since that supervised methods would require repeating the learning phase with the extended lexicons.

For this reason, in future work, it would be interesting to discover the "expiration date" of the lexicons generated. In other words, to analyze how the time difference between the extraction of the lexicons and the source of the text affects the performance on sentiment analysis tasks. This can be particularly useful to integrate the dictionaries created by our framework in supervised methods without having to create new models on, for example, a daily basis.

In addition, although the results achieved are promising, in future work to conduct our evaluation we intend to use domain-specific tweet datasets (instead of using an automatic domain disambiguation). We do believe that with a domain already defined, the performance of the lexicons will increase. We also plan to further extend our system by adding a geographical component to the dictionaries generated. We intend to use news sources as well as tweets 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 analyzed.

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.