RESEARCH

A Benchmark Comparison of State-of-the-Practice Sentiment Analysis Methods

Filipe N Ribeiro[1,2*], Matheus Araújo[1], Pollyanna Gonçalves[1], Marcos André Gonçalves[1] and Fabrício Benevenuto[1]

*Correspondence: filiperibeiro@dcc.ufmg.br
[1]Computer Science Department, Federal University of Minas Gerais, Belo Horizonte, Brazil
[2]Computer and Systems Department, Federal University of Ouro Preto, Joao Monlevade, Brazil
Full list of author information is available at the end of the article

Abstract

In the last few years thousands of scientific papers have investigated sentiment analysis, several startups that measure opinions on real data have emerged and a number of innovative products related to this theme have been developed. There are multiple methods for measuring sentiments, including lexical-based approaches and supervised machine learning methods. Despite the vast interest on the theme and wide popularity of some methods, it is unclear which method is better for identifying the polarity (i.e., positive or negative) of a message. Accordingly, there is a strong need to conduct a thorough apple-to-apple comparison of sentiment analysis methods, *as they are used in practice*, across multiple datasets originated from different data sources. Such a comparison is key for understanding the potential limitations, advantages, and disadvantages of popular methods. This article aims at filling this gap by presenting a benchmark comparison of twenty-two popular sentiment analysis methods (which we call the state-of-the-practice methods). Our evaluation is based on a benchmark of eighteen labeled datasets, covering messages posted on social networks, movie and product reviews, as well as opinions and comments in news articles. Our results highlight the extent to which the prediction performance of these methods varies considerably across datasets. Aiming at boosting the development of this research area, we open the methods' codes and datasets used in this article, deploying them in a benchmark system, which provides an open API for accessing and comparing sentence-level sentiment analysis methods.

Keywords: Sentiment Analysis; Benchmark; Methods Evaluation

Introduction

Sentiment analysis has become an extremely popular tool, applied in several analytical domains, especially on the Web and social media. To illustrate the growth of interest in the field, Figure 1 shows the steady growth on the number of searches on the topic, according to Google Trends[1], mainly after the popularization of online social networks (OSNs). More than 7,000 articles have been written about sentiment analysis and various startups are developing tools and strategies to extract sentiments from text [1].

The number of possible applications of such a technique is also considerable. Many of them are focused on monitoring the reputation or opinion of a company or a brand with

[1] `https://www.google.com/trends/explore#q=sentiment%20analysis`

the analysis of reviews of consumer products or services [2]. Sentiment analysis can also provide analytical perspectives for financial investors who want to discover and respond to market opinions [3, 4]. Another important set of applications is in politics, where marketing campaigns are interested in tracking sentiments expressed by voters associated with candidates [5].

Due to the enormous interest and applicability, there has been a corresponding increase in the number of proposed sentiment analysis methods in the last years. The proposed methods rely on many different techniques from different computer science fields. Some of them employ machine learning methods that often rely on supervised classification approaches, requiring labeled data to train classifiers [6]. Others are lexical-based methods that make use of predefined lists of words, in which each word is associated with a specific sentiment. The lexical methods vary according to the context in which they were created. For instance, LIWC [7] was originally proposed to analyze sentiment patterns in formally written English texts, whereas PANAS-t [8] and POMS-ex [9] were proposed as psychometric scales adapted to the Web context.

Overall, the above techniques are acceptable by the research community and it is common to see concurrent important papers, sometimes published in the same computer science conference, using completely different methods. For example, the famous Facebook experiment [10] which manipulated users feeds to study emotional contagion, used LIWC [7]. Concurrently, Reis *et al.* used SentiStrength [11] to measure the negativeness or positiveness of online news headlines [12, 13], whereas Tamersoy [14] explored Vader's lexicon [15] to study patterns of smoking and drinking abstinence in social media.

As the state-of-the-art has not been clearly established, researchers tend to accept any popular method as a valid methodology to measure sentiments. However, little is known about the relative performance of the several existing sentiment analysis methods. In fact, most of the newly proposed methods are rarely compared with all other pre-existing ones using a large number of existing datasets. This is a very unusual situation from a scientific perspective, in which benchmark comparisons are the rule. In fact, most applications and experiments reported in the literature make use of previously developed methods exactly how they were released with no changes and adaptations and with none or almost none parameter setting. In other words, the methods have been used as a black-box, without a deeper investigation on their suitability to a particular context or application.

To sum up, existing methods have been widely deployed for developing applications without a deeper understanding regarding their applicability in different contexts or their advantages, disadvantages, and limitations in comparison with each another. Thus, there is a strong need to conduct a thorough apple-to-apple comparison of sentiment analysis methods, *as they are used in practice*, across multiple datasets originated from different data sources.

This *state-of-the-practice* situation is what we propose to investigate in this article. We do this by providing a thorough benchmark comparison of *twenty-two state-of-the-practice* methods using *eighteen* labeled datasets. In particular, given the recent popularity of online social networks and of short texts on the Web, many methods are focused in detecting sentiments at the sentence-level, usually used to measure the sentiment of small sets of sentences in which the topic is known *a priori*. We focus on such context – thus, our datasets cover messages posted on social networks, movie and product reviews, and opinions and comments in news articles, Ted talks, and blogs. We survey an extensive literature on sentiment analysis to identify existing sentence-level methods covering several different techniques. We contacted authors asking for their codes when available or we implemented existing methods when they were unavailable but could be reproduced based on their descriptions in the original published paper.

Our experimental results unveil a number of important findings. First, we show that there is no single method that always achieves the best prediction performance for all different datasets, a result consistent with the "there is no free lunch theorem" [16]. We also show that existing methods vary widely regarding their agreement, even across similar datasets. This suggests that the same content could be interpreted very differently depending on the choice of a sentiment method. We noted that most methods are more accurate in correctly classifying positive than negative text, suggesting that current approaches tend to be biased in their analysis towards positivity. Finally, we quantify the relative prediction performance of existing efforts in the field across different types of datasets, identifying those with higher prediction performance across different datasets.

Based on these observations, our final contribution consists on releasing our gold standard dataset and the codes of the compared methods[2]. We also created a Web system through which we allow other researchers to easily use our data and codes to compare results with the existing methods. More importantly, by using our system one could easily test which method would be the most suitable to a particular dataset and/or application. We hope that our tool will not only help researchers and practitioners for accessing and comparing a wide range of sentiment analysis techniques, but can also help towards the development of this research field as a whole.

The remainder of this paper is organized as follows. In Section 2, we briefly describe related efforts. Then, in Section 3 we describe the sentiment analysis methods we compare. Section 4 presents the gold standard data used for comparison. Section 5 summarizes our results and findings. Finally, Section 6 concludes the article and discusses directions for future work.

Background and Related Work

Next we discuss important definitions and justify the focus of our benchmark comparison. We also briefly survey existing related efforts that compare sentiment analysis methods.

Focus on Sentence-Level Sentiment Analysis

Since sentiment analysis can be applied to different tasks, we restrict our focus on comparing those efforts related to detect the polarity (i.e. positivity or negativity) of a given short text (i.e. sentence-level). Polarity detection is a common function across all sentiment methods considered in our work, providing valuable information to a number of different applications, specially those that explore short messages that are commonly available in social media [1].

Sentence-level sentiment analysis can be performed with supervision (i.e. requiring labeled training data) or not. An advantage of supervised methods is their ability to adapt and create trained models for specific purposes and contexts. A drawback is the need of labeled data, which might be highly costly, or even prohibitive, for some tasks. On the other hand, the lexical-based methods make use of a pre-defined list of words, where each word is associated with a specific sentiment. The lexical methods vary according to the context

[2]Except for paid methods

in which they were created. For instance, LIWC [7] was originally proposed to analyze sentiment patterns in English texts, whereas PANAS-t [8] and POMS-ex [9] are psychometric scales adapted to the Web context. Although lexical-based methods do not rely on labeled data, it is hard to create a unique lexical-based dictionary to be used for all different contexts.

We focus our effort on evaluating unsupervised efforts as they can be easily deployed in Web services and applications without the need of human labeling or any other type of manual intervention. As described in Section 3, some of the methods we consider have used machine learning to build lexicon dictionaries or even to build models and tune specific parameters. We incorporate those methods in our study, since they have been released as black-box tools that can be used in an unsupervised manner.

Existing Efforts on Methods Comparison

Despite the large number of existing methods, only a limited number of them have performed a comparison among sentiment analysis methods, usually with restricted datasets. Overall, lexical methods and machine learning approaches have been evolving in parallel in the last years, and it comes as no surprise that studies have started to compare their performance on specific datasets and use one or another strategy as baseline for comparison. A recent survey summarizes several of these efforts [17] and conclude that a systematic comparative study that implements and evaluates all relevant algorithms under the same framework is still missing in the literature. As new methods emerge and compare themselves only against one, at most two other methods, using different evaluation datasets and experimental methodologies, it is hard to conclude if a single method triumphs over the remaining ones, or even in specific scenarios. To the best of our knowledge, our effort is the first of kind to create a benchmark that provides such thorough comparison.

An important effort worth mentioning consists of an annual workshop – The International Workshop on Semantic Evaluation (SemEval). It consists of a series of exercises grouped in tracks, including sentiment analysis, text similarity, among others, that put several together competitors against each other. Some new methods such as Umigon [18] have been proposed after obtaining good results on some of these tracks. Although, SemEval has been playing an important role for identifying relevant methods, it requires authors to register for the challenge and many popular methods have not been evaluated in these exercises. Additionally, SemEval labeled datasets are usually focused on one specific type of data, such as tweets, and do not represent a wide range of social media data. In our evaluation effort, we consider one dataset from SemEval 2013 and two methods that participated in the competition in that same year.

Finally, in a previous effort [19], we compared eight sentence-level sentiment analysis methods, based on one public dataset used to evaluate SentiStrength [11]. This article largely extends our previous work by comparing a much larger set of methods across many different datasets, providing a much deeper benchmark evaluation of current popular sentiment analysis methods. The methods used in this paper were also incorporated as part of an existing system, namely ifeel [20].

Sentiment Analysis Methods

This section provides a brief description of the twenty-two sentence-level sentiment analysis methods investigated in this article. Our effort to identify important sentence-level sentiment analysis methods consisted of systematically search for them in the main conferences in the field and then checking for papers that cited them as well as their own references. Some of the methods are available for download on the Web; others were kindly shared by their authors under request; and a small part of them were implemented by us based on their descriptions in the original paper. This usually happened when authors shared only the lexical dictionaries they created, letting the implementation of the method that use the lexical resource to ourselves.

Table 1 and Table 2 present an overview of these methods, providing a description of each method as well as the techniques they employ (L for Lexicon Dictionary and ML for Machine Learning), their outputs (e.g. -1,0,1, meaning negative, neutral, and positive, respectively), the datasets they used to validate, the baseline methods used for comparison and finally lexicon details. The methods are organized in chronological order to allow a better overview of the existing efforts over the years. We can note that the methods generate different outputs formats. We colored in blue the positive outputs, in black the neutral ones, and in red those that are negative.

Adapting Lexicons for the Sentence Level Task

Since we are comparing sentiment analysis methods on a sentence-level basis, we need to work with mechanisms that are able to receive sentences as input and produce polarities as output. Some of the approaches considered in this paper, shown in Table 2, are complex dictionaries built with great effort. However, a lexicon alone has no natural ability to infer polarity in sentence level tasks. The purpose of a lexicon goes beyond the detection of polarity of a sentence [1, 21], but it can also be used with that purpose [22, 23].

Several existing sentence-level sentiment analysis methods, like Vader [15] and SO-CAL [24], combine a lexicon and the processing of the sentence characteristics to determine a sentence polarity. These approaches make use of a series of intensifiers, punctuation transformation, emoticons, and many other heuristics.

Thus, to evaluate each lexicon dictionaries as the basis for a sentence-level sentiment analysis method, we considered the Vader's implementation. In other words, we used Vader's code for determining if a sentence is positive or not considering different lexicons.

Vader's heuristics were proposed based on qualitative analyses of textual properties and characteristics which affect the perceived sentiment intensity of the text. Vader's author identified five heuristics based on grammatical and syntactical cues to convey changes to sentiment intensity that go beyond the bag-of-words model. The heuristics include treatments for: 1) punctuation (e.g number of '!'s); 2) capitalization (e.g "I HATE YOU" is more intense than "i hate you"); 3) degree modifiers (e.g "The service here is extremely good" is more intense than "The service here is good"); 4) constructive conjunction "but" to shift the polarity; 5) Tri-gram examination to identify negation (e.g "The food here isn't really all that great."). We choose Vader as a basis for such heuristics as it is one of the most recent methods among those we considered. Moreover, it is becoming widely used, being even implemented as part of the well-known NLTK python library[3].

We applied such heuristics to the following lexicons: Emolex, EmoticonsDS, NRC Hashtag, Opinion Lexicon, PANAS-t, Sentiment 140 and SentiWordNet. We notice that those strategies drastically improved results of the lexicons for sentence-level sentiment analysis in comparison with a simple baseline approach that averages the occurrence of positive and negative words to classify the polarity of a sentence. Table 2 has also a column (Lexicon

[3]http://www.nltk.org/_modules/nltk/sentiment/vader.html

size) that describes the number of terms the proposed dictionary contains and column C (changed) indicates some methods we slightly modified to adequate their output formats to the polarity detection task, as described next.

Output Adaptations

It is worth noticing that the output of each method varies drastically depending on the goal it was developed for and the approach it employs. PANAS-t, for instance, associates each word with eleven moods as described in Table 1 and it was designed to track any increase or decrease in sentiments over time. Emolex lexicon provides the association of each word with eight sentiments. The word 'unhappy' for example is related to anger, disgust, and sadness and it is not related to joy, surprise, etc. SentiWordNet links each word with a synset (i.e. a set of synonyms) characterized by a positive and a negative score, both of them represented with a value between 0 and 1.

The aforementioned lexicons were used as dictionary input to Vader's code. We had to adapt the way the words are processed as follows. For PANAS-t we assumed that joviality, assurance, serenity, and surprise are positive affect. Fear, sadness, guilt, hostility, shyness, and fatigue are negative affect. Attentiveness was considered neutral. In the case of Emolex, we considered two other entries released by the authors. The first one defines the positiviy of a word (0 or 1) and the second characterizes the negativity (0 or 1). For SentiWordNet we calculate an overall score to the word by subtracting the positive value from negative value defined to that word. For example, the positive value for the word faithful is 0.625 while its negative score is 0.0. Then the overall score score is 0.625.

Other lexicons included in our evaluation already provide positive and negative scores like SentiWordNet or an overall score ranging from a negative to a positive value. After applying Vader's heuristics for each one of these lexicons we get scores in the same way Vader's output (see Table 2).

Other methods also required some output handling. The available implementation of OpinionFinder [4], for instance, generates polarity outputs (-1,0, or 1) for each sentiment clue found in a sentence so that a single sentence can have more than one clue. We considered the polarity of a single sentence as the sum of the polarities of all the clues. Happiness index uses scores ranging from 1 to 9 indicating the level of happiness in the text – we considered the values one to four as negative, five as neutral, and six to nine as positive.

The outputs from the remaining methods were easily adapted and converted to positive, negative or neutral. SO-CAL and Pattern.en delivery float numbers greater than a threshold, indicating positive, and lesser than the threshold, indicating negative. LIWC, SenticNet, SASA, SANN, SentiStrength, Umigon, Vader and Semantria already provide fixed outputs indicating one of three desired classes while Stanford Recursive Deep Model yields very negative and very positive which in our experiments are handled as negative and positive, respectively.

Paid Softwares

Four out of the twenty-two methods evaluated in this work are closed paid softwares: LIWC, Semantria, SenticNet e SentiStrength. Although SentiStrength is paid it has a free of charge academic license. SenticNet's authors kindly processed all datasets with the commercial version and return the polarities for us. For SentiStrenght we used the Java version

[4] http://mpqa.cs.pitt.edu/opinionfinder/

from May 2013 in a package with all features of the commercial version. For LIWC we acquired the license from 2007 version. Finally, for Semantria we used a trial account free of charge for a limited number of sentences, which was sufficient to run our experiments.

Methods not included

Despite our effort to include in our comparison most of the highly cited and important methods we could not include a few of them for different reasons. Profile of Mood States (POMS) [9] is not available on the Web or under request and could not be re-implemented based on their descriptions in the original papers. The same situation occurs with the Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification [25]. NRC SVM [26] is not available as well, although the lexical resources used by the authors are available and were considered in our evaluation resulting in the methods: NRC Hashtag and Sentiment140. The authors of the Convolutional Neural Network for Modeling Sentences [27] and of the Effective Use of Word Order for Text Categorization with Convolutional Neural Networks [28] have made their source code available but the first one lacks the train files and the second one requires a GPU to execute. There are a few other methods for sentiment detection proposed in the literature and not considered here. Most of them consists of variations of the techniques used by the above methods, such as WordNet-Affect[29] and ANEW [30].

Datasets and Comparison Among Methods

From Table 2 we can note that the validation strategy, the datasets used, and the comparison with baselines performed by these methods vary greatly, from toy examples to large labeled datasets. PANAS-t and Happiness Index use labeled examples to validate their methods, by presenting evaluations of events in which some bias towards positivity and negativity would be expected. PANAS-t is tested with unlabeled Twitter data related to Michael Jackson's death and the release of a Harry Potter movie whereas Happiness Index was used to measure song lyrics happiness from 1967 to 2007. Lexical dictionaries were validated in very different ways. AFINN[31] compared its Lexicon with other dictionaries. Emoticon Distance Supervised [32] used Pearson Correlation between human labeling and the predicted value. SentiWordNet [33] validates the proposed dictionary with comparisons with other dictionaries, but it also used human validation of the proposed lexicon. These efforts attempt to validate the created lexicon, without comparing the lexicon as a sentiment analysis method by itself. Vader [15] compared results with lexical approaches considering labeled datasets from different social media data. SenticNet [34] was compared with SentiStrength [11] with a specific dataset related to patient opinions, which could not be made available. Stanford Recursive Deep Model [35] and SentiStrength [11] were both compared with standard machine learning approaches, with their own datasets.

This scenario, where every new developed solution compares itself with different solutions using different datasets, happens because there is no standard benchmark for evaluating new methods. This problem is exacerbated because many methods have been proposed in different research communities (e.g. NLP, Information Science, Information Retrieval, Machine Learning), exploiting different techniques, with low knowledge about related efforts in other communities. Next, we describe how we created a large gold standard to properly compare all the considered sentiment analysis methods.

Gold Standard Data

A key aspect in evaluating sentiment analysis methods consists of using accurate gold standard labeled datasets. Several existing efforts have generated labeled data produced by experts or non-experts evaluators. Previous studies suggest that both efforts are valid as non-expert labeling may be as effective as annotations produced by experts for affect recognition, a very related task [36]. Thus, our effort to build a large and representative gold standard dataset consists of obtaining labeled data from trustful previous efforts that cover a wide range of sources and kinds of data. We also attempt to assess the "quality" of our gold standard in terms of the accuracy of the labeling process.

Table 3 summarizes the main characteristics of the eighteen exploited datasets, such as number of messages and the average number of words per message in each dataset. It also defines a simpler nomenclature that is used in the remainder of this paper. The table also presents the methodology employed in the classification. Human labeling was implemented in almost all datasets, usually done with the use of non-expert reviewers. Reviews_I dataset relies on five stars rates, in which users rate and provide a comment about an entity of interest (e.g. a movie or an establishment).

Labeling based on Amazon Mechanical Turk (AMT) was used in seven out of the eighteen datasets, while volunteers and other strategies that involve non-expert evaluators were used in ten datasets. Usually, an agreement strategy (i.e. majority voting) is applied to ensure that, in the end, each sentence has an agreed-upon polarity assigned to it. The number of annotators used to build the datasets is also shown in Table 3.

Tweets_DBT was the unique dataset built with a combination of AMT Labeling with Expert validation [37]. They selected 200 random tweets to be classified by experts and compared with AMT results to ensure accurate ratings. We note that the Tweets_Semeval dataset was provided as a list of Twitter IDs, due to the Twitter policies related to data sharing. While crawling the respective tweets, a small part of them could not be accessed, as they were deleted. We plan to release all gold standard datasets in a request basis, which is in agreement with Twitter policies.

In order to assess the extent to which these datasets are trustful, we used a strategy similar to the one used by Tweets_DBT. Our goal was not to redo all the performed human evaluation, but simply inspecting a small sample of them to infer the level of agreement with our own evaluation. We randomly select 1% of all sentences to be evaluated by experts (two of the authors) as an attempt to assess if these gold standard data are really trustful. It is important to mention that we did not have access to the instructions provided by the authors. We also could not get access to small amount of the raw data in a few datasets, which was discarded. Finally, our manual inspection unveiled a few sentences in idioms other than English in a few datasets, such as Tweets_STA and TED, which were obviously discarded.

Column R from Table 3 exhibits the level of agreement of each dataset in our evaluation. After a close look in the cases of disagreement with the evaluations in the Gold standard, we realized that other interpretations could be possible for the given text, finding cases of sentences with mixed polarity. Some of them are strongly linked to original context and are very hard to evaluate. Some NYT comments, for instance, are directly related to the news they were inserted to. We can also note that some of the datasets do not contain neutral messages. This might be a characteristic of the data or even a result of how annotators were instructed to label their pieces of text. Most of the cases of disagreement involve neutral messages. Thus, we considered these cases as well as the amount of disagreement we had with the gold standard data as reasonable and expected.

Comparison Results

Next, we present comparison results for the twenty-two methods considered in this paper based on the eighteen considered gold standard datasets.

Experimental details

At least three distinct approaches have been proposed to deal with sentiment analysis of sentences. The first of them, applied by OpinionFinder and Pattern.en, for instance, splits this task into two steps: (i) identifying sentences with no sentiment, also named as objective vs. neutral sentences and then (ii) detecting the polarity (positive or negative), only for the subjective sentences. Another common way to detect sentence polarity considers three distinct classes (positive, negative and neutral) in a single task, an approach used by Vader, SO-CAL, SANN and others. Finally, some methods like SenticNet and LIWC, classify a sentence as positive or negative only, assuming that only polarized sentences are presented, given the context of a given application. As an example, reviews of products are expected to contain only polarized opinion.

Aiming at providing a more thorough comparison among these distinct approaches, we perform two rounds of tests. In the first we consider the performance of methods to identify 3-class (positive, negative and neutral). The second considers only positive and negative as output and assumes that a previous step of removing the neutral messages needs to be executed firstly. In the 3-class experiments we used only datasets containing a considerable number of neutral messages (which excludes Tweets_RND_II, Amazon, and Reviews_II). Despite being 2-class methods, as highlighted in Table 4, we decided to include LIWC, Emoticons and SenticNet in the 3-class experiments to present a full set of comparative experiments. LIWC, Emoticons, and SenticNet cannot define, for some sentences, their positive or negative polarity, considering it as undefined. It occurs due to the absence in the sentence of emoticons (in the case of Emoticons method) or of words belonging to the methods' sentiment lexicon. As neutral (objective) sentences do not contain sentiments, we assumed, in the case of these 2-class methods, that sentences with undefined polarities are equivalent to neutral sentences.

The 2-class experiments, on the other hand, were performed with all datasets described in Table 3 excluding the neutral sentences. We also included all methods in these experiments, even those that produce neutral outputs. As discussed before, when 2-class methods cannot detect the polarity (positive or negative) of a sentences they usually assign it to an undefined polarity. As we know all sentences in the 2-class experiments are positive or negative, we create the coverage metric to determine the percentage of sentences a method can in fact classify as positive or negative. For instance, suppose that Emoticons' method can classify only 10% of the sentences in a dataset, corresponding to the actual percentage of sentences with emoticons. It means that the coverage of this method in this specific dataset is 10%. Note that, the coverage is quite an important metric for a more complete evaluation in the 2-class experiments. Even though Emoticons presents high accuracy for the classified phrases, it was not able to make a prediction for 90% of the sentences. More formally, coverage is calculated as the number of total sentences minus the number of undefined sentences, all of this divided by the total of sentences, where the number of undefined sentences includes neutral outputs for 3-class methods.

$$Coverage = \frac{\#Sentences - \#Undefined}{\#Sentences}$$

Comparison Metrics

Considering the 3-class comparison experiments, we used the traditional Precision, Recall, and F1 measures for the automated classification.

		Predicted		
		Positive	Neutral	Negative
	Positive	a	b	c
Actual	Neutral	d	e	f
	Negative	g	h	i

Each letter in the above table represents the number of instances which are actually in class X and predicted as class Y, where X;Y $\in$ positive; neutral; negative. The recall (R) of a class X is the ratio of the number of elements correctly classified as X to the number of known elements in class X. Precision (P) of a class X is the ratio of the number of elements classified correctly as X to the total predicted as the class X. For example, the precision of the negative class is computed as: $P(neg) = i/(c + f + i)$; its recall, as: $R(neg) = i/(g+h+i)$; and the $F1$ measure is the harmonic mean between both precision and recall. In this case, $F1(neg) = \frac{2P(neg) \cdot R(neg)}{P(neg)+R(neg)}$.

We also compute the overall accuracy as: $A = \frac{a+e+i}{a+b+c+d+e+f+g+h+i}$. It considers equally important the correct classification of each sentence, independently of the class, and basically measures the capability of the method to predict the correct output. A variation of F1, namely, Macro-F1, is normally reported to evaluate classification effectiveness on skewed datasets. Macro-F1 values are computed by first calculating F1 values for each class in isolation, as exemplified above for negative, and then averaging over all classes. Macro-F1 considers equally important the effectiveness in *each class*, independently of the relative size of the class. Thus, accuracy and Macro-F1 provide complementary assessments of the classification effectiveness. Macro-F1 is especially important when the class distribution is very skewed, to verify the capability of the method to perform well in the smaller classes.

The described metrics can easily computed for the 2-class experiments by just removing neutral columns and rows as in the table below.

		Predicted	
		Positive	Negative
	Positive	a	b
Actual	Negative	c	d

In this case, the precision of positive class is computed as: $P(pos) = a/(a+c)$; its recall as: $R(pos) = a/(a+b)$; while its F1 is $F1(pos) = \frac{2P(pos) \cdot R(pos)}{P(pos)+R(pos)}$

As we have a large number of combinations among the base methods, metrics and datasets, a global analysis of the performance of all these combinations is not an easy task. We propose a simple but informative measure to assess the overall performance ranking. The Mean Ranking is basically the sum of ranks obtained by a method in each dataset divided by the total number of datasets, as below:

$$MR = \frac{\sum_{j=1}^{nd} ri}{nd}$$

where nd is the number of datasets and ri is the rank of the method for dataset i. It is important to notice that the rank was calculated based on Macro F1.

The last evaluation metric we exploit is the Friedman's Test [38]. It allows to verify whether, in a specific experiment, the observed values are globally similar. We used this test to tell if the methods present similar performance across different datasets. To exemplify the application of this test, suppose that n restaurants are each rated by k judges. The question that arises is: are the judges ratings consistent with each other or do they follow completely different patterns? The application in our context is very similar: the datasets are the judges and the Macro-F1 achieved by a method is the rating from the judges.

The Friedman's Test is applied to rankings. Then, to proceed with this statistical test, we sort the methods in decreasing order of Macro-F1 for each dataset. More formally, the Friedman's rank test in our experiment is defined as:

$$F_R = (\frac{12}{rc(c+1)} \sum_{j=1}^{c} R_j^2) - 3r(c+1)$$

where

R_j^2 = square of the sum of rank positions of method j (j = 1,2,..., c)
r = number of datasets
c = number of methods

As the number of datasets increase, the statistical test can be approximated by using the chi-square distribution with $c-1$ degrees of freedom. Then, if the F_R computed value is greater than the critical value for the chi-square distribution the null hypothesis is rejected. This null hypothesis states that ranks obtained per dataset are globally similar. Accordingly, rejecting the null hypothesis means that there are significant differences in the ranks across datasets. It is important to note that, in general, the critical value is obtained with significance level $\alpha = 0.05$. Synthesizing, the null hypothesis should be rejected if $F_R > X_\alpha^2$, where X_α^2 is the critical value verified in the chi-square distribution table with $c-1$ degrees of freedom and α equals 0.05.

Comparing Prediction Performance

We start the analysis of our experiments by comparing the results of all previously discussed metrics for all datasets. Table 4 and Table 5 present accuracy, precision, and Macro-F1 for all methods considering four datasets for the 2-class and 3-class experiments, respectively. For simplicity, we choose to discuss results only for these datasets as they come from different sources and help us to illustrate the main findings from our analysis. Results for all the other datasets are presented in the Additional File 1. There are many interesting observations we can make from these results, summarized next.

Methods prediction performance varies considerably from one dataset to another: First, we note the same social media text can be interpreted very differently depending on the choice of a sentiment method. Overall, we note that the all methods yielded wide variation in their results across the different datasets. By analyzing Table 4 we can note that Vader works well for Tweets_STF, appearing in the second place, but it presents poor performance in Comments_BBC, Tweets_SAN, and Comments_DIGG, achieving the eleventh place for the first dataset and the ninth for other both. Although the first two datasets contain tweets, they belong to different contexts, which affects the performance of some methods like VADER. Another important aspect to be analyzed in this Table is the coverage. Although SentiStrength has presented good Macro-F1 values, its coverage is usually low as this method tends to classify a high number of instances as neutral. Note that some datasets provided by the SentiStrength's authors, as shown in Table 3, specially the Twitter datasets, have more neutral sentences than positive and negative ones. Another expected result is the good Macro-F1 values obtained by Emoticons, specially in the Twitter datasets. It is important to highlight that, in spite of achieving high accuracy and Macro-F1, the coverage of many methods, such as PANAS, VADER, and SentiStrength, is low (e.g. below 30%) as they only infer the polarity of part of the input sentences. Thus, the choice of a sentiment analysis is highly dependent on the data and application and researchers and practitioners need to account with this tradeoff between prediction performance and coverage.

The same kind of high variation on the methods's prediction performance can be noted for the 3-class experiments, as presented in Table 5. Umigon, the best method in four Twitter datasets, felt to the fifteenth place in the Comments_NYT dataset. We can also note the lower Macro-F1 values for some methods like Emoticons are due to the high number of sentences without emoticons in the datasets. Methods like Emoticons DS, PANAS, and Happiness Index, tend do classify only a small part of instances as neutral and also presented a poor performance in the 3-class experiments. Methods like SenticNet and LIWC were not developed for detecting neutral sentences and also achieved low values of Macro-F1. They do not among the best methods in the 2-class experiments, which is the task they were originally designed to.

Finally, Table 8 presents the Friedman's test results showing that there are significant differences in the mean rankings observed for the methods across all datasets. It statistically indicates that in terms of accuracy and Macro-F1 there is no single method that always achieves a consistent rank position for different datasets, which is something similar to the well-known "no-free lunch theorem" [16]. So, overall, before using a sentiment analysis method in a novel dataset, it is crucial to test different methods in a sample of data before simply choose one that is acceptable by the research community.

This last results suggests that, even with the good insights provided by this work about which methods perform better in each context, a preliminary investigation needs to be performed when sentiment analysis is used in a new dataset in order to guarantee a reasonable prediction performance. In the case in which prior tests are not feasible, this benchmark presents valuable information for researchers and companies that are planning to develop research and solutions on sentiment analysis.

Existing methods let a lot of space for improvements: We can note that the performance of the evaluated methods are ok, but there is a lot of space for improvements. For example, if we look at the Macro-F1 values only for the best method on each each dataset (See Table XX and XX), we can note that the overall prediction performance of the methods is still low – i.e. Macro-F1 values are around 0.9 only for methods with low coverage in the 2-class experiments and only 0.6 for the 3-class experiment. Considering that we looking at the performance of the best method out of 22 unsupervised tools, these numbers suggest that current sentence-level sentiment analysis methods still let space for improvements. Additionally, we also noted that the best method for each dataset varies considerable from one dataset to another. This might indicate that each method complements the others in different ways.

Most methods are better to classify positive than negative or neutral sentences: Figure 2 presents the average F1 Score for the 3-class experiments. It is easier to notice that eleven out of twenty-two methods are more accurate while classifying positive than negative or neutral messages, suggesting that some methods may be more biased towards positivity. Neutral messages showed to be even harder to detect by most methods. Recent efforts show that human language have a universal positivity bias [39]. Naturally, part of the bias is observed in sentiment prediction, an intrinsic property of some methods due to the way they are designed. For instance, [32] developed a lexicon in which positive and negative values are associated to words, hashtags, and any sort of tokens according to the frequency with which these tokens appear in tweets containing positive and negative emoticons. This method showed to be biased towards positivity due to the larger amount of positivity in the data they used to build the lexicon. The overall poor performance of this specific method is credited to its lack of treatment of neutral messages and the focus on Twitter messages.

Some methods are consistently among the best methods:

Table 9 presents the mean rank value, detailed before, for 2-class and 3-class experiments. The elements are sorted by the overall mean rank each method achieved based on Macro-F1 for all datasets. The top seven methods based on Macro-F1 for 2-class experiments are SentiStrength, Semantria, OpinionLexicon, AFINN, SO-CAL, Vader and Umigon. Except from Sentistrength, that was replaced by Pattern.en, the other six methods produce the best results across several datasets for both, 2 and 3-class tasks. These methods would be preferable in situations in which any sort of preliminary evaluation is not possible to be done. The mean rank for 2-class experiments is accompanied by the coverage metric, which is very important to avoid misinterpretation of the results. Observe that SentiStrength exhibited the best mean rank for these experiments, however it presents very low coverage, around 30%, a very poor result compared with Semantria and OpinionLexicon that achieved a worse mean rank (3.61 and 6.28, respectively) but an expressive better coverage, above 60%. Note also that SentiStrength presents poor results in the 3-class experiments which can be explained by its inclination to neutral as mentioned before. SentiStrength results show that many negative and positive sentences are classified as neutral.

Methods are often better in the datasets they were originally evaluated: We also note those methods perform better in datasets in which they were originally validated, which is somewhat expected due to fine tuning procedures. We could do this comparison only for SentiStrength and VADER, which kindly allowed the entire reproducibility of their work, sharing both methods and datasets. To understand this difference, we calculated the mean rank for these methods without their 'original' datasets and put the results in parenthesis. Note that, in some cases the rank order changes towards a lower value but it does not imply in major changes. We also note those methods often perform better in datasets in which they were originally validated, which is somewhat expected due to fine tuning procedures. We could do this comparison only for SentiStrength and VADER, which kindly allowed the entire reproducibility of their work, sharing both methods and datasets. To understand this difference, we calculated the mean rank for these methods without their 'original'

datasets and put the results in parenthesis. Note that, in some cases the rank order slightly changes but it does not imply in major changes. Overall, these observations suggest that initiatives like SemEval are key for the development of the area, as they allow methods to compete in a contest for a specific dataset. More important, it highlight that a standard sentiment analysis benchmark is needed and it needs to be constantly updated. We also emphasize that is possible that other methods, such as paid softwares, make use of some of the datasets used in this benchmark to improve their performance as most of gold standard used in this work is available in the Web or under request to authors.

Some methods showed to be better for specific contexts: In order to better understand the prediction performance of methods in types of data, we divided all datasets in three specific contexts – Social Networks, Comments, and Reviews – and calculated mean rank of the methods for each of them. Table 10 presents the contexts and the respective datasets.

Tables 11, 13 and 12 present the mean rank for each context separately. In the context of Social Networks the best method for 3-class experiments was Umigon, followed by Vader and AFINN. In the case of 2-class the winner was SentiStrength with a coverage around 30% and the second and fourth place were Emoticons and Panas-t with about 18% and 6% of coverage, respectively. This highlights the importance to analyze the 2-class results together with the coverage. Overall, when there is an emoticon on the text or a word from the psychometric scale PANAS, these methods are able to tell the polarity of the sentences, but they are not able to identify the polarity of the input text for the large majority of the input text. Recent efforts suggest these properties are useful for combination of methods [19]. Semantria, OpinionLexicon and Umigon showed to be the best alternatives for detecting only positive and negative polarities in social network data due to the high coverage and prediction performance. It is important to highlight that LIWC 2007 appears on the 15th and 18th position for the 3-class and 2-class mean rank results for the social network datasets and it is a very popular method in this community.

Similar analyses can be performed for the contexts Comments and Reviews. Sentistregth, Vader, Semantria, AFINN, and Opinion Lexicon showed to be the best alternatives for 2-class and 3-class experiments on datasets of comments whereas SO-CAL, Semantria, and SenticNet showed to be the best for the 2-class experiments for the datasets containing short reviews. Note that for the last one, the 3-class experiments have no results since datasets containing reviews have no neutral sentences nor a representative number of sentences without subjectivity.

We also calculated the Friedman's value for each of these specific contexts. Even after grouping the datasets, we still observe that there are significant differences in the observed ranks across the datasets. Although the values obtained for each context were quite smaller than Friedman' global value, they are still above the critical value. Table 14 presents the results of Friedman's test for the individual contexts in both experiments, 2 and 3-class. Recall that for the 3-class experiments, datasets with no neutral sentences or with an unrepresentative number of neutral sentences were not considered. For this reason, Friedman's results for 3-class experiments in the Reviews context presents no values.

Overall, even within the same contexts, methods' results varies largely from one dataset to another. This suggests that

Concluding Remarks

Recent efforts to analyze the moods embedded in Web 2.0 content have adopted various sentiment analysis methods, which were originally developed in linguistics and psychology. Several of these methods became widely used in their knowledge fields and have now been applied as tools to quantify moods in the context of unstructured short messages in online social networks. In this article, we present a thorough comparison of twenty-two popular sentence-level sentiment analysis methods using gold standard datasets that span different types of data sources. Our effort quantifies the prediction performance of the twenty-two popular sentiment analysis methods across eighteen datasets for two tasks: differentiating two classes (positive and negative) and three classes (positive, negative, and neutral).

Among many findings, we highlight that although our results identified a few methods that able to appear among the best methods for different datasets, we noted that the overall prediction performance still left a lot of space for improvements. More important, we show that the prediction performance of methods vary largely across datasets. For example, LIWC 2007, is among the most popular sentiment methods in the social network context and obtained a bad rank position in comparison with other datasets. This suggests that sentiment analysis methods cannot be used as "off-the-shelf" methods, specially for novel datasets. We show that the same social media text can be interpreted very differently depending on the choice of a sentiment method, suggesting that it is important that researchers and companies perform experiments with different methods before applying a method.

As a final contribution we open the datasets and codes used in this paper for the research community. We also incorporated them in a Web service from our research team called iFeel [20][5] that allow users compare the results of various sentiment analysis methods easily. We hope our effort cannot only help researchers and practitioners to compare a wide range of sentiment analysis techniques but also that it can help foster new relevant research in this area with a rigorous scientific approach.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Text for this section ...

Acknowledgements
This work is supported by grants from CAPES, Fapemig, and CNPq. ...

Author details
[1]Computer Science Department, Federal University of Minas Gerais, Belo Horizonte, Brazil. [2]Computer and Systems Department, Federal University of Ouro Preto, Joao Monlevade, Brazil.

References
1. Feldman, R.: Techniques and applications for sentiment analysis. Commun. ACM **56**(4), 82–89 (2013). doi:10.1145/2436256.2436274
2. Hu, M., Liu, B.: Mining and summarizing customer reviews. KDD '04, pp. 168–177 (2004). http://doi.acm.org/10.1145/1014052.1014073
3. Oliveira, N., Cortez, P., Areal, N.: On the predictability of stock market behavior using stocktwits sentiment and posting volume. In: Correia, L., Reis, L.P., Cascalho, J. (eds.) EPIA. Lecture Notes in Computer Science, vol. 8154, pp. 355–365. Springer, ??? (2013)
4. Bollen, J., Mao, H., Zeng, X.-J.: Twitter Mood Predicts the Stock Market. CoRR **abs/1010.3003** (2010)
5. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In: International AAAI Conference on Weblogs and Social Media (ICWSM) (2010)
6. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: ACL Conference on Empirical Methods in Natural Language Processing, pp. 79–86 (2002)
7. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: Liwc and computerized text analysis methods. Journal of Language and Social Psychology **29**(1), 24–54 (2010)
8. Gonçalves, P., Benevenuto, F., Cha, M.: PANAS-t: A Pychometric Scale for Measuring Sentiments on Twitter **abs/1308.1857v1** (2013)

[5]http://www.ifeel.dcc.ufmg.br

9. Bollen, J., Pepe, A., Mao, H.: Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. CoRR **abs/0911.1583** (2009)
10. Kramer, A.D.I., Guillory, J.E., Hancock, J.T.: Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences of the United States of America **111**(24), 8788–90 (2014). doi:10.1073/pnas.1320040111
11. Thelwall, M.: Heart and soul: Sentiment strength detection in the social web with SentiStrength. `http://sentistrength.wlv.ac.uk/documentation/SentiStrengthChapter.pdf` (2013)
12. Reis, J., Goncalves, P., Vaz de Melo, P., Prates, R., Benevenuto, F.: Magnet news: You choose the polarity of what you read. In: International AAAI Conference on Web-Blogs and Social Media (2014)
13. Reis, J., Benevenuto, F., Vaz de Melo, P., Prates, R., Kwak, H., An, J.: Breaking the news: First impressions matter on online news. In: Proceedings of the 9th International AAAI Conference on Web-Blogs and Social Media (ICWSM) (2015)
14. Tamersoy, A., De Choudhury, M., Chau, D.H.: Characterizing smoking and drinking abstinence from social media. In: Proceedings of the 26th ACM Conference on Hypertext and Social Media (HT) (2015)
15. Hutto, C., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International AAAI Conference on Weblogs and Social Media (ICWSM) (2014)
16. Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. IEEE Transactions on Evlutionary Computation **1**(1), 67–82 (1997)
17. Tsytsarau, M., Palpanas, T.: Survey on mining subjective data on the web. Data Min. Knowl. Discov. **24**(3), 478–514 (2012). doi:10.1007/s10618-011-0238-6
18. Levallois, C.: Umigon: sentiment analysis for tweets based on terms lists and heuristics. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 414–417. Association for Computational Linguistics, Atlanta, Georgia, USA (2013). http://www.aclweb.org/anthology/S13-2068
19. Gonçalves, P., Araujo, M., Benevenuto, F., Cha, M.: Comparing and combining sentiment analysis methods. In: Proceedings of the 1st ACM Conference on Online Social Networks (COSN'13) (2013)
20. Araujo, M., Gonçalves, P., Benevenuto, F., Cha, M.: ifeel: A system that compares and combines sentiment analysis methods. In: WWW (Companion Volume). International World Wide Web Conference (WWW'14), ??? (2014)
21. Liu, B.: Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies **5**(1), 1–167 (2012). doi:10.2200/s00416ed1v01y201204hlt016
22. Godbole, N., Srinivasaiah, M., Skiena, S.: Large-scale sentiment analysis for news and blogs. In: Proceedings of the International Conference on Weblogs and Social Media (ICWSM) (2007)
23. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: The good the bad and the omg! In: Int'l AAAI Conference on Weblogs and Social Media (ICWSM) (2011)
24. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)
25. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B.: Learning sentiment-specific word embedding for twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers, pp. 1555–1565 (2014)
26. Mohammad, S.M., Kiritchenko, S., Zhu, X.: Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In: Proceedings of the Seventh International Workshop on Semantic Evaluation Exercises (SemEval-2013), Atlanta, Georgia, USA (2013)
27. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (2014)
28. Johnson, R., Zhang, T.: Effective use of word order for text categorization with convolutional neural networks. In: NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015, pp. 103–112 (2015)
29. Valitutti, R.: Wordnet-affect: an affective extension of wordnet. In: In Proceedings of the 4th International Conference on Language Resources and Evaluation, pp. 1083–1086 (2004)
30. Bradley, M.M., Lang, P.J.: 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)
31. Nielsen, F.Å.: A new anew: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903 (2011)
32. Hannak, A., Anderson, E., Barrett, L.F., Lehmann, S., Mislove, A., Riedewald, M.: Tweetin' in the rain: Exploring societal-scale effects of weather on mood. In: Int'l AAAI Conference on Weblogs and Social Media (ICWSM) (2012)
33. Esuli, Sebastiani: Sentiwordnet: A publicly available lexical resource for opinion mining. In: International Conference on Language Resources and Evaluation (LREC), pp. 417–422 (2006)
34. Cambria, E., Speer, R., Havasi, C., Hussain, A.: Senticnet: A publicly available semantic resource for opinion mining. In: AAAI Fall Symposium Series (2010)
35. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1631–1642 (2013)
36. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '08 (2008)
37. Diakopoulos, N.A., Shamma, D.A.: Characterizing debate performance via aggregated twitter sentiment. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, pp. 1195–1198 (2010). ACM

38. Berenson, M.L., Levine, D.M., Szabat, K.A.: Basic Business Statistics - Concepts and Applications, 13 edn., p. 840. Pearson, ??? (2014)
39. Dodds, P.S., Clark, E.M., Desu, S., Frank, M.R., Reagan, A.J., Williams, J.R., Mitchell, L., Harris, K.D., Kloumann, I.M., Bagrow, J.P., Megerdoomian, K., McMahon, M.T., Tivnan, B.F., Danforth, C.M.: Human language reveals a universal positivity bias. Proceedings of the National Academy of Sciences **112**(8), 2389–2394 (2015). doi:10.1073/pnas.1411678112. http://www.pnas.org/content/112/8/2389.full.pdf
40. Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., Patwardhan, S.: Opinionfinder: a system for subjectivity analysis. In: HLT/EMNLP on Interactive Demonstrations, pp. 34–35 (2005)
41. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: ACL Conference on Empirical Methods in Natural Language Processing, pp. 347–354 (2005)
42. Dodds, P.S., Danforth, C.M.: Measuring the happiness of large-scale written expression: songs, blogs, and presidents. Journal of Happiness Studies **11**(4), 441–456 (2009). doi:10.1007/s10902-009-9150-9
43. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) LREC. European Language Resources Association, ??? (2010)
44. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
45. Mohammad, S.: #emotional tweets. In: *SEM 2012: 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 2012), pp. 246–255. Association for Computational Linguistics, Montréal, Canada (2012). http://www.aclweb.org/anthology/S12-1033
46. De Smedt, T., Daelemans, W.: Pattern for python. The Journal of Machine Learning Research **13**(1), 2063–2067 (2012)
47. Wang, H., Can, D., Kazemzadeh, A., Bar, F., Narayanan, S.: A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In: ACL System Demonstrations, pp. 115–120 (2012)
48. Watson, D., Clark, L.: Development and validation of brief measures of positive and negative affect: the panas scales. Journal of Personality and Social Psychology **54**(1), 1063–1070 (1985)
49. Mohammad, S., Turney, P.D.: Crowdsourcing a word-emotion association lexicon. Computational Intelligence **29**(3), 436–465 (2013)
50. Plutchik, R.: A general psychoevolutionary theory of emotion, pp. 3–33. Academic press, New York (1980)
51. Pappas, N., Popescu-Belis, A.: Sentiment analysis of user comments for one-class collaborative filtering over ted talks. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 773–776 (2013). ACM
52. Lexalytics: Sentiment extraction - measuring the emotional tone of content. Technical report, Lexalytics (October 2015)
53. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language Resources and Evaluation **1**(2), 0 (2005)
54. Stone, P.J., Dunphy, D.C., Smith, M.S., Ogilvie, D.M.: The General Inquirer: A Computer Approach to Content Analysis. MIT Press, ??? (1966). http://www.webuse.umd.edu:9090/
55. Biever, C.: Twitter mood maps reveal emotional states of america. The New Scientist **207** (2010)
56. Taboada, M., Anthony, C., Voll, K.: Methods for creating semantic orientation dictionaries. In: Conference on Language Resources and Evaluation (LREC), pp. 427–432 (2006)
57. Mohammad, S., Dunne, C., Dorr, B.: 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, pp. 599–608. Association for Computational Linguistics, Stroudsburg, PA, USA (2009). http://dl.acm.org/citation.cfm?id=1699571.1699591
58. Taboada, M., Anthony, C., Voll, K.: Methods for creating semantic orientation dictionaries. In: Conference on Language Resources and Evaluation (LREC), pp. 427–432 (2006)
59. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring user influence in twitter: The million follower fallacy. In: International AAAI Conference on Weblogs and Social Media (ICWSM) (2010)
60. Strapparava, C., Mihalcea, R.: Semeval-2007 task 14: Affective text. In: Proceedings of the 4th International Workshop on Semantic Evaluations. SemEval '07, pp. 70–74. Association for Computational Linguistics, Stroudsburg, PA, USA (2007). http://dl.acm.org/citation.cfm?id=1621474.1621487
61. Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., Wilson, T.: SemEval-2013 Task 2: Sentiment Analysis in Twitter (2013)
62. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: In Proceedings of the ACL, pp. 271–278 (2004)
63. Narr, S., Hülfenhaus, M., Albayrak, S.: Language-independent twitter sentiment analysis. Knowledge Discovery and Machine Learning (KDML), 12–14 (2012)
64. Aisopos, F.: Manually Annotated Sentiment Analysis Twitter Dataset NTUA. `www.grid.ece.ntua.gr` (2014)
65. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Processing **-**, 1–6 (2009)
66. Sanders, N.: Twitter Sentiment Corpus by Niek Sanders. `http://www.sananalytics.com/lab/twitter-sentiment/` (2011)

Figures

Tables

Additional Files

Additional file 1 — Full results of prediction performance

In the Additional file 1, we present the full results of prediction performance of all twenty-two sentiment analysis methods on all labeled datasets googletrends

Figure 1 Searches on Google for the Query: "Sentiment Analysis". This figure shows the steady grouth on the number of searches on the topic, according to Google Trends, mainly after the popularization of online social networks (OSNs)

F1byclass

Figure 2 Average F1 Score for each class. This figure presents the average F1 of positive and negative class and as we can see, methods use to achieve better prediction performance on positive messages

Table 1 Overview of the sentence-level methods available in the literature.

Name	Description	L	ML
Emoticons [19]	Messages containing positive/negative emoticons are positive/negative. Messages without emoticons are not classified.	✓	
Opinion Lexicon [2]	Focus on Product Reviews. Builds a Lexicon to predict polarity of product features phrases that are summarized to provide an overall score to that product feature.	✓	
Opinion Finder (MPQA) [40] [41]	Performs subjectivity analysis trough a framework with lexical analysis former and a machine learning approach latter.	✓	✓
Happiness Index [42]	Quantifies happiness levels for large-scale texts as lyrics and blogs. It uses ANEW words [30] to rank the documents.	✓	
SentiWordNet [33] [43]	Construction of a lexical resource for Opinion Mining based on WordNet [44]. The authors grouped adjectives, nouns, etc in synonym sets (synsets) and associated three polarity scores (positive, negative and neutral) for each one.	✓	✓
LIWC [7]	Text analysis paid tool to evaluate emotional, cognitive, and structural components of a given text. It uses a dictionary with words classified into categories (anxiety, health, leisure, etc).	✓	
SenticNet [34]	Uses dimensionality reduction to infer the polarity of common sense concepts and hence provide a resource for mining opinions from text at a semantic, rather than just syntactic level.	✓	
AFINN [31] - A new ANEW	Builds a Twitter based sentiment Lexicon including Internet slangs and obscene words.	✓	
SO-CAL [24]	Creates a new Lexicon with unigrams (verbs, adverbs, nouns and adjectives) and multi-grams (phrasal verbs and intensifiers) hand ranked with scale +5 (strongly positive) to -5 (strongly negative). Authors also included part of speech processing, negation and intensifiers.	✓	
Emoticons DS (Distant Supervision)[32]	Creates a scored lexicon based on a large dataset of tweets. Its based on the frequency each lexicon occurs with positive or negative emotions.	✓	
NRC Hashtag [45]	Builds a lexicon dictionary using a Distant Supervised Approach. In a nutshell it uses known hashtags (i.e #joy, #sadness, #happy etc) to “classify” the tweet. Afterwards, it verifies frequency each specific n-gram occurs in a emotion and calculates its Strong of Associaton with that emotion.	✓	
Pattern.en [46]	Python Programming Package (toolkit) to deal with NLP, Web Mining and Sentiment Analysis. Sentiment analysis is provided through averaging scores from adjectives in the sentence according to a bundle lexicon of adjective.	✓	
SASA [47]	Detects public sentiments on Twitter during the 2012 U.S. presidential election. It is based on the statistical model obtained from the classifier Naïve Bayes on unigram features. It also explores emoticons and exclamations.		✓
PANAS-t [8]	Detects mood fluctuations of users on Twitter. The method consists of an adapted version (PANAS) Positive Affect Negative Affect Scale [48], well-known method in psychology with a large set of words, each of them associated with one from eleven moods(surprise, fear, joviality, assurance, serenity, sadness, guilt, hostility, shyness, fatigue, and attentiveness).	✓	
EmoLex [49]	Builds a general sentiment Lexicon crowdsourcing supported. Each entry lists the association of a token with 8 basic sentiments: joy, sadness, anger, etc defined by [50]. Proposed Lexicon includes unigrams and bigrams from Macquarie Thesaurus and also words from GI and Wordnet.	✓	
SANN [51]	Infer additional reviews user ratings by performing sentiment analysis (SA) of user comments and integrating its output in a nearest neighbor (NN) model that provides multimedia recommendations over TED Talks.	✓	✓
Sentiment140 Lexicon [26]	Creation of a lexicon dictionary in a similar way to [45] but authors used the occurrency of emoticons to classify the tweet as positive or negative. Then, the n-gram score was calculated based on the frequency it occurs in each class of tweets.	✓	
SentiStrength [11]	Builds a lexicon dictionary annotated by humans and improved with the use of Machine Learning.	✓	✓
Stanford Recursive Deep Model [35]	Proposes a model called Recursive Neural Tensor Network (RNTN) that processes all sentences dealing with their structures and compute the interactions between them. This approach is interesting since RNTN take into account the order of words in a sentence, which is ignored in most of methods.	✓	✓
Umigon [18]	Disambiguates tweets using lexicon with heuristics to detect negations plus elongated words and hashtags evaluation.	✓	
Vader [15]	It is a human-validated sentiment analysis method developed for twitter and social media contexts. Vader was created from a generalizable, valence-based, human-curated gold standard sentiment lexicon.	✓	
Semantria [52]	It is a paid tool that employs multi-level analysis of sentences. Basically it has four levels: part of speech, assignment of previous scores from dictionaries, application of intensifiers and finally machine learning techniques to delivery a final weight to the sentence.	✓	✓

Table 2 Overview of the sentence-level methods available in the literature.

Name	Output	Validation	Compared To	Lexicon size
Emoticons	-1, 1	-	-	79
Opinion Lexicon	Provides polarities for lexicons.	Product Reviews from Amazon and CNet	-	6,787
Opinion Finder (MPQA)	**Negative, Objective, Positive**	MPQA [53]	Compared to itself in different versions.	20,611
Happiness Index	**1, 2, 3, 4, 5, 6, 7, 8, 9**	Lyrics, Blogs,STU Messages [6], British National Corpus [7],	-	1,034
SentiWordNet	Provides positive, negative and objective scores for each word (0.0 to 1.0)	-	General Inquirer (GI)[54]	117,658
LIWC	**negEmo, posEmo**	-	-	4,500
SenticNet	**Negative, Positive**	Patient Opinions (Unavailable)	SentiStrength [11]	15,000
AFINN	Provides polarity score for lexicons (-5 to 5).	Twiter [55]	OpinonFinder [40], ANEW [30], GI [54] and Sentistrength [11]	2,477
SO-CAL	**[$<$0), 0, ($>$0]**	Epinion [56], MPQA[53], Myspace[11],	MPQA[53], GI[54], SentiWordNet [33],"Maryland" Dict. [57], Google Generated Dict. [58]	9,928
Emoticons DS (Distant Supervision)	Provides polarity score for lexicons	Validation with unlabeled twitter data [59]	-	1,162,894
NRC Hashtag	Provides polarities for lexicons	Twitter (SemEval-2007 Affective Text Corpus) [60]	WordNet Affect [60]	679,468
Pattern.en	**Objective, $<$0.1, $\geq$0.1]**	Product reviews, but the source was not specified	-	2,973
SASA [47]	**Negative, Neutral, Unsure, Positive**	"Political" Tweets labeled by "turkers" (AMT) (unavailable)	-	-
PANAS-t	Provides association for each word with eleven moods (joviality, attentiveness, fear, etc)	Validation with unlabeled twitter data [59]	-	50
EmoLex	Provides polarities for lexicons	-	Compared with existing gold standard data but it was not specified	141,820
SANN	**neg, neu, pos**	Their own dataset - Ted Talks	Comparison with other multimedia recommendation approaches.	MPQA (8,226) / Their own (9,176)
Sentiment140	Provides polarity scores for lexicon	Twitter and SMS from Semeval 2013, task 2 [61].	Other Semeval 2013, task 2 approaches	1,220,176
SentiStrength	**-1,0,1**	Their own datasets - Twitter, Youtube, Digg, Myspace, BBC Forums and Runners World.	The best of nine Machine Learning techniques for each test.	2,698
Stanford Recursive Deep Model	**very negative, negative, neutral, positive, very positive**	Movie Reviews [62]	Naïve Bayes and SVM with bag of words features and bag of bigram features.	227,009
Umigon	**Negative, Neutral, Positive**	Twitter and SMS from Semeval 2013, task 2 [61].	[26]	1,053
Vader	**[$<$-0,05), (-0,05..0,05), ($>$0,05]**	Their own datasets - Twitter, Movie Reviews, Technical Product Reviews, NYT User's Opinions.	(GI)[54], LIWC, [7], SentiWordNet [33], ANEW [30], SenticNet [34] and some Machine Learning Approaches.	7,517
Semantria	**negative, neutral, positive**	not available	not available	not available

Table 3 Labeled datasets.

Dataset	Nomeclature	# Msgs	# Pos	# Neg	# Neu	Average # of phrases	Average # of words	Annotators Expertise	# of Annotators	R (%)
Comments (BBC) [11]	**Comments_BBC**	1,000	99	653	248	3,98	64,39	Non Expert	3	87,0
Comments (Digg) [11]	**Comments_Digg**	1,077	210	572	295	2,50	33,97	Non Expert	3	88,0
Comments (NYT) [15]	**Comments_NYT**	5,190	2,204	2,742	244	1,01	17,76	AMT	20	88,0
Comments (TED) [51]	**Comments_TED**	839	318	409	112	1	16,95	Non Expert	6	82,0
Comments (Youtube) [11]	**Comments_YTB**	3,407	1,665	767	975	1,78	17,68	Non Expert	3	90,0
Movie-reviews [62]	**Reviews_I**	10,662	5,331	5,331	-	1,15	18,99	User Rating	-	66,0
Movie-reviews [15]	**Reviews_II**	10,605	5,242	5,326	37	1,12	19,33	AMT	20	97,0
Myspace posts [11]	**Myspace**	1,041	702	132	207	2,22	21,12	Non Expert	3	91,0
Product reviews [15]	**Amazon**	3,708	2,128	1,482	98	1,03	16,59	AMT	20	94,0
Tweets (Debate) [37]	**Tweets_DBT**	3,238	730	1249	1259	1,86	14,86	AMT + Expert	Undef.	60
Tweets (Random) [11]	**Tweets_RND_I**	4,242	1,340	949	1953	1,77	15,81	Non Expert	3	88,0
Tweets (Random) [15]	**Tweets_RND_II**	4,200	2,897	1,299	4	1,87	14,10	AMT	20	97,5
Tweets (Random) [63]	**Tweets_RND_III**	3,771	739	488	2,536	1,54	14,32	AMT	3	90,0
Tweets (Random) [64]	**Tweets_RND_IV**	500	139	119	222	1,90	15,44	Expert	Undef.	90,0
Tweets (Specific domains w/ emot.) [65]	**Tweets_STF**	359	182	177	-	1,0	15,1	Non Expert	Undef.	97,0
Tweets (Specific topics) [66]	**Tweets_SAN**	3737	580	654	2503	1,60	15,03	Expert	1	97,0
Tweets (Semeval2013 Task2) [61]	**Tweets_Semeval**	6,087	2,223	837	3027	1,86	20,05	AMT	5	100,0
Runners World forum [11]	**RW**	1,046	484	221	341	4,79	66,12	Non Expert	3	86,0

Table 4 2-classes experiments results with 4 datasets

Dataset	Method	Accur.	Posit. Sentiment P	Posit. Sentiment R	Posit. Sentiment F1	Negat. Sentiment P	Negat. Sentiment R	Negat. Sentiment F1	MacroF1	Coverage
Tweets_SAN	AFINN	77.85	72.73	86.72	79.11	84.83	69.54	76.43	77.77	69.94
	Emolex	72.06	65.42	85.67	74.19	82.61	60.06	69.55	71.87	60.04
	Emoticons	85.71	87.50	89.74	88.61	82.61	79.17	80.85	84.73	5.77
	Emoticons DS	47.20	47.13	99.21	63.90	55.56	0.88	1.74	32.82	98.08
	Happiness Index	64.23	59.63	89.92	71.70	78.81	38.11	51.38	61.54	45.10
	NRC Hashtag	69.77	72.01	58.71	64.69	68.36	79.63	73.57	69.13	93.68
	LIWC	77.27	71.37	92.05	80.40	88.38	62.10	72.95	76.67	63.70
	Opinion Finder	68.60	63.54	71.35	67.22	73.80	66.35	69.87	68.55	34.74
	Opinion Lexicon	81.77	78.18	86.74	82.24	85.98	77.05	81.27	81.75	65.35
	PANAS-t	84.62	80.00	80.00	80.00	87.50	87.50	87.50	83.75	2.38
	Pattern.en	74.62	68.46	90.59	77.98	86.40	58.90	70.04	74.01	72.59
	SANN	70.02	64.40	85.95	73.63	80.46	54.90	65.27	69.45	45.55
	SASA	61.18	61.74	74.71	67.61	60.12	45.16	51.58	59.59	43.45
	Semantria	78.91	76.42	82.85	79.50	81.79	75.08	78.29	78.90	57.38
	SenticNet	66.51	65.31	57.73	61.29	67.33	73.96	70.49	65.89	77.45
	Sentiment140	72.45	0.00	0.00	0.00	72.45	100.00	84.02	42.01	53.90
	SentiStrength	89.47	88.70	91.81	90.23	90.41	86.84	88.59	89.41	29.61
	SentiWordNet	67.49	64.25	75.63	69.48	71.90	59.70	65.23	67.35	59.21
	SO-CAL	79.70	74.74	84.55	79.34	85.11	75.56	80.05	79.70	68.19
	Stanford DM	62.72	87.60	22.60	35.93	59.35	97.25	73.71	54.82	92.94
	Umigon	82.41	83.66	80.49	82.04	81.25	84.32	82.76	82.40	67.74
	Vader	77.18	71.81	88.15	79.15	85.34	66.59	74.81	76.98	78.74
Tweets_STF	AFINN	84.42	80.62	91.49	85.71	89.66	77.04	82.87	84.29	76.88
	Emolex	79.65	76.09	88.98	82.03	85.23	69.44	76.53	79.28	62.95
	Emoticons	85.42	80.65	96.15	87.72	94.12	72.73	82.05	84.89	13.37
	Emoticons DS	51.96	51.41	100.00	67.91	100.00	2.27	4.44	36.18	99.72
	Happiness Index	65.93	58.72	94.39	72.40	88.89	40.34	55.49	63.95	62.95
	NRC Hashtag	71.30	73.05	70.93	71.98	69.51	71.70	70.59	71.28	92.20
	LIWC	64.29	63.75	76.12	69.39	65.22	50.85	57.14	63.27	70.39
	Opinion Finder	80.77	81.16	76.71	78.87	80.46	84.34	82.35	80.61	43.45
	Opinion Lexicon	86.10	83.67	91.11	87.23	89.29	80.65	84.75	85.99	72.14
	PANAS-t	94.12	88.89	100.00	94.12	100.00	88.89	94.12	94.12	4.74
	Pattern.en	77.85	75.69	85.09	80.12	80.95	69.86	75.00	77.56	85.52
	SANN	73.21	69.35	82.69	75.44	78.82	63.81	70.53	72.98	58.22
	SASA	68.52	65.65	78.90	71.67	72.94	57.94	64.58	68.12	60.17
	Semantria	88.45	89.15	88.46	88.80	87.70	88.43	88.07	88.43	69.92
	SenticNet	70.49	71.31	63.50	67.18	69.88	76.82	73.19	70.18	80.22
	Sentiment140	75.53	0.00	0.00	0.00	75.53	100.00	86.06	43.03	52.37
	SentiStrength	95.33	95.18	96.34	95.76	95.52	94.12	94.81	95.29	41.78
	SentiWordNet	72.99	73.17	78.95	75.95	72.73	65.98	69.19	72.57	58.77
	SO-CAL	87.36	82.89	93.33	87.80	92.80	81.69	86.89	87.35	77.16
	Stanford DM	66.56	87.69	36.31	51.35	61.24	95.18	74.53	62.94	89.97
	Umigon	86.99	91.73	81.88	86.52	83.02	92.31	87.42	86.97	81.34
	Vader	94.12	100.00	90.48	95.00	86.67	100.00	92.86	93.93	9.47
Comments_Digg	AFINN	70.94	47.01	81.82	59.72	91.17	67.05	77.27	68.49	74.81
	Emolex	61.71	34.60	75.83	47.52	88.93	57.53	69.87	58.69	67.14
	Emoticons	73.08	72.22	86.67	78.79	75.00	54.55	63.16	70.97	3.32
	Emoticons DS	28.24	27.30	100.00	42.89	100.00	1.77	3.48	23.19	98.72
	Happiness Index	42.32	27.44	91.45	42.21	91.53	27.62	42.44	42.32	64.96
	NRC Hashtag	74.69	51.01	40.64	45.24	80.80	86.48	83.54	64.39	92.97
	LIWC	46.15	27.44	58.40	37.34	72.49	41.52	52.79	45.07	58.18
	Opinion Finder	71.14	43.04	64.76	51.71	86.88	73.13	79.42	65.56	56.27
	Opinion Lexicon	71.82	47.45	86.43	61.27	93.40	66.75	77.86	69.56	69.44
	PANAS-t	68.00	12.50	50.00	20.00	94.12	69.57	80.00	50.00	3.20
	Pattern.en	66.72	43.49	77.44	55.70	88.25	62.75	73.35	64.53	77.62
	SANN	60.04	35.56	84.96	50.13	91.83	52.33	66.67	58.40	61.13
	SASA	65.54	40.26	66.91	50.27	84.82	65.06	73.64	61.95	68.29
	Semantria	82.46	62.72	88.33	73.36	94.81	80.25	86.93	80.14	56.14
	SenticNet	69.40	46.30	72.46	56.50	86.77	68.25	76.40	66.45	96.55
	Sentiment140	85.45	0.00	0.00	0.00	85.45	100.00	92.15	46.08	54.48
	SentiStrength	92.09	78.69	92.31	84.96	97.40	92.02	94.64	89.80	27.49
	SentiWordNet	62.17	36.86	77.68	50.00	88.84	57.18	69.58	59.79	58.82
	SO-CAL	76.55	52.86	77.08	62.71	90.65	76.37	82.90	72.81	71.99
	Stanford DM	79.45	66.67	37.58	48.06	81.57	93.63	87.19	67.63	83.38
	Umigon	83.37	66.22	75.38	70.50	90.72	86.23	88.42	79.46	63.04
	Vader	68.65	45.29	85.63	59.24	92.31	62.50	74.53	66.89	83.63
Comments_BBC	AFINN	66.56	23.08	81.08	35.93	96.32	64.66	77.38	56.65	85.11
	Emolex	59.64	21.52	89.04	34.67	97.38	55.62	70.80	52.73	80.72
	Emoticons	33.33	0.00	0.00	0.00	100.00	33.33	50.00	25.00	0.40
	Emoticons DS	13.33	13.10	100.00	23.17	100.00	0.31	0.61	11.89	99.73
	Happiness Index	41.81	15.65	95.52	26.89	98.41	35.03	51.67	39.28	79.52
	NRC Hashtag	84.45	33.33	25.27	28.75	89.76	92.83	91.27	60.01	97.47
	LIWC	50.10	15.38	58.33	24.35	88.00	48.78	62.77	43.56	69.55
	Opinion Finder	74.43	21.74	62.50	32.26	94.93	75.72	84.24	58.25	76.46
	Opinion Lexicon	74.14	29.81	84.93	44.13	97.24	72.66	83.17	63.65	80.72
	PANAS-t	58.73	20.00	75.00	31.58	93.94	56.36	70.45	51.02	8.38
	Pattern.en	61.09	20.00	70.73	31.18	93.48	59.72	72.88	52.03	87.50
	SANN	54.34	19.09	88.06	31.38	96.88	49.80	65.78	48.58	75.13
	SASA	61.61	23.50	66.20	34.69	90.80	60.77	72.81	53.75	61.30
	Semantria	83.43	40.00	84.75	54.35	97.64	83.26	89.88	72.11	67.42
	SenticNet	66.07	24.44	74.16	36.77	94.24	64.83	76.81	56.79	88.96
	Sentiment140	92.56	0.00	0.00	0.00	92.56	100.00	96.14	48.07	51.86
	SentiStrength	93.93	64.29	78.26	70.59	97.72	95.54	96.61	83.60	32.85
	SentiWordNet	57.49	20.00	88.06	32.60	97.13	53.45	68.96	50.78	76.33
	SO-CAL	75.28	28.93	80.28	42.54	96.71	74.64	84.25	63.40	82.85
	Stanford DM	89.45	63.16	40.91	49.66	91.81	96.52	94.11	71.88	92.02
	Umigon	79.37	39.13	61.02	47.68	92.10	82.72	87.15	67.42	50.93
	Vader	62.19	22.12	85.54	35.15	96.77	59.02	73.32	54.23	92.15

Table 5 3-classes experiments results with 4 datasets

Dataset	Method	Accur.	Posit. Sentiment P	Posit. Sentiment R	Posit. Sentiment F1	Negat. Sentiment P	Negat. Sentiment R	Negat. Sentiment F1	Neut. Sentiment P	Neut. Sentiment R	Neut. Sentiment F1	MacroF1
	AFINN	62.36	61.10	70.09	65.28	44.08	55.56	49.15	71.43	58.57	64.37	59.60
	Emolex	48.74	48.15	62.71	54.47	31.27	38.59	34.55	57.90	41.30	48.21	45.74
	Emoticons	52.88	72.83	11.34	19.62	55.56	5.38	9.80	34.05	96.53	50.34	26.59
	Emoticons DS	36.59	36.55	100.00	53.53	75.00	0.36	0.71	100.00	0.03	0.07	18.10
	Happiness Index	48.81	43.61	65.27	52.29	36.96	18.28	24.46	36.82	45.16	40.56	39.10
	NRC Hashtag	36.95	42.04	75.03	53.88	24.57	56.03	34.16	53.33	3.70	6.92	31.65
	LIWC	39.54	36.52	42.33	39.21	15.14	13.02	14.00	48.64	44.83	46.66	33.29
Tweets _Semeval	Opinion Finder	57.63	67.57	27.94	39.53	40.75	33.69	36.89	58.20	86.06	69.44	48.62
	Opinion Lexicon	60.37	62.09	62.71	62.40	41.19	52.81	46.28	66.41	60.75	63.46	57.38
	PANAS-t	53.08	90.95	9.04	16.45	51.56	3.94	7.33	51.65	99.01	67.89	30.55
	Pattern.en	50.19	58.07	68.47	62.84	24.68	55.79	34.22	67.73	35.22	46.34	47.80
	SANN	54.77	52.72	47.59	50.02	38.91	29.99	33.87	58.95	66.90	62.67	48.86
	SASA	50.63	46.34	47.77	47.04	33.07	20.31	25.17	56.39	61.12	58.66	43.62
	Semantria	61.54	67.28	57.35	61.92	39.57	52.81	45.24	65.98	67.03	66.50	57.89
	SenticNet	49.68	51.85	1.26	2.46	29.79	1.67	3.17	49.82	98.51	66.17	23.93
	Sentiment140	42.25	0.00	0.00	0.00	26.79	68.10	38.45	50.57	66.14	57.31	31.92
	SentiStrength	57.83	78.01	27.13	40.25	47.80	23.42	31.44	55.49	89.89	68.62	46.77
	SentiWordNet	48.33	55.54	53.44	54.47	19.67	37.51	25.81	61.22	47.57	53.54	44.61
	SO-CAL	58.83	58.89	59.02	58.95	40.39	54.24	46.30	39.89	59.96	47.91	51.05
	Stanford DM	22.54	72.14	18.17	29.03	14.92	90.56	25.61	47.19	6.94	12.10	22.25
	Umigon	65.88	75.18	56.14	64.28	39.66	55.91	46.41	70.65	75.78	73.13	61.27
	Vader	60.05	56.08	79.26	65.68	44.13	59.74	50.76	76.88	46.02	57.57	58.01
	AFINN	64.41	40.81	72.12	52.13	49.67	62.50	55.35	85.95	62.54	72.40	59.96
	Emolex	54.76	31.67	59.95	41.44	40.14	47.54	43.53	77.48	54.64	64.08	49.68
	Emoticons	70.22	70.06	16.78	27.07	65.62	8.61	15.22	41.29	97.56	58.02	33.44
	Emoticons DS	20.34	19.78	99.46	33.00	62.07	3.69	6.96	53.85	0.55	1.09	13.68
	Happiness Index	55.16	29.13	61.98	39.64	50.65	23.98	32.55	43.35	59.16	50.03	40.74
	NRC Hashtag	30.47	28.25	77.40	41.39	24.18	72.54	36.27	79.08	8.77	15.78	31.15
	LIWC	46.88	21.85	38.43	27.86	19.18	18.24	18.70	69.51	54.83	61.31	35.95
Tweets _RND_III	Opinion Finder	71.55	57.48	32.75	41.72	49.85	34.63	40.87	75.95	89.90	82.34	54.98
	Opinion Lexicon	63.86	40.65	66.17	50.36	48.84	56.15	52.24	81.96	64.66	72.29	58.30
	PANAS-t	68.79	79.49	8.39	15.18	48.57	3.48	6.50	68.75	98.86	81.10	34.26
	Pattern.en	53.57	36.25	76.86	49.26	35.19	59.43	44.21	84.20	45.68	59.23	50.90
	SANN	66.88	42.70	48.71	45.51	46.35	36.48	40.83	77.99	77.99	77.99	54.78
	SASA	55.37	29.42	54.53	38.22	42.46	47.34	44.77	78.30	57.15	66.08	49.69
	Semantria	68.89	48.86	63.73	55.31	49.82	55.53	52.52	82.02	72.96	77.22	61.68
	SenticNet	29.97	31.08	74.83	43.92	20.98	73.98	32.68	79.70	8.49	15.35	30.65
	Sentiment140	55.05	0.00	0.00	0.00	28.14	81.35	41.81	71.14	66.00	68.47	36.76
	SentiStrength	73.80	70.94	41.95	52.72	57.53	25.82	35.64	75.35	92.26	82.95	57.10
	SentiWordNet	55.85	37.42	58.19	45.55	24.04	35.86	28.78	79.25	59.00	67.64	47.33
	SO-CAL	66.51	43.06	68.88	52.99	51.84	60.66	55.90	45.77	66.94	54.37	54.42
	Stanford DM	31.90	64.48	38.57	48.26	15.58	85.04	26.33	75.64	19.77	31.35	35.32
	Umigon	74.12	57.67	70.23	63.33	48.83	68.44	57.00	88.80	76.34	82.10	67.47
	Vader	59.82	37.52	81.73	51.43	47.99	65.98	55.57	89.26	52.28	65.94	57.64
	AFINN	50.10	16.22	60.61	25.59	82.62	56.05	66.79	40.11	30.24	34.48	42.29
	Emolex	44.10	15.51	65.66	25.10	83.19	45.48	58.81	35.27	31.85	33.47	39.13
	Emoticons	24.60	0.00	0.00	0.00	33.33	0.15	0.30	19.77	98.79	32.95	11.09
	Emoticons DS	10.00	9.85	98.99	17.92	66.67	0.31	0.61	0.00	0.00	0.00	6.18
	Happiness Index	33.60	11.83	64.65	20.00	84.93	28.48	42.66	26.46	34.68	30.02	30.89
	NRC Hashtag	64.00	20.72	23.23	21.90	70.20	91.27	79.36	52.50	8.47	14.58	38.62
	LIWC	33.00	11.11	42.42	17.61	67.69	33.69	44.99	22.90	27.42	24.95	29.18
Comments _BBC	Opinion Finder	51.80	14.96	35.35	21.02	78.76	60.18	68.23	33.71	36.29	34.95	41.40
	Opinion Lexicon	55.00	20.67	62.63	31.08	85.27	59.42	70.04	40.82	40.32	40.57	47.23
	PANAS-t	27.10	16.67	6.06	8.89	75.61	4.75	8.93	25.35	94.35	39.97	19.26
	Pattern.en	46.00	14.39	58.59	23.11	77.30	52.68	62.66	38.16	23.39	29.00	38.26
	SANN	40.10	14.50	59.60	23.32	79.49	37.98	51.40	33.45	37.90	35.54	36.75
	SASA	38.20	17.03	47.47	25.07	70.75	36.29	47.98	25.19	39.52	30.77	34.60
	Semantria	56.00	28.90	50.51	36.76	83.82	57.12	67.94	35.86	55.24	43.49	49.40
	SenticNet	47.10	17.74	66.67	28.03	72.87	57.58	64.33	25.89	11.69	16.11	36.16
	Sentiment140	50.60	0.00	0.00	0.00	73.23	55.28	63.00	28.60	58.47	38.41	33.80
	SentiStrength	44.20	47.37	18.18	26.28	86.64	32.77	47.56	29.37	84.68	43.61	39.15
	SentiWordNet	42.40	14.90	59.60	23.84	81.63	41.50	55.03	34.56	37.90	36.15	38.34
	SO-CAL	55.50	20.88	57.58	30.65	80.47	63.09	70.73	28.57	34.68	31.33	44.23
	Stanford DM	65.50	43.37	36.36	39.56	71.01	89.28	79.10	37.50	14.52	20.93	46.53
	Umigon	45.70	28.35	36.36	31.86	76.35	41.04	53.39	29.31	61.69	39.74	41.66
	Vader	49.10	15.96	71.72	26.10	82.57	55.13	66.12	50.42	24.19	32.70	41.64
	AFINN	42.45	64.81	41.79	50.81	80.29	39.82	53.24	7.89	77.87	14.32	39.46
	Emolex	42.97	55.12	53.72	54.41	75.35	33.33	46.22	7.22	54.10	12.74	37.79
	Emoticons	4.68	0.00	0.00	0.00	0.00	0.00	0.00	4.47	99.59	8.56	2.85
	Emoticons DS	42.58	42.55	99.77	59.66	78.57	0.40	0.80	0.00	0.00	0.00	20.15
	Happiness Index	31.81	48.42	50.18	49.29	71.70	15.06	24.89	5.36	54.10	9.76	27.98
	NRC Hashtag	54.84	55.38	45.74	50.10	61.55	65.68	63.55	8.33	15.16	10.76	41.47
	LIWC	24.35	42.88	27.72	33.67	53.42	19.07	28.11	4.67	53.28	8.58	23.45
Comments _NYT	Opinion Finder	29.38	68.77	18.78	29.51	76.52	32.68	45.80	6.29	88.11	11.75	29.02
	Opinion Lexicon	44.57	65.95	43.15	52.17	79.81	43.11	55.98	7.94	73.77	14.34	40.83
	PANAS-t	5.88	69.23	1.23	2.41	62.07	1.31	2.57	4.75	99.18	9.07	4.68
	Pattern.en	45.39	55.15	44.69	49.37	63.65	45.92	53.35	7.85	45.90	13.41	38.71
	SANN	27.92	56.74	29.40	38.73	78.02	22.14	34.49	5.93	79.51	11.04	28.09
	SASA	30.04	49.92	30.13	37.58	59.11	27.21	37.26	5.74	61.07	10.49	28.44
	Semantria	44.59	70.60	41.83	52.54	80.54	44.24	57.11	7.53	73.36	13.65	41.10
	SenticNet	4.70	0.00	0.00	0.00	0.00	0.00	0.00	4.70	100.00	8.98	2.99
	Sentiment140	34.66	0.00	0.00	0.00	65.76	59.88	62.68	5.83	64.34	10.69	24.46
	SentiStrength	18.17	78.51	8.62	15.54	81.12	18.96	30.74	5.41	95.49	10.24	18.84
	SentiWordNet	32.20	57.35	34.53	43.10	70.31	26.95	38.97	6.08	70.08	11.19	31.09
	SO-CAL	50.79	64.36	51.13	56.99	77.25	49.16	60.08	8.68	65.98	15.34	44.14
	Stanford DM	51.93	73.39	21.14	32.83	59.48	77.90	67.46	9.65	38.11	15.40	38.56
	Umigon	24.08	68.76	16.38	26.46	68.78	24.51	36.14	5.88	88.93	11.04	24.54
	Vader	48.84	61.96	52.40	56.78	80.09	44.02	56.81	9.51	70.90	16.77	43.46

Table 6 Best Method for each Dataset - 2-class experiments

Dataset	Method	F1-Pos	F1-Neg	Macro-F1	Coverage
Comments_BBC	SentiStrength	70.59	96.61	83.60	32.85
Comments_Digg	SentiStrength	84.96	94.64	89.80	27.49
Comments_NYT	SentiStrength	70.11	86.52	78.32	17.63
Comments_TED	Emoticons	85.71	94.12	89.92	1.65
Comments_YTB	SentiStrength	96.94	89.62	93.28	38.24
Reviews_I	SenticNet	97.39	93.66	95.52	69.41
Reviews_II	SenticNet	94.15	93.87	94.01	94.25
Myspace	PANAS-t	97.99	72.73	85.36	9.59
Amazon	Semantria	86.08	72.85	79.46	59.45
Tweets_DBT	Semantria	72.07	78.86	75.46	49.97
Tweets_RND_I	PANAS-t	88.57	80.00	84.29	4.81
Tweets_RND_II	Vader	99.25	98.33	98.79	94.61
Tweets_RND_III	PANAS-t	96.12	87.18	91.65	6.85
Tweets_RND_IV	Emoticons	94.74	94.29	94.51	78.78
Tweets_STF	PANAS-t	94.12	94.12	94.12	4.74
Tweets_SAN	SentiStrength	90.23	88.59	89.41	29.61
Tweets_Semeval	Emoticons	95.27	78.26	86.77	10.52
RW	SentiStrength	90.04	75.79	82.92	23.12

Table 7 Best Method for each Dataset - 3-class experiments

Dataset	Method	F1-Pos	F1-Neg	F1-Neu	Macro-F1
Comments_BBC	Semantria	36.76	67.94	43.49	49.40
Comments_Digg	Umigon	49.62	62.04	44.27	51.98
Comments_NYT	SO-CAL	56.99	60.08	15.34	44.14
Comments_TED	Opinion Lexicon	64.95	56.59	30.77	50.77
Comments_YTB	Vader	73.22	54.67	42.54	56.81
Myspace	Vader	79.23	43.88	39.02	54.05
Tweets_DBT	Opinion Lexicon	43.44	47.71	48.84	46.66
Tweets_RND_I	Umigon	60.53	51.39	65.22	59.05
Tweets_RND_III	Umigon	63.33	57.00	82.10	67.47
Tweets_RND_IV	Umigon	75.86	76.33	71.54	74.58
Tweets_SAN	Umigon	44.16	45.95	70.45	53.52
Tweets_Semeval	Umigon	64.28	46.41	73.13	61.27
RW	Semantria	61.38	46.30	44.53	50.74

Table 8 Friedman's Test Results

2-class experiments		3-class experiments	
FR	253.15	FR	198.56
Critical Value	31.41	Critical Value	31.41
Reject null hypothesis		**Reject null hypothesis**	

Table 9 Mean Rank Table for All Datasets

3-Classes			2-Classes			
Pos	**Method**	**Mean Rank**	**Pos**	**Method**	**Mean Rank**	**Coverage (%)**
1	Vader	3.31(3.42)	1	SentiStrength	2.00 (2.50)	29.30 (28.91)
2	AFINN	3.62	2	Semantria	3.61	62.34
3	Opinion Lexicon	3.92	3	Opinion Lexicon	5.28	69.50
4	Semantria	4.08	4	SO-CAL	6.00	72.64
5	Umigon	5.00	5	AFINN	6.61	73.05
6	SO-CAL	5.85	6	Vader	7.33 (7.71)	79.00 (78.72)
7	Pattern.en	8.08	7	Umigon	7.89	64.11
8	Emolex	9.92	8	PANAS-t	8.72	5.10
9	SANN	10.38	9	Emoticons	9.00	10.69
10	Opinion Finder	11.38	10	Pattern.en	10.22	82.03
11	SentiWordNet	11.69	11	Emolex	12.11	66.12
12	SentiStrength	12.08(11.71)	12	SenticNet	12.17	73.86
13	SASA	13.15	13	Opinion Finder	12.61	46.63
14	SenticNet	13.69	14	SANN	12.61	55.97
15	Stanford DM	13.69	15	Stanford DM	14.50	87.57
16	LIWC	14.85	16	NRC Hashtag	14.78	93.52
17	NRC Hashtag	15.00	17	SentiWordNet	15.00	61.77
18	Happiness Index	15.69	18	SASA	16.22	60.12
19	Sentiment140	17.08	19	LIWC	16.83	61.82
20	Emoticons	19.31	20	Happiness Index	17.67	60.75
21	PANAS-t	20.08	21	Sentiment140	20.67	43.50
22	Emoticons DS	21.15	22	Emoticons DS	21.17	99.36

Table 10 Contexts' Groups

Context Groups	
Social Networks	Myspace, Tweets_DBT, Tweets_DBT, Tweets_RND_I, Tweets_RND_II, Tweets_RND_III, Tweets_RND_IV, Tweets_STF, Tweets_SAN, Tweets_Semeval
Comments	Comments_BBC, Comments_DIGG, Comments_NYT, Comments_TED, Comments_YTB, RW
Reviews	Reviews_I, Reviews_I, Amazon

Table 11 Mean Rank Table for Datasets of Social Networks

3-Classes			2-Classes			
Pos	**Method**	**Mean Rank**	**Pos**	**Method**	**Mean Rank**	**Coverage (%)**
1	Umigon	2.29	1	SentiStrength	1.78(2.00)	31.54(32.18)
2	Vader	3.43(3.43)	2	Emoticons	4.00	18.04
3	AFINN	3.71	3	Semantria	4.33	61.98
4	Opinion Lexicon	4.29	4	PANAS-t	5.22	5.87
5	Semantria	4.29	5	Opinion Lexicon	5.78	66.56
6	SO-CAL	7.00	6	Umigon	6.33	71.67
7	Pattern.en	7.86	7	Vader	6.56(7.25)	75.67(73.31)
8	SANN	10.14	8	SO-CAL	6.89	67.81
9	Emolex	10.43	9	AFINN	7.00	73.37
10	SentiStrength	10.57(9.60)	10	Pattern.en	9.89	81.33
11	Opinion Finder	11.14	11	Emolex	11.78	62.63
12	SentiWordNet	11.57	12	Opinion Finder	12.00	39.58
13	SenticNet	12.57	13	SANN	12.56	50.75
14	SASA	13.14	14	SenticNet	14.89	66.26
15	LIWC	14.00	15	SentiWordNet	15.67	61.41
16	Happiness Index	15.14	16	NRC Hashtag	16.33	94.20
17	Sentiment140	16.43	17	SASA	16.44	58.57
18	Emoticons	17.71	18	LIWC	16.56	61.24
19	Stanford DM	17.86	19	Happiness Index	16.89	58.19
20	NRC Hashtag	18.29	20	Stanford DM	19.56	89.06
21	PANAS-t	19.57	21	Sentiment140	21.22	42.83
22	Emoticons DS	21.57	22	Emoticons DS	21.33	99.28

Table 12 Mean Rank Table for Datasets of Comments

3-Classes			2-Classes			
Pos	**Method**	**Mean Rank**	**Pos**	**Method**	**Mean Rank**	**Coverage (%)**
1	Vader	3.17(3.40)	1	SentiStrength	1.17(1.50)	28.29(24.02)
2	AFINN	3.50	2	Semantria	2.50	61.02
3	Opinion Lexicon	3.50	3	Opinion Lexicon	5.17	71.59
4	Semantria	3.83	4	AFINN	5.50	74.21
5	SO-CAL	4.50	5	SO-CAL	6.00	74.59
6	Umigon	8.17	6	Vader	7.67(7.80)	82.93(86.17)
7	Pattern.en	8.33	7	Umigon	9.00	57.87
8	Stanford DM	8.83	8	Emoticons	10.33	4.99
9	Emolex	9.33	9	Opinion Finder	11.33	55.66
10	SANN	10.67	10	Stanford DM	11.67	85.18
11	NRC Hashtag	11.17	11	Pattern.en	11.83	81.39
12	Opinion Finder	11.67	12	SANN	12.33	61.75
13	SentiWordNet	11.83	13	SenticNet	12.33	78.66
14	SASA	13.17	14	NRC Hashtag	12.67	93.43
15	SentiStrength	13.83(17.00)	15	Emolex	12.83	69.69
16	SenticNet	15.00	16	PANAS-t	13.50	5.10
17	LIWC	15.83	17	SentiWordNet	14.83	63.32
18	Happiness Index	16.33	18	SASA	15.33	61.91
19	Sentiment140	17.83	19	LIWC	17.33	62.24
20	Emoticons DS	20.67	20	Happiness Index	18.33	65.54
21	PANAS-t	20.67	21	Sentiment140	20.17	46.01
22	Emoticons	21.17	22	Emoticons DS	21.17	99.31

Table 13 Mean Rank Table for Datasets of Reviews

3-Classes			2-Classes			
Pos	**Method**	**Mean Rank**	**Pos**	**Method**	**Mean Rank**	**Coverage (%)**
1	-	-	1	SO-CAL	3.33	83.20
2	-	-	2	Semantria	3.67	66.04
3	-	-	3	SenticNet	3.67	87.05
4	-	-	4	Opinion Lexicon	4.00	74.14
5	-	-	5	SentiStrength	4.33 (4.33)	24.56 (24.56)
6	-	-	6	Stanford DM	5.00	87.89
7	-	-	7	AFINN	7.67	69.77
8	-	-	8	Pattern.en	8.00	85.43
9	-	-	9	Vader	9.00 (11.00)	81.11 (84.76)
10	-	-	10	PANAS-t	9.67	2.80
11	-	-	11	Umigon	10.33	53.90
12	-	-	12	Emolex	11.67	69.47
13	-	-	13	SANN	13.33	60.07
14	-	-	14	SentiWordNet	13.33	59.73
15	-	-	15	NRC Hashtag	14.33	91.64
16	-	-	16	LIWC	16.67	62.75
17	-	-	17	Opinion Finder	17.00	49.73
18	-	-	18	SASA	17.33	61.22
19	-	-	19	Happiness Index	18.67	58.84
20	-	-	20	Sentiment140	20.00	40.50
21	-	-	21	Emoticons DS	20.67	99.71
22	-	-	22	Emoticons	21.33	0.04

Table 14 Friedman's Test Results per Contexts

Context: Social Networks			
2-class experiments		3-class experiments	
FR	161.85	FR	114.77
Critical Value	31.41	Critical Value	31.41
Reject null hypothesis		**Reject null hypothesis**	
Context: Comments			
2-class experiments		3-class experiments	
FR	85.39	FR	100.09
Critical Value	31.41	Critical Value	31.41
Reject null hypothesis		**Reject null hypothesis**	
Context: Reviews			
2-class experiments		3-class experiments	
FR	56.01	FR	-
Critical Value	31.41	Critical Value	-
Reject null hypothesis		**Reject null hypothesis**	