UM BENCHMARK DE COMPARAÇÃO DE MÉTODOS PARA ANÁLISE DE SENTIMENTOS

POLLYANNA DE OLIVEIRA. GONÇALVES

UM BENCHMARK DE COMPARAÇÃO DE MÉTODOS PARA ANÁLISE DE SENTIMENTOS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Fabrício Benevenuto de Souza

Belo Horizonte

Agosto de 2015

POLLYANNA DE OLIVEIRA. GONÇALVES

UM BENCHMARK PARA COMPARAÇÃO DE MÉTODOS PARA ANÁLISE DE SENTIMENTOS

Dissertation presented to the Graduate Program in Computer Science of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Fabrício Benevenuto de Souza

Belo Horizonte

August 2015

Gonçalves, Pollyanna de Oliveira.

G635b Um benchmark para comparação de métodos para análise de sentimentos / Pollyanna de Oliveira. Gonçalves. — Belo Horizonte, 2015
xxii, 93 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Fabrício Benevenuto de Souza

1. Computação — Teses. 2. Redes sociais online — Teses. 3. Mineração de dados (Computação)–Teses. I. Orientador. II. Título.

CDU 519.6*06(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Um benchmark para comparação de métodos para análise de sentimentos

POLLYANNA DE OLIVEIRA GONÇALVES

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Fabrício Benevenuto de Souza
PROF. FABRÍCIO BENEVENUTO DE SOUZA - Orientador
Departamento de Ciência da Computação - UFMG

PROF. ADRIANO CÉSAR MACHADO PEREIRA
Departamento de Ciência da Computação - UFMG

PROF. ALEXANDRE PLASTINO DE CARVALHO
Instituto de Computação - UFF

Marcos André Gonçalves
PROF. MARCOS ANDRÉ GONÇALVES
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 14 de agosto de 2015.

For my wonderful family. Thanks for always being there for me.

Acknowledgments

This study becomes a reality with the kind support and help of many individuals. I humble extend my sincere thanks to all of them. Foremost, I would like to express my gratitude towards the Universidade Federal de Minas Gerais (UFMG) for letting me fulfill my dream of being a student here. I would also like to thank the Departamento de Ciência da Computação (DCC) and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for giving me the opportunity to conclude this study.

For my committee members, Adriano Pereira, Marcos Gonçalves and Alexandre Plastino, I am extremely grateful for your assistance and suggestions throughout my project. Most of all, I am fully indebted to my advisor Fabrício Benevenuto, for his understanding, wisdom, patience, enthusiasm, encouragement and for pushing me further than I thought I could go. And also for the N.E.R.D.S team who helped in this research, especially to Matheus Araújo for the collaboration in many papers.

For my parents and siblings who is always on my side when times I needed most. Thank you for the encouragement which helped me in completion of this journey. Last but not the least, I would like to express my gratitude to all my friends and classmates for helping me and supporting me in the hard moments, filling the past years with amazing experiences and moments of joy.

"There are only two mistakes one can make along the road to truth; not going all the way, and not starting." – Buddha

Resumo

Nos últimos anos, milhares de artigos científicos vêm explorando análise de sentimentos, várias *startups* que medem opiniões em tempo real também surgiram, assim como um número de produtos inovadores que vêm sendo desenvolvidos na área. Existem diversos métodos para medir sentimentos, incluindo abordagens léxicas e métodos de aprendizado de máquina. Apesar do grande interesse no tema e da alta popularidade de alguns desses métodos, ainda não está claro qual deles possui melhor performance na identificação de polaridade (positivo, negativo ou neutro) de uma mensagem. Tal comparação é crucial para o entendimento de potenciais limitações, vantagens e desvantagens de métodos populares. Esse estudo tem como objetivo preencher essa lacuna apresentando um *benchmark* de comparação de 21 métodos e ferramentas muito utilizados na análise de sentimentos para melhor entender suas performances. Nossa avaliação é baseada em um *benchmark* que consiste em 21 datasets rotulados, abrangendo mensagens compartilhadas em redes sociais online, *reviews* de filmes e produtos, assim como opiniões e comentários em notícias. Nossos resultados realçam limitações, vantagens e desvantagens dos métodos existentes, mostrando que suas performances variam através das bases de dados. Por fim, propomos um esforço inicial na combinação desses métodos com o objetivo de maximizar os resultados de classificação de sentimentos. Apesar da tentativa introdutória, mostramos que essa é uma estratégia promissora e que precisa de maiores investigações.

Palavras-chave: Análise de sentimentos, Mineração de opinião, Redes sociais online.

Abstract

In the last few years thousands of scientific papers have explored sentiment analysis, several startups that measures 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, negative or neutral) of a message. Such a comparison is key for understanding the potential limitations, advantages, and disadvantages of popular methods. This study aims at filling this gap by presenting a benchmark comparison of 21 widely used sentiment analysis methods and tools to better understand their strengths and weaknesses. Our evaluation is based on a benchmark of 21 labeled datasets, covering messages posted on social networks, movie and product reviews, as well as opinions and comments in news articles. Our results highlight limitations, advantages, and disadvantages of existing methods, showing that their performances varied widely across datasets. Finally, we propose initial efforts in combining these methods with the aim of maximize the results of sentiment classification. Despite of this introductory attempt, we show that this is a promising strategy that needs further investigation.

Palavras-chave: Sentiment analysis, Opinion mining, Online social networks.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Given the recent popularity of Web applications, which can be defined as the semantic Web technologies integrated into, or powering, large-scale Web applications [36], sentiment analysis has become an important research topic, mainly when considering short and informal texts, a challenging scenario. More than 7,000 articles have been written about sentiment analysis, and various start-ups are developing tools and strategies to extract sentiments from text [28]. As an example of the popularity of this area we searched for the expression “sentiment analysis” on Google Trends, a Google’s online search tool that allows the user to see how often specific keywords, subjects and phrases have been queried over a specific period of time, and we observed the growing search of this expression, as presented by Figure 1.1.

Applications of sentiment analysis include the monitoring of reviews or opinions about a company, product or a brand [37], and political analysis, including the tracking of sentiments expressed by voters about candidates for an election [96], and even analysis of stock market fluctuations [9, 65], to cite a few. Due to its applicability and importance, many studies have been recently reported and there are many researchers and companies currently developing tools and strategies to extract sentiments from texts [28].

Online Social Networks (OSNs) have become popular communication platforms for the public to logs thoughts, opinions, and sentiments about everything from social events to daily chatter. The size of the active user bases and the volume of data created daily on friendships OSNs such as Facebook[1] or Twitter[2], on professional OSNs such as LinkedIn[3], or on OSNs for share videos such as Youtube[4] are massive. Only Twitter,

[1] www.facebook.com
[2] www.twitter.com
[3] www.linkedin.com
[4] www.youtube.com

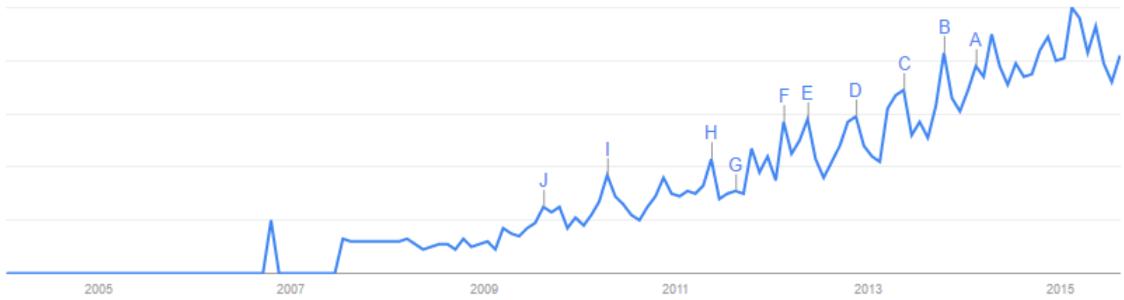


Figure 1.1. Search by the expression "sentiment analysis" at Google Trends

popular micro-blogging site, has 280 million active users, who post more than 500 million tweets[5] a day [19]. Another example is Facebook, one of the most famous online social networks, that surpassed 1 billion users registered on the website [12].

Millions of individual users are sharing the information they discover over the Web, making it an important source of breaking news during emergencies like revolutions, epidemics, and disasters [30, 48, 81]. Not surprisingly, when noteworthy events occur, users present their personal take on the events, expressing how such events were able to affect their feelings. Thus, as some messages express information about their author's emotional state, we hypothesize that messages containing feelings related to a certain event are able to unveil public sentiment about that event.

There is a number of methods for sentiment analysis that rely 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 [68]. 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 [93] was originally proposed to analyze sentiment patterns in formally written English texts, whereas PANAS-t [32] and POMS-ex [10] were proposed as psychometric scales adapted to the Web context. Other techniques include deep-learning based methods [86] and natural language processing approaches [14].

Overall, all the above techniques are acceptable by the research community and it is common to see in a single computer science conference papers that use completely different methods. However, little is known about how various sentiment methods work in the context of OSNs. In practice, sentiment methods have been widely used for developing applications without an understanding either of their applicability in the context

[5] Messages with no more than 140 characters shared on the online social network Twitter.

of OSNs, or their advantages, disadvantages, and limitations in comparison with one another. In fact, many of these methods were proposed for complete sentences, not for real-time short messages, yet little effort has been paid to apple-to-apple comparison of the most widely used sentiment analysis methods.

1.1 Objectives

The main objective of this work is to provide a comparison of many sentence-level sentiment analysis methods aiming at analyzing their advantages, disadvantages, and possible limitations. In this work, we perform a comparison among 21 sentiment analysis methods: LIWC [93], Happiness Index [25], SentiWordNet [26], SASA [99], PANAS-t [32], Emoticons [31], Emoticons DS [34], SenticNet [14], SentiStrength [94], Stanford Recursive Deep Model [86], NRC Hashtag Lexicon [54], EmoLex [56], Sentiment140 Lexicon [57] , OpinionLexicon [37], VADER [38], OpinionFinder [103], AFINN [64], SO-CAL [92], Pattern.en [22], SANN [70] and Umigon [44]. As most of the methods we compare are public available in the Web or under request to the authors, they have been increasingly used as black box for any sort of task, and this is the exactly scenery we would like to investigate in this study. We also propose initial efforts in demonstrating the feasibility of building combined methods that have the main objective of combining several of the methods considered in this study in order to maximize goals (i.e., accuracy and Macro-F1).

1.2 Results and Contributions

To address the problem of comparing and combining sentiment analysis methods, we created a benchmark that consists of 21 labeled and one unlabeled dataset that cover messages posted on social networks, movie and product reviews, and opinions and comments in news articles, TED talks, and blogs. We then survey an extensive literature on sentiment analysis to identify existing sentence-level (where each sentence of a document is individually analyzed) methods that covers several different techniques for identifying polarity (ex.: positive, negative or neutral) and we contacted authors asking their codes or we even implemented existing methods when they were unavailable, but could be reproduced from a published paper.

Our results unveil a number of important findings. First, we show that there is no single method that always achieves the best prediction performance for different datasets. We also show that existing methods varied widely in 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 existing approaches tend to be bias in their analysis towards positivity. Also, we show that methods varied widely in time performance and memory usage. Finally, we quantify relative prediction performance of existing effort in the field across different types of datasets, identified those with higher prediction performance and that can correctly classify positive, neutral, and negative messages accurately across different datasets.

As a second contribution of this work, we propose initial efforts in developing a combined method aiming at combining the outputs of all 21 methods. Despite this method is based on simple combining technique, our results show that these are promising strategies that needs further investigation.

1.3 Organization

The rest of this document is organized as it follows:

- **Chapter 2 - Sentiment Analysis.** This chapter presents an overview of the main concepts and terminologies related to sentiment analysis area, as well as a description of the levels of granularity of sentiment detection commonly used, and also a discussion about possible applications of sentiment analysis. Furthermore, we describe existing approaches and techniques on the literature, and we also describe the 21 methods for sentiment analysis considered in this study.

- **Chapter 3 - Datasets.** This chapter presents our effort to build a large and representative standard dataset consists of obtaining labeled data from trustful previous efforts that cover a wide range of sources and kinds of data.

- **Chapter 4 - Methodology.** This chapter presents our methodology of the comparison and combination processes of all 21 sentiment analysis methods, including a description of the measures used in this task, highlighting advantages, disadvantages, limitations and possible improvements.

- **Chapter 5 - Results and Discussions.** This chapter presents the results of the comparison among all sentiment analysis methods analyzing the proposed measures of prediction performance, percentage of agreement among all methods, winning number score and polarity detection in global events, highlighting the advantages, disadvantages and possible limitations of each method. We also present in this chapter the results of the proposed combined method, highlighting its limitations and possible improvements.

- **Chapter 6 - Conclusions and Future Work.** This chapter presents the conclusions of this study, highlighting its main contributions and prospects for future work.

1.4 Publications

As we show, a few papers were published since the beginning of this study. Some results presented in this papers are not part of this thesis, but they contributed to build it.

- Pollyanna Gonçalves, Wellington Dores, and Fabrício Benevenuto. “Panas-t: Uma Escala Psicométrica para Análise de Sentimentos no Twitter”. I Brazilian Workshop on Social Network Analysis and Mining (BraSNAM). 2012.

- Pollyanna Gonçalves, Fabrício Benevenuto, and Meeyoung Cha. “Panas-t: A Psychometric Scale for Measuring Sentiments on Twitter”. CoRR arXiv:1308.1857. 2013.

- Pollyanna Gonçalves, Fabrício Benevenuto, and Virgílio Almeida. “O Que Tweets Contendo Emoticons Podem Revelar sobre Sentimentos Coletivos?”. II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM). 2013.

- Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha. “Comparing and Combining Sentiment Analysis Methods”. In Proceedings of the first ACM conference on Online social networks (COSN). ACM, New York, 27-38. DOI=10.1145/2512938.2512951. 2013

- Pollyanna Gonçalves, Daniel Hasan Dalip, Júlio Reis, Johnnatan Messias, Filipe Ribeiro, Philipe Melo, Leandro Araújo, Fabrício Benevenuto, and Marcos Gonçalves. “Bazinga! Caracterizando e Detectando Sarcasmo e Ironia no Twitter”. IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM). 2015.

- Pollyanna Gonçalves, Matheus Araújo, Filipe Ribeiro, and Fabrício Benevenuto. “A Benchmark Comparison of Sentence-Level Sentiment Analysis Methods”. ACM Computer Surveys. 2015. (submitted)

Chapter 2

Sentiment Analysis

Sentiment analysis have been applied in many studies, products, services and domains. Therefore, it is clear that a comparison among strategies and techniques proposed in the literature is necessary for better understand sentiment analysis area. For that reason, in this chapter we introduce fundamental concepts and approaches presenting often used terminologies and applications examples. We also discuss about the 21 methods considered in this work and its strategies for measuring sentiment on texts from Web.

2.1 Definitions and Terminologies

Due to the recent popularity of sentiment analysis topic, many terms have been used to describe same tasks in detecting sentiments. In order to present this different definitions and terminologies and also situate this study, we describe it as follow:

- **Polarity:** This term represents the degree of positivity, negativity or neutrality of a sentence.

- **Emotion:** This term indicates the sentiment or a mood that the author has related to a specific subject (e.g.: surprise, anger, happiness, etc.) [46].

- **Strength:** This term represents the intensity of an emotion, a feeling or a specific polarity.

- **Subjectivity:** This term is used by methods that are focused on the classification of the subjectivity of a message. For example, informal texts (e.g.: texts from OSNs) are more subjective than formal texts (e.g.: texts from articles and news).

- **Opinion:** This term represents a personal point of view of the author about a specific subject (e.g.: a review of a movie, of a brand, or of a product) [95].

We explored a wide range of tools and methods proposed for this task and observed that they are proposed for different levels of granularity of a document. The granularity level says that the classification given by a method may be attached to whole documents (for document-based sentiment), to individual sentences (for sentence-based sentiment) or to specific aspects of entities (for aspect-based sentiment) [28]. In other words, the lower the granularity, the more specific the sentiment classification is. Next, we better describe these three levels of granularity:

- ***Sentence-level*:** This granularity level is based on the fact that in a single document there are multiple polarity involved [67]. This level is often uses when we want to have a more fine-grained view of the different opinions expressed in the document about the entities [28]. Most approaches using this granularity in sentiment analysis are either based on supervised learning [53] or on unsupervised learning [111].

- ***Aspect-level*:** This granularity level is based in the hypothesis that in many cases people talk about entities that have many aspects (attributes) and they have a different opinion about each of the aspects [28]. In other words, in this level a sentence can be judged by different entities and may have different polarities associated with it [67]. This strategy of often used for reviews. For example, the sentence "This hotel, despite the great room, have a terrible customer service" has two different polarities associated with "room" and "customer service" for the same hotel. While "room" has a positive polarity associated with it, "customer service" is judged in a negative way. Many researchers have been using this approach to the sentiment detection task ([33, 73, 107])

- ***Document-level*:** At this granularity level, the polarity classification occurs at the document level, in order to detect polarity of a whole text at once. This is considered the simplest form of sentiment analysis and assumes that all the document is related to a single entity, such as a specific product or topic and consequently, associated with a single polarity [95]. Pang et al. [68] show that even in this simple granularity level, good accuracy can be achieved.

In this study, we are focused in the polarity (positivity, negativity or neutrality) detection of messages shared on many Web domains and in the in sentence-level granularity. Next, we discuss some practical applications for the sentiment analysis.

2.2 Applications for Sentiment Analysis

With the recent advent of sentiment analysis as a hot topic on scientific researches, many applications have been proposed on areas such as commerce, tourism, political, economics and health. The most common application is in the area of reviews analysis. There are many online Web tools that provide automated information of reviews about products, services or brands (e.g..: Trackur[1], Sendible[2] and Meltwater[3]), helping companies to monitoring public opinions.

Sentiment analysis can also be used in the development of tools for monitoring and prediction of stock market behavior. In this systems, models for predict key stock market variables can be developed using sentiment analysis strategies on data from Web. As said by Nuno O. et al. [65], the community of users that utilizes these microblogging services to share information about stock market issues has grown and is potentially more representative of all investors. Many models ([8]) and online tools (e.g.: StockFluence[4] and TheySay[5]) were proposed in this area.

In health-care, although the health professional is the expert in diagnosing, sentiment analysis have been used in the development of systems that monitor mental diseases such as postpartum depression on online social networks ([21, 84]).

Next we present various techniques and methods proposed by literature and also describe the 21 methods considered in this work.

2.3 Existing Approaches for Sentiment Analysis

In this section, we describe methods for sentiment analysis proposed in the literature and that will be used in this work. Methods can be divided in three types: (i) machine learning-based methods; (ii) lexical-based methods; and (iii) hybrid methods.

In the following sections, we describe each of these types and also describe the methods that will be used in this work.

2.3.1 Machine Learning Approaches

Machine-learning-based methods relies on well-known machine learning algorithms to solve the sentiment analysis task as a regular text classification problem. Machine

[1] www.trackur.com
[2] www.sendible.com
[3] www.meltwater.com
[4] www.stockfluence.com
[5] www.theysay.io

learning algorithms varies depending of the type of features that will be extracted from a sentence in order to classify sentiments and also the amount of labeled data available. Machine learning methods are suitable for applications that need content-driven or adaptive polarity identification models.

2.3.1.1 Supervised Learning

Supervised learning depend on the existence of labeled documents, in our case, sentences where the positive/negative label is already linked to it. Supervised learning can be divided in the following types.

- **Probabilistic classifiers:** These classifiers are able to predict a probability distribution over a set of classes, rather than just provide a class for a given sentence.

  - **Naïve Bayes classifier (NB):** The simplest and most commonly used classifier. Based on the distribution of the words in a sentence, NB algorithm calculates the posterior probability of a class. This classifier was used by [40] in order to solve a problem where polarity detection using lexicon dictionaries has a positive bias. They evaluate the NB classifier in a dataset with restaurants reviews and improved the recall and precision rates in at least 10%.
  - **Maximum Entropy Classifier (ME):** This type of classifier provides the least biased estimate possible based on the given information. As said by [39], the ME classifier "is maximally noncommittal with regards to missing information". ME classifier was used by [42] to deal with natural language processing tasks, particularly statistical machine translation.

- **Linear classifiers:** These classifiers are known for create a linear predictor that separate hyperplanes among different classes using a normalized document word frequency and vectors of linear coefficients with same dimensionality as the feature space.

  - **Support Vector Machine classifier (SVM):** A SVM model is a representation of instances as points in space, mapped so that the instances of each class are divided by a space that is as wide as possible. New instances are then mapped in the same space and predicted as belonging to a class based on which side of the space they are placed. SVM was used by [45] as a sentiment polarity classifier and proposed a framework that provides a numeric summarization of opinions on micro-blogs. Is important to remember that SVM also works if the data set does not allow classification by a linear classifier. In this case the SMV non-linear maps every data point into a higher dimensional space via some transformation where the data training is separable.

  - **Neural Network (NN):** NN classifiers are networks of "neurons" based on the neural structure of the brain. Each information processed by a neuron generates a weight depending on the result. Neurons get scores when achieve hits and lost scores when make mistakes. The errors from the initial classification of the first record is fed back into the network, and used to modify the networks algorithm the second time around, and so on for many iterations. SVM and NN were used by [98] to solve the problem of mark relationships between two persons as positive, neutral, or unknown, one person being a topic pf a biography and the other being mentioned in this biography.

- **Decision tree classifiers:** These classifiers provides a hierarchical decomposition of the training data space in which a condition on the attribute value is used to divide the data [76]. Classifiers based on decision tree are similar to if-then rules. Each node of a decision tree is a value test and each branch of this node is identified with the possible test values. This type of classifier was used by [110] to solve the problem of mining the content structures of topical terms in sentence-level contexts to discover the links among a specific topical term and its context words.

2.3.1.2 Weakly, Semi and Unsupervised Learning

Although the proved efficacy of supervised learning, a major drawback of this methods based on this type of learning is the need of labeled data. In text classification, it is sometimes difficult to create these labeled datasets for training. The unsupervised approach overcome this difficulty. Unsupervised learning classifiers try to find relevant patterns on data in order divide it in clusters, and then each cluster is the representation of a single class. One of the simplest unsupervised classifier is called K-means, that consists of find closer instances of fixed centroids and then recalculated each centroid until they convergence. Unsupervised approach was used by [108] to solve the problem of automatically discovering sentiments associated with aspects on Chinese social reviews.

Weakly and semi-supervised learning are a class of supervised learning task and techniques that make use of a small amount of labeled data with a large amount of unlabeled data. Despite the small number of labeled data, many researches showed that this technique can produce considerable improvement in learning efficacy. He Y. and Zhou D. [35] proposed a framework for sentiment detection that weakly-learned from a pseud-labeled examples bases on a prior information extracted from an existing sentiment lexicon. Authors showed that the proposed method outperforms existing weakly-supervised sentiment classification methods despite using no labeled documents.

2.3.1.3 Meta classifiers

This is a classifier that does not implement a classification algorithm on its own, but uses another classifier to do the actual work. In other words, a meta classifier is focused on predicting the right algorithm for a particular problem based on characteristics of the dataset or based on the performance of other, simpler learning algorithms [85].

In [58], authors investigated investigate hybrid approaches, developed as a combination of the learning and lexical algorithms. The authors did not obtain significant improvements over the individual techniques for this particular domain. By analyzing different datasets and considering much more techniques as part of our ensembles, we noted that it is possible to obtain significant improvements over existing techniques depending on the domain.

Wu et. al[112] explored an entity-level sentiment analysis method specific to the Twitter data. A sentiment analysis in the entity-level granularity provides sentiment associated with a specific entity in the data (e.g. about a single product). In that work, authors combined lexicon-based and learning-based methods in order to increase the recall rate of individual methods in Twitter data. Similarly, [60] proposed *pSenti*, a method for sentiment analysis developed as a combination of lexicon and learning approaches for a different granularity level, the concept-level (semantic analysis of text by means of web ontology or semantic networks).

In this work, we used three methods that rely on machine-learning approaches:

- **Sail/Ail Sentiment Analyzer (SASA):**

  We consider one more machine learning-based method called the SASA [99]. The open source tool was evaluated by the Amazon Mechanical Turk (AMT) [6], where "turkers" were invited to label tweets as positive, negative, neutral, or undefined, resulting in a dataset of about 17,000 labelled tweets.

[6]www.mturk.com

SASA was originally proposed to be a real-time method that detects public sentiments on Twitter during the 2012 U.S. presidential election. Authors built a sentiment method based in the use of the statistical classifier Naïve Bayes on unigram features. These features were calculated from tokenization of the tweets that attempts to preserve punctuation that may signify sentiment (e.g.; emoticons and exclamation points) [99]. SASA classify messages in a range of [-1, 1], with -1 and 1 being the most negative and most positive score. In this work, we will consider scores less than zero as negative, equals to zero as neutral and greater than zero as positive.

We include SASA in particular because it is an open source tool and further because there had been no comparison of this method against others in the literature. We used the SASA python package version 0.1.3, which is available at `https://pypi.python.org/pypi/sasa/0.1.3`.

- **Stanford Recursive Deep Model:**

  Stanford Recursive Deep Model, simple called here as SRDM, is a method for sentiment detection proposed by [86]. The method was proposed using a dataset with almost 11,000 sentences from online movie reviews, where half of which were considered negative and the other half positive. First of all, authors used the Stanford Parser [43] to create random sentences from the original dataset, resulting in other 215,000 phrases. Then, “turkers” from Amazon Mechanical Turk[7] labeled each sentence in a scale range from very negative to very positive, passing through the neutral sentiment.

  Then, authors proposed a new model called Recursive Neural Tensor Network (RNTN) that processes all sentences dealing with the structures of each sentence and compute the interactions among them. This approach is interesting since RNTN deals with the order of words in a sentence, which is ignored in most of methods. For instance, in the sentence “This movie was actually neither that funny, nor super witty”, shared by authors in their paper, most of methods would labeled it as a positive sentence, because of the words “funny” and “witty”. But, besides the method proposed learned that funny and witty are positive, it can realize that the sentence is actually negative. Stanford Recursive Deep Model classify messages as “Negative”, “Very Negative”, “Neutral”, “Positive” and “Very positive”, in this work we will consider “Negative” and “Very Negative” to be negative, and “Positive” and “Very positive” to be positive.

[7] www.mturk.com

Stanford Recursive Deep Model is integrated into Stanford CoreNLP as of version 3.3.0 and is available in `http://nlp.stanford.edu/software/corenlp.shtml`.

- **Pattern.en:**

  Pattern [22] is a package for Python programming language with components for web mining, natural language processing, machine learning and network analysis in English texts. Pattern is organized in separated modules that covers its functionalities. For example, *pattern.search* is used for do queries by syntax and semantics and *pattern.vector* is used to train a classifier. In this study, we are focused in the sentiment analysis use of the package, possible with the use of the *pattern.en* module.

  Pattern.en module was built to be a fast, regular expressions-based shallow parser for English using a finite state part-of-speech tagger extended with a tokenizer, lemmatisation and chunker [22]. This module also offers functions for sentiment analysis based on the WordNet corpus [51].

  Pattern.en is integrated into Pattern package and is available in `http://www.clips.ua.ac.be/pages/pattern`.

2.3.2 Lexicon-based Approaches

Differently from machine learning approaches, strategies based on lexical dictionaries do not require training and, consequently, do not require labeled datasets.

2.3.2.1 Dictionary-based approach

Dictionary-based methods utilize a provided list of pre-defined words to identify sentiments in texts. This list, which is commonly called as dictionary, is usually collected manually with known orientations, and then is increased associating synonyms or related words using corpora such as WordNet [51]. Qiu et. al [75] used this approach to identify sentiment sentences in contextual advertising.

The main disadvantage of this approach is in the fact that the lexicon must be reconstructed in order to adapt itself to a new dataset context. Therefore, lexical methods hardly have high performance rates in different databases.

In this work, we used five methods that rely on dictionary-based approach:

- **Emoticons:**

  The simplest to detect the way polarity (i.e., positive and negative affect) of a message is based on the emoticons it contains. Emoticons have become popular in recent years, to the extent that some (e.g. <3) are now included in English Oxford Dictionary [27]. Emoticons are primarily face-based and represent happy or sad feelings, although a wide range of non-facial variations exist: for instance, <3 represents a heart and expresses love or affection.

To extract polarity from emoticons, we utilize a set of common emoticons from [20, 50, 59] proposed in a previous work ([31]) and listed in Table 2.1. This table also includes the popular variations that express the primary polarities of positive and negative. Messages with more than one emoticon were associated to the polarity of the first emoticon that appeared in the text, although we encountered only a small number of such cases in the data.

Table 2.1. Emoticons symbols and its variations

Emoticon	Polarity	Symbols
	Positive	:) :] :} :o) :o] :o} :-] :-) :-} =) =] =} =^] =^) =^} :B :^B :^D :^B =B =^B =^D :') :'] :'} :-B :^D =') ='] ='} :-D <3 ^.^ ^-^ ^_^ ^^ :* =* :-* ;) ;] ;} :-p :-P :-b :^p :^P :^b =P =p \o\ /o/ :P :p :b =b =^p =^P =^b \o/

	Negative	D: D= D-: D^: D^= :(:[:{ :o(:o[:^(:^[:^{ =^(=^{
		:-[:-(=(=[={ =^[>:-=(>=[>=(>=[>={ >=(>:-{ >:-[
		>:-(>=^[>:-(:'(:'[:'{ ='{ ='(='[=\ :\ =/ :/ =$
		o.O O_o Oo :$:-{ >:-{ >=^(>=^{ :o{

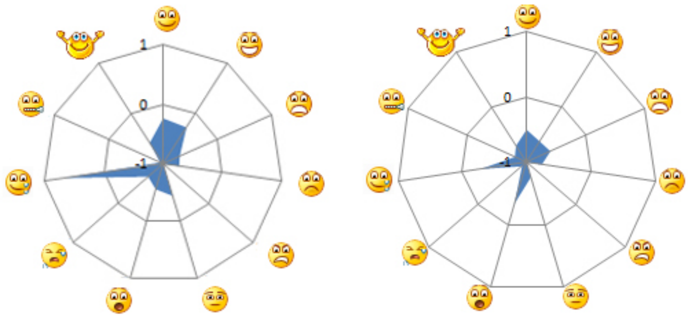
	Neutral	:| =| :-| >.< >< >_< :o :O =O :@ =@ :^o :^@ -.- -.-' -_- -_-' :x =X =# :-x :-@ :-# :^x :^# :#

This method was evaluated using a large dataset consisting of global events filtered from Twitter where sentiments related to them are easy to be assumed. Figures 2.1(a) and 2.1(b) show the sentiments calculated by Emoticons on Twitter for the Susan Boyle appearance on a TV' show and for the Obama's presidential inauguration, in 2009. In this figures, we can see that users tended to use more emoticons associated with happiness (considered as positive by the method) in the first, and surprise (considered as neutral by the method) in the second event.

As one may expect, the rate of OSN messages containing at least one emoticon is very low compared to the total number of messages that could express emotion. A recent work has identified that this rate is less than 10% [71] in Twitter. Therefore, emoticons have been often used in combination with other techniques for building a training dataset in supervised machine learning techniques [77].

- **Emoticons DS:**

On their study, Hannak A. et al [34] made an effort to construct automatically a large sentiment scored word list using a corpus of over 1.5 billion tweets collected



(a) 2009Susan Boyle's appearance (b) 2009Obama's presidential inauguration

Figure 2.1. E valuationof Emoticons method on two global events filtered from Twitter

by [15]. The methodology used to associate polarity to terms extracted from each tweet consisted on classify checking the existence of what authors called clearly positive and negative emoticons. The score given to each word extracted after the tokenizer of the tweet is calculated as the relative fraction of times each token occurs with a positive or negative emoticon. At the end, each token's score ranges between -1 and 1 indicating the polarity of it.

The process of evaluation consisted in the using of AMT for labeling 1,000 tweets, each one rated by 10 "turkers". The final correlation coefficient of the word list was 0.651, what authors considered a good result.

Since authors did not named the list, in this work will be defined as Emoticons DS method. Emoticons DS list used in this work was kindly sent to us by authors.

- **NRC Hashtag Sentiment Lexicon:**

  The NRC Hashtag Sentiment Lexicon [54] is a lexicon dictionary of Twitter's hashtags with associations to eight sentiments: joy, sadness, anger, fear, trust, disgust, anticipation and surprise. Just like EmoLex, from these sentiments we consider joy and trust as positive, sadness, anger, fear and disgust as negative, and anticipation and surprise as neutral.

  The dictionary of up to 32,000 hashtags was created from a collection of 775,310 tweets posted between April and December 2012 that had a positive or a negative hashtag, such as #good and #excellent. Results of the referenced paper showed that emotion hashtags assigned to tweets are efficient for detecting emotion in other tweets.

In this work, we used the NRC Hashtag Sentiment Lexicon version 0.2, which the authors kindly sent to us. We grouped sentiments as positive and negative as we did for Emolex.

- **EmoLex:**

  The EmoLex [56], or NRC Emotion Lexicon, is lexical method with up 10,000 word-sense pairs. Each entry lists the association of the a word-sense pair with 8 basic sentiments: joy, sadness, anger, fear, trust, disgust, anticipation and surprise, defined by [72]. From these sentiments we consider joy and trust as positive, sadness, anger, fear and disgust as negative, and anticipation and surprise as neutral. The method was built using a large dataset consisting of words labeled using Amazon Mechanical Turk[8] service, and also words from General Inquirer [88] and WordNet Affect Lexicon (WAL) [97].

  We used NRC Emotion Lexicon version 0.92, which was available from the authors of the method.

- **OpinionLexicon:**

  OpinionLexicon [37], also known as Sentiment Lexicon, is a lexical method that measures the polarity of a sentence. It consists of two lists with 2,006 positive words and 4,783 negative words. The dictionary was built using data mining techniques in consumers reviews about products sold online, and then labeling it as positive or negative. OpinionLexicon includes slang, misspellings, morphological variants, and social-media markups. In this work, each message classified will receive label 1 if positive, -1 if negative and 0 if neutral (in the case that OpinionLexicon could not find any word of the dictionary associated in the message).

  OpinionLexicon is available for download at `http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`.

- **Valence Aware Dictionary for Sentiment Reasoning (VADER):**

  Proposed by [38], VADER is a human-validated sentiment analysis method developed for twitter and social media contexts. VADER is focused in detecting sentiments on social media style text, and it requires no training data. According to authors, VADER was constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon.

  Authors constructed and empirically validate a list of candidate lexical features associated with sentiment intensity measures, including a full list of Western-style

[8] www.mturk.com

emoticons [9], sentiment-related acronyms and initialisms [10], and commonly used slang [11]. All these features were analyzed with respect to its applicability using the wisdom-of-the-crowd (WotC) approach [90], collecting ratings on each of their candidate lexical features from ten independent human raters in AMT [3]. In this work, each message classified will receive label 1 if positive, -1 if negative and 0 if neutral (in the case that VADER could not find any word of the dictionary associated in the message).

The validation process of the method consisted of the analysis of its prediction performance in four labeled dataset collected by authors, consisting of movie-reviews, Amazon product reviews, New York Times opinion news editorials/articles and tweets. VADER is available for download at `http://comp.social.gatech.edu/papers/`.

- **Happiness Index:**

  Happiness Index [25] is a sentiment scale that uses the popular Affective Norms for English Words (ANEW) [11]. ANEW is a collection of 1,034 words commonly used associated with their affective dimensions of valence, arousal, and dominance. Happiness Index was constructed based on the ANEW terms and has scores for a given text between 1 and 9, indicating the amount of happiness existing in the text. The authors calculated the frequency that each word from the ANEW appears in the text and then computed a weighted average of the valence of the ANEW study words. The validation of the Happiness Index score is based on examples. In particular, the authors applied it to a dataset of song lyrics, song titles, and blog sentences. They found that the happiness score for song lyrics had declined from 1961 to 2007, while the score for blog posts in the same period had increased.

  In order to adapt Happiness Index for detecting polarity, in this work we consider any text that is classified with this method in the range of [1..5) to be negative and in the range of [5..9] to be positive.

- **AFINN:**

  Created by [64], AFINN consist of a list with English words associated with a integer between minus five (negative) and plus five (positive). The first version of the list (AFINN-96 [63]) contains 1,468 words and phrases manually labeled

[9] `http://en.wikipedia.org/wiki/List_of_emoticons#Western`
[10] `http://en.wikipedia.org/wiki/List_of_acronyms`
[11] `http://www.internetslang.com`

by the author and was built using tweets about the United Nation Climate Conference (COP15). The newest version (AFINN-111) was increased and consists of 2,477 words and phrases. This version was built using not only tweets but also words from the public domain Original Balanced Affective Word List [12], internet slangs and acronyms (such as "WTF", "LOL" and "ROFL") from Urban Dictionary [13], The Compass DeRose Guide to Emotion Words [14] and the Microsoft Web n-gram similarity Web service ("Clustering words based on context similarity" [15]). Author explain that the word list to have a bias towards negative words (68%), and compares it to OpinionFinder's bias (64%).

AFINN-111 was compared to ANEW, General Inquirer, OpinionFinder and SentiStrength in a dataset with 1,000 tweets labeled with AMT and collected by Alan Mislove for the Twitter-Mood
"Pulse of Nation" [16] study [6].

In this work, we used AFINN-111 version that is available for download at `http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010`.

- **Semantic Orientation CALculator (SO-CAL):**

  SO-CAL [92] is a dictionary-based method proposed to classify the polarity of texts taking into consideration the semantic orientation (SO) of words. Determine the semantic orientation of a word consists of label it as positive or negative towards a particular subject matter and also rate the strength of this polarity. This method uses dictionaries of words annotated with their semantic orientation (polarity and strength), and incorporates intensification and negation.

  The method first extract of sentiment-bearing words (e.g.: adjectives, verbs, nouns, and adverbs) and then use these words to calculate semantic orientation, taking into account valence shifters (intensifiers, downtoners, negation and irrealis markers). Authors explain that adjectives have been used in sentiment analysis as the primary source of subjective content in a document. They also say that, generally, the semantic orientation of a text is the combined effect of the adjectives or relevant words found within, based upon a dictionary of word scores. SO-CAL was built using a dataset with 400 reviews collected from Epinions [17] that includes books, cars, computers, cookware, hotels, movies, music, and phones reviews.

[12] http://www.sci.sdsu.edu/CAL/wordlist/origwordlist.html
[13] http://www.urbandictionary.com/
[14] http://www.derose.net/steve/resources/emotionwords/ewords.html
[15] http://web-ngram.research.microsoft.com/similarity/
[16] http://www.ccs.neu.edu/home/amislove/twittermood/
[17] www.epinions.com

SO-CAL was compared with other dictionaries (manually and automatically created) such as OpinionFinder MPQA, General Inquirer, SentiWordNet, Maryland Dictionary [55] and with a previous version of the method [91].

The SO-CAL version used in this work was kindly sent to us by authors.

- **Umigon:**

  Umigon [44] belongs to the family o lexicon-based method and was proposed to detect sentiments on tweets and also indicates subjectivity markers. The method classify tweets in 4 steps: (i) Detection of semantic features using onomatopes,exclamations such as "yeaaaaaaaah" and emoticons; (ii) Hashtag evaluation with the use of techniques for decomposing hashtags like'#greatstuff" and "#notveryexciting"; (iii) Decomposition in n-grams (up to 4-grams); and (iv) Post-procession. In the last step, a series of heuristics that were defined using the techniques used in previous steps are applied in order to output a single polarity. Lists was created for positive, negative, strengthen and negation words, each one with different heuristics for classification.

  The method was evaluated in a semantic evaluation task proposed by SemEval2013 [18] with a dataset of 3,813 tweets labeled as positive, negative or neutral. Umigon was also compared with Sentiment140 Lexicon.

  Umigon is a open source method and available for download at `https://github.com/seinecle/Umigon` in the version 2.0.

  The definition of psychometric scales come from psychology and refers to a set of techniques used to measure human behaviors. Psychometric scales are commonly applied in the form of questionnaires (psychological tests) where interviewed expose their opinion, usually in the form of scores, associated with a feeling about a specific context. Such questionnaires are previously scientific tested by a medical society in order to prove its efficiency in detecting human behaviors [41].

  In this work, we used one method that relies on psychometric scale-based approach:

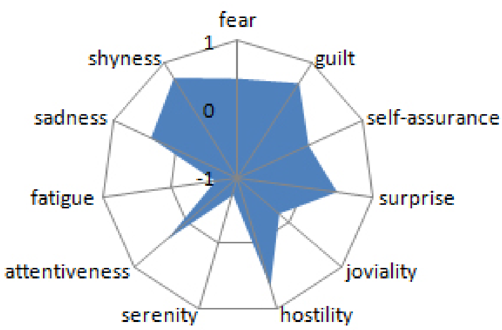
- **PANAS-t:**

  PANAS-t is a lexical method proposed in a previous work [32] to detect mood fluctuations of users on Twitter. The method consists of an adapted version of the psychometric scale Positive Affect Negative Affect Scale (PANAS [100]) extended version (PANAS-ex [101]), which is a well-known method in psychology.

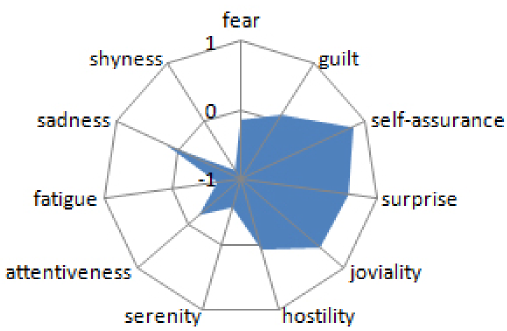
[18] https://www.cs.york.ac.uk/semeval-2013/

The definition of psychometric scales come from psychology and refers to a set of techniques used to measure human behaviors. Psychometric scales are commonly applied in the form of questionnaires (psychological tests) where interviewed expose their opinion, usually in the form of scores, associated with a feeling about a specific context. Such questionnaires are previously scientific tested by a medical society in order to prove its efficiency in detecting human behaviors [41].

The PANAS-t is based on a set of words associated with eleven moods: joviality, assurance, serenity, surprise, fear, sadness, guilt, hostility, shyness, fatigue, and attentiveness. This method was designed to track any increase or decrease in sentiments over time. The method was evaluated using a large dataset consisting of global events filtered from Twitter where sentiments related to them are easy to be assumed. Figures 2.2(a) and 2.2(b) show the sentiments calculated by PANAS-t on Twitter for the Samoa's earthquake and for the Obama's presidential inauguration, in 2009. As we can see, users tended to use words associated with fear and sadness (considered negative by the method) for the first event, and words associated with self-assurance and joviality (considered positive by the method) in the second one. The original method only considers messages that contains the expressions "i am", "feeling", "me", "myself" and its variations. In this work, this restriction was removed since there are datasets where this kind of expressions may not appear. PANAS-t assumes joviality, assurance, serenity, and surprise to be positive affect, fear, sadness, guilt, hostility, shyness, and fatigue to be negative affect, and attentiveness to be neutral.



(a) 2009Samoa's earhquake



(b) 2009Obama's presidential inauguration

Figure 2.2. PANAS-t evaluation on two global events filtered from Twitter

A few studies have been adapting PANAS-ex in order to measure human affective states in social media [17, 18], not only as a lexical-based method but also as a training corpus for a supervised method.

2.3.2.2 Corpus-based approach

Differently of the dictionary-based approach, that typically use synsets and hierarchies to acquire opinion words, corpus-bases approach often use a double propagation among opinion words and the items they modify. In other words, these methods depend on syntactic patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus [1]. This approach use conventions or connectives (e.g.: AND, OR, BUT, etc) to identify opinion words.

The corpus-based approach is performed using statistical or semantic approach, as described next:

- **Statistical approach:** Corpus-based oriented methods can use statistical techniques to the task of find co-occurrence patterns or seed opinion words. As proposed by [47], this could be done by deriving posterior polarities using the co-occurrence of adjectives in a corpus, so, the polarity of a word could be identified by studying the ocurrence frequency of a word another text [78].

- **Semantic Approach:** This approach represent methods that extract semantic features associated with specific sentiments to detect polarity in documents. Features consist of semantic concepts (eg.: person, company, etc.) that represent entities (eg.: Steve Jobs, Vodafone, etc.) extracted from documents [80]. The idea behind this approach is that certain entities and concepts could have a consistent correlation with positive or negative polarities.

This approach is often used when we want to find domain and context specific opinion words and domain dependent orientations (positive, negative or neutral). However, the main disadvantage of this approach (and also the dictionary-based approach) is in the fact that is hard to prepare a huge corpus to cover all words.

In this work, we used two methods that rely on corpus-based approach:

- **LIWC:**

  LIWC (Linguistic Inquiry and Word Count) [93] is a text analysis tool that evaluates emotional, cognitive, and structural components of a given text based on the use of a dictionary containing words and their classified categories. In addition to detecting positive and negative affects in a given text, LIWC provides other sets of sentiment categories. For example, the word “agree" belongs to the following word categories: assent, affective, positive emotion, positive feeling, and cognitive process. In this work, we will consider messages that obtained greater positive affect score than negative affect score as positive, less positive affect score than negative affect score as negative, and neutral otherwise.

The LIWC software is commercial and provides optimization options such as allowing users to include customized dictionaries instead of the standard ones. For this work, we used the LIWC2007 version and its English dictionary, which is the most current version and contains labels for more than 4,500 words and 100 word categories. LIWC is available at `http://www.liwc.net/`. In order to measure polarity, we examined the relative rate of positive and negative affects in the feeling categories.

- **SenticNet:**

  SenticNet [14] is a method of opinion mining and sentiment analysis that explores Web semantic techniques. The goal of SenticNet is to infer the polarity of common sense concepts from natural language text at a semantic level, rather than at the syntactic level. The method uses Natural Language Processing (NLP) techniques to create a polarity for nearly 14,000 concepts. For instance, to interpret a message “Boring, it’s Monday morning", SenticNet first tries to identify concepts, which are “boring" and “Monday morning" in this case. Then it gives polarity score to each concept, in this case, -0.383 for “boring", and +0.228 for “Monday morning". The resulting sentiment score of SenticNet for this example is -0.077, which is the average of these values. In this work, we will consider scores less than zero as negative, equals to zero as neutral and greater than zero as positive

  SenticNet was tested and evaluated as a tool to measure the level of polarity in opinions of patients about the National Health Service in England [13]. The authors also tested SenticNet with data from LiveJournal blogs, where posts were labeled by the authors with over 130 moods, then categorized as either positive or negative [77, 87].

  We use SenticNet version 2.0, which is available at `http://sentic.net/`.

2.3.3 Hybrid Approaches

In hybrid techniques both combination of machine learning and lexicon base approaches are used [16]. There are many sentiment analysis methods that combines lexical and learning techniques. Researchers often use this type of strategy in order obtain the best of both worlds (ie.: accuracy as well as macroF1) and consequently improve the performance of a classifier.

In this work, we used five methods that relies on hybrid approach for sentiment analysis:

- **SentiWordNet:**

  SentiWordNet [26] is a tool that is widely used in opinion mining, and is based on an English lexical dictionary called WordNet [52]. This method groups adjectives, nouns, verbs and other grammatical classes into synonym sets called synsets using a semi-supervised learning step. SentiWordNet associates three scores with synset from the WordNet dictionary to indicate the sentiment of the text: positive, negative, and objective (neutral). The scores, which are in the values of [0, 1] and add up to 1, are obtained using a semi-supervised machine learning method. For example, suppose that a given synset $s = [bad, wicked, terrible]$ has been extracted from a tweet. SentiWordNet then will give scores of 0.0 for positive, 0.850 for negative, and 0.150 for objective sentiments, respectively. In this work, we will consider scores less than zero as negative, equals to zero as neutral and greater than zero as positive

  In this work, we used SentiWordNet version 3.0, which is available at `http://sentiwordnet.isti.cnr.it/`. To assign polarity based on this method, we considered the average scores of all associated synsets of a given text and consider it to be positive, if the average score of the positive affect is greater than that of the negative affect. Scores from objective sentiment were not used in determining polarity.

- **Sentiment140 Lexicon:**

  Sentiment140 Lexicon [57] is a dictionary of words with associations to positive and negative sentiments. The dictionary of Sentiment140 Lexicon consists of up to 66,000 unigrams (single words), 677,000 bigrams (two-word sequence) and 480,000 of unigram–unigram pair, unigram–bigram pair, bigram–unigram pair, or a bigram–bigram pair and was built using a SVM classifier that analyzed features such as number and categories of emoticons and sum of the sentiment scores for all tokens (calculated with lexicons). This combinations were extracted from tweets from Stanford Twitter Corpus [29]. In this work, each message classified will receive label 1 if positive, -1 if negative and 0 if neutral (in the case that Sentiment140 Lexicon could not find any word of the dictionary associated in the message).

  We used the Sentiment140 Lexicon version 0.1, available at `http://www.saifmohammad.com/WebPages/ResearchInterests.html`.

- **SentiStrength:**

  The most comprehensive work [94] consists of a lexicon dictionaty with labels annotated by humans and improved with the use of many machine learning strategies, including simple logistic regression, SVM, J48 classification tree, JRip rule-based classifier, SVM regression, AdaBoost, Decision Table, Multilayer Perception, and Naïve Bayes. The core classification of this work relies on the set of words in the LIWC dictionary [93], and the authors expanded this baseline by adding new features for the OSN context. The features added include a list of negative and positive words, a list of booster words to strengthen (e.g., "very") or weaken (e.g., "somewhat") sentiments, a list of emoticons, and the use of repeated punctuation (e.g., "Cool!!!!") to strengthen sentiments. For evaluation, the authors used labeled text messages from six different Web 2.0 sources, including MySpace, Twitter, Digg, BBC Forum, Runners World Forum, and YouTube Comments.

  SentiStrength classify positive (from 1 to 5) and negative (from -1 to -5) sentiment strength separately as the default setup of the method, used unless binary (positive/negative), trinary (positive/negative/neutral) or scale (-4 to +4) is set. Since we would like to evaluate methods including in neutral messages, in this work we will consider the trinary classification. This mode receive a message as input and outputs three values corresponding to the positivity, negativity and neutral score. For example, for the message "I love you" the result in the trinary mode would be 3 -1 1, this is: (+ve classification) (-ve classification) (trinary classification). So, the trinary classification is the final polarity of that instance.

  In this work, we used SentiStrength version 2.0, which is available at `http://sentistrength.wlv.ac.uk/Download`.

- **OpinionFinder:**

  OpinionFinder is a system that performs subjectivity analysis, automatically identifying when opinions, sentiments, speculations, and other private states are present in text [103]. The tool is considered as a hybrid approach since it performs subjectivity analysis trough a framework with lexical analysis former and a machine learning approach latter. The subjective analysis of OpinionFinder has four components: (i) Naïve Bayes classifier that distinguishes between subjective and objective sentence; (ii)Identification of speech events (e.g., "said", "according to") and direct subjective expressions (e.g., "fears", "is happy"'); (iii) Opinion source identification (the source of a speech event is the speaker; the source of a subjective expression is the experience of the private state) using MPQA Opinion Corpus [19] as features source to training; and (iv) Sentiment expression classification. The last component, consists of two classifiers to identify words with positive or negative sentiments trained with BoosTexter [83] and MPQA Opinion Corpus. The first classifier focuses on identifying sentiment expressions and the second classifier takes the sentiment expressions and identifies those that are positive and negative.

In this work, we used OpinionFinder version 2.0, which is available at `http://mpqa.cs.pitt.edu/opinionfinder/opinionfinder_2`.

- **SANN:**

  The fifth and last hybrid method considered in this study is called Sentiment-aware Narest Neighbor Model (SANN). SANN was proposed by [70] with the purpose of infer additional user ratings by performing sentiment analysis (SA) of user comments and integrating its output in a nearest neighbor (NN) model. The classifier uses the MPQA polarity lexicon and can deal with negation, intensifiers, and polarity shifters. SANN is considered as a hybrid method since it was built using dictionary-based methods and specifically on an extension of the rule-based unsupervised sentiment classifier proposed on a previous study [69]

Table 2.2 and 2.3 present an overview of previous discussed methods, providing a brief description of each one as well as their outputs (e.g. -1, 0, 1, meaning negative, neutral, and positive, respectively), the datasets they used to validate and finally, the baseline methods used for comparison. The methods are organized in chronological order to allow a better overview of the existing efforts along the years. We can note that the methods generate different outputs formats. We colored as blue the positive outputs, as black the neutral ones, and as red those that are negative.

In the next chapter, we present dataset considered in this study for our experimental analysis.

[19] The MPQA Opinion Corpus is available at http://nrrc.mitre.org/NRRC/publications.htm

Table 2.2. Overview of the sentence-level methods available in the literature (table continues).

Nome	Description	Output	Validation	Compared To
Emoticons	Messages containing pos/neg emoticons are pos/neg. Messages without emoticons are not classified.	**-1, 1**	-	-
Opinion Lexicon [37]	Focus on product reviews. Built a lexicon to predict polarity of product features phrases that are summarized to provide an overall score to it.	**Negative, Positive**	Product reviews from Amazon and CNet	-
Opinion Finder (MPQA) [104] [105]	Performs subjectivity analysis trough a framework with lexical analysis former and a machine learning approach latter.	**Negative, Neutral, Positive**	MPQA [102]	Compared to itself in different versions.
Happiness Index [25]	Quantifies happiness levels for large-scale texts like lyrics and blogs. Uses ANEW [11] to rank the documents.	**1, 2, 3, 4, 5, 6, 7, 8, 9**	Lyrics, blogs, STUmessages [20], British National Corpus [21]	-
SentiWordNet [26] [5]	Construction of a lexical resource based on WordNet [52]. Authors grouped adjectives, nouns, etc in synonym sets (synsets) and associated polarity scores (positive, negative and neutral) for each one.	**[-1..0), 0, (0..1]**	-	General Inquirer (GI)[88]
LIWC [93]	Commercial tool to evaluate emotional, cognitive, and structural components of a given text.	**negEmo, posEmo**	-	-
SenticNet [14]	Uses dimensionality reduction to infer the polarity of common sense concepts and hence provide a public resource for mining opinions from natural language text at a semantic, rather than just syntactic level.	**[-1..0), 0, (0..1]**	Patient opinions	SentiStrength [94]
AFINN [64]	Twitter based sentiment lexicon that includes internet slangs and obscene words.	**[-5..) ,-1..1, (..5]**	Twitter [7]	OpinonFinder [104], ANEW [11], GI [88] and Sentistrength [94]
SO-CAL [92]	Creates lexicon with unigrams and multi-grams hand ranked with scale +5 (strongly positive) to -5 (strongly negative). Includes part of speech processing, negation and intensifiers.	**[<0), 0, (>0]**	Epinion [91], MPQA[102], Myspace[94],	MPQA[102], GI[88], SentiWordNet [26],"Maryland" Dict. [55], Google Generated Dict. [91]
Emoticons DS (Distant Supervision)[34]	Creates a scored lexicon based on a large dataset of tweets. Based on the frequency each term occurrence with positive or negative emotions.	**-1, 1**	Unlabeled Twitter data [15]	-
NRC Hashtag [54]	Builds a lexicon dictionary using a Distant Supervised. Used hashtag to classify tweet (i.e #joy, #sadness, etc). Then, it verifies the occurrence of each specific n-gram in that emotion.	**sadness, anger, fear, disgust, anticipation, surprise, joy, trust**	Twitter (SemEval-2007 Affective Text Corpus) [89]	-
Pattern.en [22]	Python Programming Package (toolkit) to deal with NLP, web mining and Sentiment Analysis.	**[-1..0), 0.1, (0.1..1]**	Product reviews, but the source was not specified	-

Table 2.3. Overview of the sentence-level methods available in the literature.

Nome	Description	Output	Validation	Compared To
SASA [99]	Based on the statistical model obtained from the classifier Naïve Bayes on unigram features. It also explores emoticons and exclamations.	**Negative, Neutral, Unsure, Positive**	Political tweets labeled with AMT	-
PANAS-t [32]	Adapted version (PANAS) Positive Affect Negative Affect Scale [100], well-known method in psychology with a large set of words associated with eleven moods.	**fear, sadness, guilt, hostility, shyness, fatigue, attentiveness, joviality, assurance, serenity, surprise**	Unlabeled global events data from Twitter [15]	-
EmoLex [56]	General sentiment lexicon crowdsourcing supported. Each entry lists the association of a token with 8 basic sentiments defined by [72]. Includes unigrams and bigrams from Macquarie Thesaurus, General Iquirer and WordNet.	**sadness, anger, fear, disgust, anticipation, surprise, joy, trust**	-	Compared with existing gold standard data but it was not specified
SANN [70]	Infer additional reviews user ratings by performing sentiment analysis of user comments and integrating its output in a Nearest Neighbor model.	**neg, neu, pos**	TED Talks	Comparison with other multimedia recommendation approaches.
Sentiment140 Lexicon [57]	Creation of a lexicon dictionary in a similar way to [54] and a SVM Classifier with features like: number and categories of emoticons, sum of the sentiment scores for all tokens (calculated with lexicons), etc.	**Negative, Neutral, Positive**	Twitter and SMS from Semeval 2013-task 2 [61].	Other Semeval 2013-task 2 approaches
SentiStrength [94]	Lexicon dictionary annotated by humans and improved with the use of Machine Learning.	**[-5..) ,-1..1, (..5]**	Twitter, Youtube, Digg, Myspace, BBC Forums and Runners World.	The best of nine Machine Learning techniques for each test.
Stanford Recursive Deep Model [86]	Proposes a model called Recursive Neural Tensor Network that processes all sentences dealing with their structures and compute the interactions among them.	**very negative, negative, neutral, positive, very positive**	Movie Reviews [66]	Naïve Bayes and SVM's with bag of words features and bag of bigram features.
Umigon [44]	Disambiguated tweets using lexicon and heuristics.	**Negative, Neutral, Positive**	Twitter and SMS from Semeval 2013-task 2 [61].	[57]
VADER [38]	Human-validated sentiment analysis method developed for Twitter and social media contexts. Created from a generalizable, valence-based, human-curated gold standard sentiment lexicon.	**-1, 0, 1**	Twitter, Movie Reviews, Technical Product Reviews, NYT User's Opinions.	(GI)[88], LIWC, [93], SentiWordNet [26], ANEW [11], SenticNet [14] and some Machine Learning Approaches.

Chapter 3

Datasets

To make the comparison among methods possible, we considered several datasets of many domains from Web. In this study, we employed labeled and unlabeled dataset, which will be described next.

3.1 Unlabeled data: Near-complete Twitter logs

The first set of dataset is a near-complete log of Twitter messages posted by all users from March 2006 to August 2009 [15]. This dataset contains 54 million users who had 1.9 billion follow links among themselves and posted 1.7 billion tweets over the course of 3.5 years. This dataset is appropriate for the purpose of this work, as it contains all users who set their account publicly available (excluding those users who set their accounts private) and their tweets, which is not based on sampling and hence alleviates any sampling bias. Additionally, this dataset allows us to study the reactions to noteworthy past events and evaluate our methods on data from real scenarios.

We chose six events covered by Twitter users[1]. These events, summarized in Table 3.1, span topics related to tragedies, product and movie releases, politics, health and sports events. To extract tweets relevant to these events, we first identified the sets of keywords describing the topics by consulting news websites, blogs, Wikipedia, and informed individuals. Given our selected list of keywords, we identified the topics by searching for keywords in the tweet dataset. This process is very similar to the way in which mining and monitoring tools to crawl data about specific topics.

We limited the duration of each event because popular keywords are typically hijacked by spammers after a certain amount of time. Table 3.1 displays the keywords

[1]Top Twitter trends at `http://tinyurl.com/yb4965e`

Table 3.1. Summary information of the six major topics events

Topic	Period	Keywords	#Messages
AirFrance	06.01–06.2009	victims, passengers, a330, 447, crash, airplane, airfrance.	10,000
2008US-Elect	11.02–06.2008	voting, vote, candidate, campaign, mccain, democrat*, republican*, obama, bush.	10,000
2008Olympics	08.06–26.2008	olympics, medal*, china, beijing, sports, peking, sponsor.	10,000
Susan Boyle	04.11–16.2009	susan boyle, I dreamed a dream, britain's got talent, les miserables.	10,000
H1N1	06.09–26.2009	outbreak, virus, influenza, pandemi*, h1n1, swine, world health organization.	10,000
Harry-Potter	07.13–17.2009	harry potter, half-blood prince, rowling.	10,000

used and the total number of tweets used in this study for each topic. The first column contains a short name for the event, which we use to refer to them in the rest of the paper. While the table does not show the ground truth sentiment of the six events, we can utilize these events to compare the predicted sentiments across different methods.

3.2 Labeled data: Multi-domain logs

The second set of datasets consists of sets of messages labeled as positive, negative or neutral (some datasets does not include this polarity), with a total of 21 labeled subsets. Yelp is a business review service where users give ratings and write reviews about businesses and services. These information help other Yelp users to evaluate a business or a service and make a choice. From the Yelp Challenge Dataset, available in [109], we filtered five thousand reviews for these businesses from the greater Phoenix, AZ metropolitan area. Since each review comes with a star rating given by users in the moment they evaluate some place, we could use this score to infer the sentiment of that review. Thus, we would be able to label the reviews in negative (1 star) or positive (5 stars). For example, for the review “I really enjoy this place, they have the best hamburger in the world!” was given a 5 star rating, so we considered it as a positive message.

The Stanford Twitter Corpus is a labeled dataset of tweets collected in [29]. Authors labeled a set of 177 negative tweets and 182 positive tweets extracted from the Twitter API. Tweets was collected searching for specific queries such as companies (AIG, AT&T), people (Bobby Flay, Warren Buffet) and consumer products (Kindle2, iPhone, etc.).

The third dataset is six sets of messages labeled as positive and negative by humans, and was made available in the SentiStrength research [94]. This dataset include a wide range of social web texts from: MySpace, Twitter, Digg, BBC forum, Runners World forum, and YouTube comments. Each line of this dataset consists of a message and its positive and negative score. In order to have a single score that summarizes both, we considered the message as positive if its positive score is higher than the negative score, negative if its negative score is higher than the positive score, and neutral if the scores are equal.

The fourth dataset consists of sentiment judgment from the first 2008 U.S. Presidential debate collected from Twitter by [23]. Authors labeled all 3,238 tweets collected with Amazon Mechanical Turk [2] as positive, negative, mixed (tweets included those that contained both positive and negative components) and other (a category included to catch non-evaluative statements or questions). For the purpose of this work, we filtered 750 positive and 750 negative tweets.

The fifth dataset is a set of movie reviews of different categories written before 2002 collected by [66]. All reviews were labeled as positive or negative based on the number of stars or some numerical value that indicates the acceptance rate of the movie.

The sixth dataset, collected by [38] and used by authors to validate the VADER method, consists of labeled messages from Twitter's public timeline, sentence-level snippets from New York Times opinion news editorials/articles, snippets of movie reviews from Rotten Tomatoes [3] and customer reviews about different products on Amazon.

Another labeled dataset considered in this study consists of comments from TED Talks [70]. TED [4] is a popular online repository of public talks and user-contributed material. The next four set of datasets consist of random ([2, 62]) and specific topics' ([82]) tweets, and also tweets collected by the SemEval 2013 Task-2 [61] posted on the online social network.

Finally, from our near-complete Twitter logs, we also built a "tricky dataset" consisting of messages containing sarcastic and ironic content. This dataset consists of 150 tweets with the hashtag "#sarcasm" and 150 tweets with the hashtag "#irony" filtered from [15]. All tweets were manually inspected in order to filter only those with positive words contrasted in a negative situation.

Tables 3.2 and 3.3 summarize the main characteristics of 21 datasets such as number of messages, the average number of words found in all messages in each dataset. It also defines a simpler nomenclature that will be 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. Two datasets, Reviews_I and YELP, rely on five stars rates, in which users

[2] www.mturk.com
[3] www.rotten.tomatoes.com
[4] `http://ted.com`

rate and provide a comment about a content (e.g. a movie or an establishment).

Amazon Mechanical Turk Labeling (AMT) was used in seven out of 21 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 the correct polarity assigned to it. The number of annotators used to build the datasets is also shown in tables. Tweets_DBT was the unique dataset that was built with the use of AMT Labeling plus Expert validation. 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. When we crawled these tweets we could not access a small part of them as they were deleted. To avoid these sharing problems, 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 similar strategy used by Tweets_DBT. Our goal is not to redo all the human evaluation these efforts already did, but simple to inspect a small sample of them to infer our level of agreement with our gold standard data. We random select 1% of all sentences to be evaluated by experts (collaborators of this study) as an attempt to asses if these gold standard data are really trustful. It is important to mention that we do not have access to the instructions provided by the authors and a small amount of the data could not be evaluated and were discarded. For example, this manual inspection unveiled a few sentences in other idioms different than English, in the Tweets_STA and TED datasets, which were discarded. We also attempted to identify the messages that were suspect to be in different languages in the rest of the datasets. Then we manually inspected the suspected ones and removed those that are not in English.

Column R from the table exhibits the agreement of each dataset in our own evaluation. After a close look in the cases we disagree with the evaluations in the Gold standard, we understand that other interpretations could be given to the text, finding cases of sentences with mixed polarity. Some of then are strongly linked to context and 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, messages in languages other than English, or even messages with specific contexts (e.g.: Tweets_DBT). Thus, we considered these cases as well as the amount of disagreement we had with the gold standard data as reasonable and expected. Since the datasets Irony and Sarcasm were built by us, they

Table 3.2. Labeled datasets (table continuous).

Dataset	Nomeclature	# Msgs	# Pos	# Neg	# Neu	Average # of phrases	Average # of words	Annotat. Expertise	# of Annotat.	R (%)
Comments (BBC) [94]	**Comments_BBC**	1,000	99	653	248	3,98	64,39	Non Expert	3	87
Comments (Digg) [94]	**Comments_Digg**	1,077	210	572	295	2,50	33,97	Non Expert	3	88
Comments (NYT) [38]	**Comments_NYT**	5,190	2,204	2,742	244	1,01	17,76	AMT	20	88
Comments (TED) [70]	**Comments_TED**	839	318	409	112	1	16,95	Non Expert	6	82
Comments (Youtube) [94]	**Comments_YTB**	3,407	1,665	767	975	1,78	17,68	Non Expert	3	90
Movie reviews [66]	**Reviews_I**	10,662	5,331	5,331	-	1,15	18,99	User Rating	-	66
Movie reviews [38]	**Reviews_II**	10,605	5,242	5,326	37	1,12	19,33	AMT	20	97
Myspace posts [94]	**Myspace**	1,041	702	132	207	2,22	21,12	Non Expert	3	91
Product reviews [38]	**Amazon**	3,708	2,128	1,482	98	1,03	16,59	AMT	20	94
Tweets (Political debate) [23]	**Tweets_DBT**	1,488	741	747	-	1	13,82	AMT + Expert	Undef.	60
Tweets (Irony) (Labeled by us)	**Irony**	100	38	43	19	1,01	17,44	Expert	3	-

were not evaluated in this table.

Finally, we included as part of our gold standard data two small datasets containing tweets with the hashtag #sarcasm and #irony. These tweets were obtained as a random sample from a one-year dataset obtained in 2014 that contains a sample of 1% of all tweets produced in that period. These datasets were then labeled by two of us, considered as experts in the topic. A third evaluator was used in cases of disagreement.

In the next section, we introduce the methodology of the work presented in this study.

Table 3.3. Labeled datasets.

Dataset	Nomeclature	# Msgs	# Pos	# Neg	# Neu	Average # of phrases	Average # of words	Annotat. Expertise	# of Annotat.	R (%)
Tweets (Sarcasm) (Labeled by us)	**Sarcasm**	100	38	38	24	1	15,55	Expert	3	-
Tweets (Random) [94]	**Tweets_RND_I**	4,242	1,340	949	1,953	1,77	15,81	Non Expert	3	88
Tweets (Random) [38]	**Tweets_RND_II**	4,200	2,897	1,299	4	1,87	14,10	AMT	20	97
Tweets (Random) [62]	**Tweets_RND_III**	3,771	739	488	2,536	1,54	14,32	AMT	3	90
Tweets (Random) [2]	**Tweets_RND_IV**	500	139	119	222	1,90	15,44	Expert	Undef.	90
Tweets (Specific domains w/ emot.) [29]	**Tweets_STF**	359	182	177	-	1,0	15,1	Non Expert	Undef.	97
Tweets (Specific topics) [82]	**Tweets_SAN**	3,737	580	654	2,503	1,60	15,03	Expert	1	97
Tweets (Semeval) (Task2) [61]	**Tweets_Semeval**	6,087	2,223	837	3,027	1,86	20,05	AMT	5	100
Runners World forum [94]	**RW**	1,041	702	132	207	2,22	21,12	Non Expert	3	86
Yelp Dataset [109]	**YLP**	5,000	2,500	2,500	-	1	131,44	User Rating	-	94

Chapter 4

Methodology

In this chapter, we present evaluation methodology for comparing and combining the 21 sentiment analysis methods.

4.1 Comparing Methods

Comparing methods in order to highlight its advantages, disadvantages and possible limitations is not a easy task since methods varies in many particulars. Therefore, we considered different metrics to analyze the prediction performance method, as illustrated by Figure

In this section, we describe measures to compare the performance of the 21 methods.

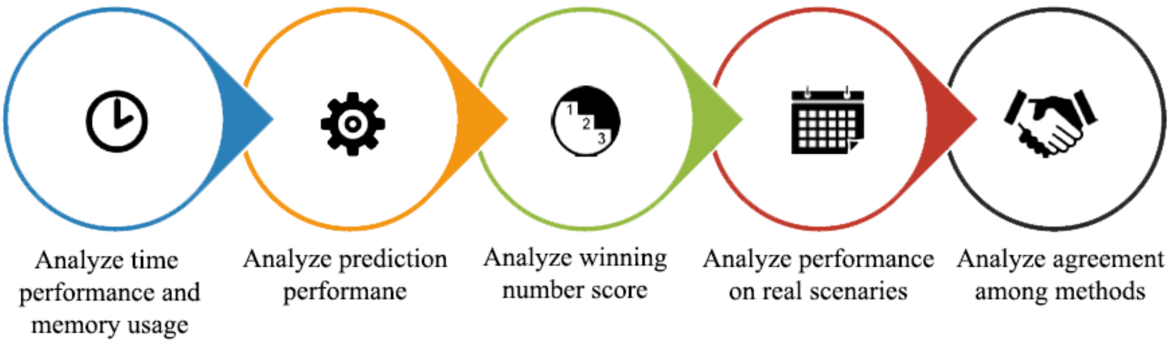


Figure 4.1. Methodology experiments illustrated by steps

4.1.1 Time Performance and Memory Usage

We would like to compare the time and memory usage performance of all methods in order to highlight their possible limitations when dealing with big datasets. This analysis is important since it can demonstrate, for example, which methods could be implemented in a mobile or in a real time applications, very needy environments nowadays.

To this analysis, we grouped unlabeled random tweets from our Twitter near-complete dataset in subsets of 10 thousand, 100 thousand, 1 million and 10 million sentences. The 21 methods were tested and compared among them in these subsets, allowing us to analyze the faster and slowly method, and also the method with less and high memory usage. All tests were executed on a Dell Desktop, with Intel(R) Xeon(R) Processor (2.53GHz) with 24 Cores, and 96 Gigabytes of RAM, in a Ubuntu version 12.04.3.

The results of this analysis will be presented in next chapters. Next, we introduce the prediction performance measure that will also be used to compare methods.

4.1.2 Prediction Performance

Considering the classification strategy when sentiment analysis results contain three classes, positive, neutral, and negative, we consider the following metrics:

		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 text instances which are actually in class X and predicted in class Y, where X;Y $\in$ positive; neutral; negative. The recall (R) of a class X is the ratio of the number of users correctly classified to the number of instances in class X. Precision (P) of a class X is the ratio of the number of instances classified correctly to the total predicted as instances of class X. For example, the precision of negative class is computed as:

$$P(neg) = i/(c + f + i) \tag{4.1}$$

Its recall as:

$$R(neg) = i/(g + h + i) \tag{4.2}$$

And its $F1$ measure is the harmonic mean between both precision and recall. In this case:

$$F1(neg) = \frac{2P(neg) \cdot R(neg)}{P(neg) + R(neg)} \tag{4.3}$$

We also compute the overall accuracy as:

$$Acc = \frac{a + e + i}{a + b + c + d + e + f + g + h + i} \tag{4.4}$$

It considers equally important the correct classification of each piece of text, independently of the class, and basically measures the capability of the method to predict the correct input. A variation of F1, namely, macro-F1, is normally reported to evaluate classification effectiveness when the classes are unbalanced. 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 results of this analysis will be presented in next chapters. Next, we introduce the winning number measure that will also be used to compare methods about their prediction performance.

4.1.3 Winning Number

As we have a large number of combination among base methods, baselines and datasets, a global analysis of the performance of all these combinations is not an easy task. For this, we resort to a performance measure proposed in [74], called *winning number*. This measure tries to assess the most competitive methods among a series of candidates, given a large series of pre-defined tasks they have to perform. That is, the *winning number* of a method i in the context of a performance measure M, is given as:

$$S_i(M) = \sum_{j=1}^{n} \sum_{k=1}^{n} \mathbf{1}_{M_i(j) > M_k(j)} \tag{4.5}$$

Where k is different from i, j is the dataset index (21 datasets) , i and k are the methods' index (21 methods), $M_i(j)$ is the performance of the $i-th$ method on $j-th$

dataset in terms of measure M, and $\mathbf{1}_{M_i(j)>M_k(j)}$ is the indicator function:

$$\mathbf{1}_{M_i(j)>M_k(j)} = \begin{cases} 1 & \text{if } M_i(j) > M_k(j), \\ 0 & \text{otherwise.} \end{cases} \tag{4.6}$$

Thus, the larger $S_i(M)$ is, the better the $i-th$ method performs compared to the others. In the next section, we introduce our initial efforts in combining all these 21 sentiment analysis methods.

The results of this analysis will be presented in next chapters. Next, we introduce the agreement measure that will also be used to compare methods.

4.1.4 Agreement Among Methods

In this study, we also would like to examine the degree to which different methods agree on the polarity of the content when they correctly classify polarity. For instance, when two or more methods detect sentiments in the same message (and this is the sentiment indicating by the ground truth) it is important to check whether these sentiments are the same. This analysis would strengthen the confidence in the polarity classification and can be done computing the intersections of polarity proportion given by each method.

In this analysis, each pair of method will be compared in relation to their output in the 21 labeled datasets and will receive a percentage that indicate what fraction of these sentences they agree. So, pair of methods with low percentage of agreement indicate that these methods did not match in the polarity classification in most of the sentences.

The results of this analysis will be presented in next chapters. Next, we introduce the polarity in global events measure that will also be used to compare methods.

4.1.5 Polarity in Global Events

Thus far, we have introduce measures of time performance, memory usage, prediction performance, winning number score and the agreement of the sentiment analysis methods. However, we would also like to analyze the 21 sentiment analysis on how polarity measured for each method varies across different global events filtered from our unlabeled Twitter dataset. With this analysis, we could show how methods behave in datasets related to real scenery, where no label is provided.

4.2 Combining Methods

Combining methods could be a important strategy for increase the prediction performances of sentiment analysis task because this task can group good qualities of many methods in one. This type of technique is already being used in other segments of Computer Science such as Search Engine [49, 79]. In this section, we aim at evaluate the viability of combining sentiment analysis methods with the final goal of maximize results of prediction performance.

An intuitive way to combine methods for sentiment analysis is to assign as the polarity of a message the most frequent polarity detected by all methods, this method could be called Majority Voting Method. More specifically, this combined method works as follows:

1. Execute all methods on a chosen dataset;
2. Take the result of each method for each message from the dataset;
3. Check the most frequent polarity given by all methods for each message;
4. Assign the most frequent polarity as the final polarity of this message.

As we can note, this approach consists of applying a majority voting algorithm considering the hypothesis that when most of methods agree in a polarity, this polarity should be the most likely to be the right one for a single message. In the remainder of this study, we named the Majority Voting Method as Combined I.

Chapter 5

Results and Discussions

In this chapter, we present comparison results for the 21 methods considered in this paper based on the 21 gold standard datasets considered. We highlight that comparing methods is a very complicated task, as these methods were developed with different goals. As most of the methods, we compare are public available in the Web or under request to the authors, they have been increasingly used as black box for any sort of task. We also present the results of prediction performance of the two combined methods proposed in this work.

5.1 Comparison Results

In order to understand the advantages, disadvantages, and limitations of the various sentiment analysis methods, we present comparison results among them. Next, we describe the results of the comparison in terms of the previous discussed metrics> time and memory usage performance, prediction performance, winning number, agreement and polarity in global events.

5.1.1 Time and Memory Usage Performance

In this section, we begin investigating which of the 21 methods have its execution time and memory usage performance affected when the volume of data input increases. As we said in previous chapters, this analysis is important due the increasing need to develop applications for mobile or Internet of Things (IoT) systems, which requires low memory usage and fast executions. As said in previous chapters, all tests were executed on a Dell Desktop, with Intel(R) Xeon(R) Processor (2.53GHz) with 24 Cores, and 96 Gigabytes of RAM, in a Ubuntu version 12.04.3. Table 5.1 and 5.2 shows the execution

Table 5.1. Execution time of all methods in files with increasing number of messages

	Size of the input file			
Method	**10,000**	**100,000**	**1,000,000**	**10,000,000**
SANN	400.0000	-	-	-
SO-CAL	31.6333	-	-	-
Stanford Deep Model	8.5532	96.0000	-	-
SentiStrength	0.5375	0.5377	0.5370	0.5372
Umigon	0.2121	14.2167	-	-
SentiWordNet	0.0517	0.5860	7.3077	35.3408
PANAS-t	0.0194	0.0030	0.0262	0.3095
SenticNet	0.0102	0.1094	1.0728	10.3909
SASA	0.0009	0.0223	0.5185	5.3235
VADER	0.0006	0.0053	0.0399	0.4699
OpinionFinder	0.0006	0.0053	-	-
Pattern.en	0.0006	0.0053	6.8000	33.6667
Emoticon DS	0.0006	0.0053	0.0399	0.4699
AFINN	0.0006	0.0053	0.0399	0.4699
Emoticons	0.0003	0.0020	0.0192	0.1757
NRC Hashtag	0.0001	0.0007	0.0047	0.0414
EmoLex	0.0001	0.0004	0.0025	0.0212
Sentiment140	0.0001	0.0030	0.0022	0.0192
OpinionLexicon	0.0000	0.0002	0.0017	0.0166
LIWC	0.0000	0.0031	0.0231	0.0000
Happiness Index	0.0000	0.0003	0.0026	0.0224

time and memory usage the datasets of 10 thousand, 100 thousand, 1 million and 10 million sentences.

We note that methods have varying degrees of execution time and memory usage performance. The method SANN and SO-CAL was not able to finish the execution in datasets bigger than 10 thousand sentences in time for this work. As well as Umigon and OpinionFinder that would not able to finish the execution on time in files bigger than 100 thousand sentences. Most methods have a constant memory usage, however, some methods achieved a prohibitive memory usage. It is the case of OpinionFinder, that use almost 24Gb of memory to execute a file with 100 thousand sentences.

These results is crucial for the efficacy of each method, because it can limit the number of messages that can be computed. The execution time varied from one method to another. In the context of Twitter, recent statistics showed that a person tweets an average of 1.85 tweets per day in the social networks. Thus, 10,000 tweets might be a good representation of the content for this domain.

Next we present a comparative performance evaluation of each method in terms

Table 5.2. Memory usage of all methods in files with increasing number of messages

	Size of the input file			
Method	**10.000**	**100.000**	**1.000.000**	**10.000.000**
OpinionFinder	9.6780	23.5080	-	-
Umigon	7.4750	8.9560	-	-
SANN	0.4000	-	-	-
SASA	0.2305	0.2507	0.2507	0.2507
Pattern.en	0.2305	0.2507	0.2507	0.2507
Stanford Deep Model	0.1359	0.1359	-	-
SentiWordNet	0.1007	0.1317	0.1317	0.1317
SenticNet	0.0597	0.0597	0.0597	0.0597
LIWC	0.0500	0.0500	0.2040	0.0000
VADER	0.0359	0.0359	0.0359	0.0359
EmoLex	0.0190	0.0190	0.0190	0.0190
Sentiment140	0.0128	0.0129	0.0128	0.0128
NRC Hashtag	0.0090	0.0090	0.0090	0.0090
Happiness Index	0.0061	0.0061	0.0061	0.0061
SentiStrength	0.0057	0.0057	0.0057	0.0057
SO-CAL	0.0046	-	-	-
Emoticon DS	0.0046	0.0046	0.0046	0.0046
AFINN	0.0046	0.0046	0.0046	0.0046
PANAS-t	0.0042	0.0042	0.0042	0.0042
Emoticons	0.0028	0.0028	0.0028	0.0029
OpinionLexicon	0.0024	0.0024	0.0024	0.0024

of correctly predicting polarity for the 21 methods.

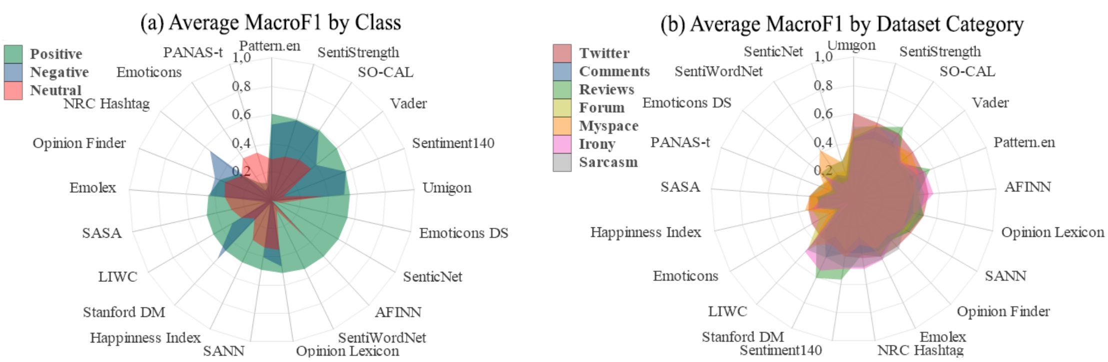


Figure 5.1. Average Macro-F1 by class of all methods

5.1.2 Prediction Performance

We start the analysis of our experiments by comparing the results of all metrics discussed previously for all labeled datasets. First, we note that existing methods varied widely in their prediction. This suggests that the same social media text could be interpreted very differently depending on the choice of a sentiment method. A few methods obtain worst results than a random method (i.e. a method that would randomly chooses among positive, neutral, or negative as output). This usually happened when a method is biased towards one or more classes. As example, emoticons showed to be a good method for detecting positive and negative messages when the input data has an emoticon. However, it considers most of the instances as neutral, leading to a bad performance for most of the datasets. However, we note that this bias can be used to construct ensemble approaches. For example, when emoticons position itself towards a positive or negative classification, it should be highly considered as it is usually correct. This can clearly be extended to other methods which showed a similar kind of bias.

We start the analysis of our experiments by comparing the results of all metrics discussed previously for all datasets. We selected some of these results to help us summarize our main findings (we point the reader to the complete material with the results of all methods in the 21 labeled datasets, available on Appendix A). Table 5.3 and 5.4 show the results of accuracy, precision, recall and F1 by class and general macroF1 for four of all the datasets. These tables also present the results of the combined method that will be described in next chapters.

First of all, is important to highlight that some datasets do not have neutral messages. In this case, the calculation of the MacroF1 metrics will consider the only the two remaining classes, positive and negative. We begin the analysis observing that in terms of accuracy and Macro-F1, there is no single method that always achieves the best prediction performance for different datasets, which is similar to the well-known "no-free lunch theorems" [106]. This suggests that at least a preliminary investigation should be performed when sentiment analysis is used in a new dataset in order to guarantee a reasonable prediction performance. As example, Pattern.en works well for Tweets_STF, appearing among the top 3 methods, but it presented poor prediction performance for Comments_YTB.

In a second finding, showed in Figure 5.1, we can see that most methods are more accurate in correctly classifying positive than negative text, suggesting that methods can lead to bias in their analysis towards positivity. Neutral showed to be even harder to be detected by most of them. Recent efforts show that human language have a

Table 5.3. Prediction performance of all methods in Comments_YTB and Tweets_STF datasets, including the combined method

Dataset	Method	Acc.	Posit. sentiment			Negat. sentiment			Neut. sentiment			MacroF1
			P	R	F1	P	R	F1	P	R	F1	
Comments_YTB	AFINN	0.54	0.69	0.61	0.64	0.59	0.34	0.43	0.38	0.58	0.46	0.51
	Emolex	0.43	0.64	0.37	0.47	0.47	0.34	0.40	0.32	0.61	0.42	0.43
	Emotic.	0.33	0.75	0.11	0.19	0.37	0.02	0.04	0.29	0.94	0.45	0.23
	Emot. DS	0.48	0.49	0.93	0.64	0.68	0.02	0.05	0.28	0.07	0.11	0.27
	H. Index	0.43	0.51	0.58	0.55	0.00	0.00	0.00	0.33	0.51	0.40	0.32
	LIWC	0.41	0.53	0.53	0.53	0.17	0.27	0.21	0.40	0.31	0.35	0.36
	NRC H.	0.37	0.70	0.21	0.33	0.34	0.72	0.46	0.27	0.35	0.30	0.36
	Op.Finder	0.42	0.70	0.31	0.43	0.42	0.32	0.36	0.33	0.70	0.45	0.41
	Opin. Lex.	0.48	0.69	0.46	0.55	0.54	0.36	0.43	0.34	0.62	0.44	0.47
	PANAS-t	0.31	0.70	0.05	0.09	0.49	0.04	0.08	0.29	0.96	0.45	0.20
	Pattern.en	0.58	0.71	0.73	0.72	0.48	0.48	0.48	0.42	0.39	0.40	0.53
	SANN	0.49	0.67	0.52	0.59	0.48	0.29	0.36	0.36	0.62	0.46	0.47
	SASA	0.47	0.52	0.73	0.60	0.00	0.00	0.00	0.36	0.39	0.37	0.33
	SO-CAL	0.57	0.74	0.62	0.68	0.54	0.52	0.53	0.40	0.53	0.46	0.55
	SWN	0.47	0.49	0.89	0.63	0.00	0.00	0.00	0.31	0.12	0.17	0.27
	S.Strength	0.61	0.75	0.75	0.75	0.49	0.61	0.54	0.47	0.38	0.42	0.57
	SenticNet	0.49	0.49	1.00	0.66	0.00	0.00	0.00	0.00	0.00	0.00	0.22
	Sentim.140	0.47	0.59	0.65	0.62	0.35	0.59	0.44	0.25	0.07	0.11	0.39
	Stanf. DM	0.47	0.82	0.47	0.60	0.33	0.72	0.45	0.35	0.27	0.30	0.45
	Umigon	0.57	0.79	0.62	0.70	0.44	0.51	0.47	0.43	0.54	0.48	0.55
	Vader	0.56	0.78	0.59	0.67	0.68	0.30	0.41	0.39	0.72	0.51	0.53
	Combined I	0.58	0.64	0.76	0.70	0.47	0.56	0.51	0.57	0.40	0.47	0.56
Tweets_STF	AFINN	0.53	0.76	0.62	0.68	0.88	0.45	0.59	0.00	0.00	0.00	0.64
	Emolex	0.37	0.70	0.41	0.51	0.82	0.33	0.47	0.00	0.00	0.00	0.49
	Emotic.	0.11	0.83	0.14	0.24	0.94	0.09	0.16	0.00	0.00	0.00	0.20
	Emot. DS	0.52	0.53	1.00	0.69	1.00	0.03	0.05	0.00	0.00	0.00	0.37
	H. Index	0.42	0.59	0.55	0.57	0.89	0.27	0.42	0.00	0.00	0.00	0.50
	LIWC	0.46	0.57	0.64	0.60	0.34	0.65	0.45	0.00	0.00	0.00	0.53
	NRC H.	0.50	0.81	0.24	0.37	0.76	0.77	0.77	0.00	0.00	0.00	0.57
	Op.Finder	0.35	0.81	0.31	0.45	0.80	0.40	0.53	0.00	0.00	0.00	0.49
	Opin. Lex.	0.46	0.77	0.50	0.61	0.93	0.42	0.58	0.00	0.00	0.00	0.60
	PANAS-t	0.07	0.80	0.07	0.12	0.86	0.07	0.13	0.00	0.00	0.00	0.13
	Pattern.en	0.67	0.76	0.75	0.75	0.81	0.58	0.67	0.00	0.00	0.00	0.71
	SANN	0.43	0.69	0.47	0.56	0.78	0.39	0.52	0.00	0.00	0.00	0.54
	SASA	0.30	0.50	0.60	0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.28
	SO-CAL	0.67	0.83	0.69	0.75	0.93	0.66	0.77	0.00	0.00	0.00	0.76
	SWN	0.49	0.51	0.97	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.34
	S.Strength	0.69	0.82	0.68	0.74	0.84	0.69	0.76	0.00	0.00	0.00	0.75
	SenticNet	0.51	0.51	1.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.34
	Sentim.140	0.75	0.77	0.71	0.74	0.75	0.78	0.76	0.00	0.00	0.00	0.75
	Stanf. DM	0.60	0.88	0.31	0.46	0.61	0.89	0.73	0.00	0.00	0.00	0.60
	Umigon	0.71	0.92	0.67	0.77	0.83	0.75	0.79	0.00	0.00	0.00	0.78
	Vader	0.45	0.88	0.54	0.67	0.91	0.36	0.51	0.00	0.00	0.00	0.59
	Combined I	0.60	0.63	0.86	0.73	0.56	0.93	0.70	0.00	0.00	0.00	0.72

universal positivity bias [24]. So, part of the bias we observe for sentiment prediction might be related to characteristics of human language, which is intrinsic leveraged to some methods by the way they are constructed. For example, [34] developed a lexical resource in which positive and negative values are associated to words, hashtags, and any sort of tokens according to the frequency these tokens appear together with tweets containing positive and negative emoticons. As a consequence, this method showed to be biased towards positivity due to the larger amount of positivity in the data they used to build the lexicon resource. The overall poor performance of this specific method is credited to its lack of treatment to neutral messages and better performance mostly

Table 5.4. Prediction performance of all methods in Tweets_RND_III and Irony datasets, including the combined method

Dataset	**Method**	**Acc.**	**Posit. sentiment**			**Negat. sentiment**			**Neut. sentiment**			**MacroF1**
			P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	
	AFINN	0.64	0.41	0.65	0.50	0.49	0.48	0.48	0.81	0.67	0.73	0.57
	Emolex	0.64	0.38	0.41	0.40	0.43	0.41	0.42	0.76	0.75	0.75	0.52
	Emotic.	0.70	0.70	0.17	0.27	0.66	0.09	0.15	0.70	0.98	0.82	0.41
	Emot. DS	0.21	0.20	0.98	0.33	0.90	0.04	0.07	0.60	0.02	0.04	0.15
	H. Index	0.53	0.27	0.65	0.38	0.00	0.00	0.00	0.77	0.59	0.67	0.35
	LIWC	0.47	0.38	0.22	0.28	0.18	0.19	0.18	0.55	0.70	0.62	0.36
	NRC H.	0.51	0.39	0.30	0.34	0.25	0.80	0.39	0.78	0.52	0.62	0.45
	O.Finder	0.72	0.57	0.33	0.42	0.50	0.35	0.41	0.76	0.90	0.82	0.55
	Opin. Lex.	0.70	0.48	0.50	0.49	0.56	0.43	0.48	0.78	0.81	0.80	0.59
	PANAS-t	0.70	0.77	0.13	0.22	0.55	0.06	0.11	0.70	0.98	0.82	0.38
Tweets	Pattern.en	0.54	0.36	0.77	0.49	0.35	0.59	0.44	0.84	0.46	0.59	0.51
_RDN_III	SANN	0.67	0.43	0.49	0.46	0.46	0.36	0.40	0.78	0.78	0.78	0.55
	SASA	0.52	0.26	0.67	0.37	0.00	0.00	0.00	0.78	0.57	0.66	0.34
	SO-CAL	0.67	0.43	0.69	0.53	0.52	0.61	0.56	0.84	0.67	0.75	0.61
	SWN	0.32	0.24	0.71	0.36	0.23	0.49	0.32	0.75	0.17	0.28	0.32
	S.Strength	0.65	0.45	0.80	0.58	0.42	0.73	0.54	0.92	0.58	0.71	0.61
	SenticNet	0.20	0.20	1.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.11
	Sentim.140	0.29	0.24	0.72	0.36	0.28	0.76	0.41	0.81	0.07	0.13	0.30
	Stanf. DM	0.32	0.64	0.39	0.48	0.16	0.85	0.26	0.76	0.20	0.31	0.35
	Umigon	0.74	0.58	0.70	0.63	0.49	0.68	0.57	0.89	0.76	0.82	0.67
	Vader	0.73	0.54	0.65	0.59	0.68	0.41	0.51	0.81	0.82	0.81	0.64
	Combined I	0.77	0.67	0.60	0.63	0.55	0.69	0.61	0.85	0.84	0.84	0.70
	AFINN	0.56	0.65	0.59	0.62	0.86	0.42	0.56	0.35	0.88	0.50	0.56
	Emolex	0.47	0.53	0.45	0.49	0.86	0.42	0.56	0.24	0.63	0.35	0.47
	Emotic.	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.20	1.00	0.33	0.11
	Emot. DS	0.32	0.29	1.00	0.45	1.00	0.02	0.05	0.75	0.19	0.30	0.26
	H. Index	0.33	0.41	0.64	0.50	0.00	0.00	0.00	0.28	0.81	0.41	0.30
	LIWC	0.49	0.64	0.50	0.56	0.30	0.87	0.45	0.81	0.34	0.48	0.50
	NRC H.	0.59	0.44	0.18	0.26	0.71	0.84	0.77	0.38	0.50	0.43	0.49
	O.Finder	0.38	0.70	0.32	0.44	0.89	0.19	0.31	0.26	1.00	0.41	0.39
	Opin. Lex.	0.44	0.53	0.36	0.43	0.88	0.33	0.47	0.28	0.88	0.42	0.44
Irony	PANAS-t	0.21	0.00	0.00	0.00	0.50	0.02	0.04	0.20	1.00	0.34	0.13
	Pattern.en	0.53	0.63	0.77	0.69	0.76	0.30	0.43	0.35	0.81	0.49	0.54
	SANN	0.41	0.41	0.41	0.41	1.00	0.23	0.38	0.29	0.88	0.43	0.41
	SASA	0.25	0.31	0.55	0.39	0.00	0.00	0.00	0.19	0.50	0.28	0.22
	SO-CAL	0.56	0.59	0.59	0.59	0.83	0.47	0.60	0.34	0.75	0.47	0.55
	SWN	0.27	0.28	1.00	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.14
	S.Strength	0.56	0.53	0.45	0.49	0.65	0.60	0.63	0.41	0.56	0.47	0.53
	SenticNet	0.27	0.27	1.00	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.14
	Sentim.140	0.53	0.42	0.68	0.52	0.63	0.56	0.59	0.57	0.25	0.35	0.49
	Stanf.DM	0.63	0.77	0.45	0.57	0.64	0.84	0.73	0.42	0.31	0.36	0.55
	Umigon	0.42	0.53	0.41	0.46	0.64	0.33	0.43	0.26	0.69	0.38	0.42
	Vader	0.42	0.71	0.45	0.56	0.89	0.19	0.31	0.28	1.00	0.43	0.43
	Combined I	0.51	0.50	0.69	0.58	0.35	0.88	0.50	0.94	0.31	0.47	0.52

in Twitter related datasets. This behavior will be analyzed again in this work in the next sections.

Next, we present the results of winning number achieved for all sentiment analysis methods.

5.1.3 Winning Number

In this section we present the results of the winning number score achieved for all method in the labeled datasets. As discussed before, the winning number measure tries to assess the most competitive methods among a series of candidates, given a

Table 5.5. Winning Points Ranking for MacroF1 and Accuracy

Ranking	MacroF1 Winning score	Accuracy Winning score
SO-CAL	379	350
SentiStrength	369	351
Umigon	326	295
Pattern.en	322	309
Opinion Lexicon	301	287
Vader	290	263
Stanford DM	267	262
AFINN	260	241
Sentiment140	250	273
SANN	247	229
Emolex	230	213
Opinion Finder	213	214
NRC Hashtag	202	226
LIWC	196	195
SASA	156	61
SentiWordNet	149	202
SenticNet	115	108
Happinness Index	111	63
PANAS-t	109	143
Emoticons	104	137
Emoticons DS	101	183

large series of pre-defined tasks they have to perform. By Equation 4.5, the highest winning number that could be achieved by each method is 420. Table 5.5 present the results of winning score, in which we consider the performance metric MacroF1 and Accuracy.

As we can observe by Table 5.5, the top three methods in terms of winning numbers for MacroF1 are SO-CAL, SentiStrength and Umigon, and SentiStrength, SO-CAL and Pattern.en in terms of Accuracy. This means that these methods are in general good across datasets to correctly identify three classes: positive, neutral, and negative. This suggests that these methods would be preferable in situations in which any sort of preliminary evaluation can be performed. However, it is important to note that the overall unsupervised classification results are considerable low, leaving a still large space for the development of better techniques. We also note that methods are usually doing better in the datasets in which they were originally validated, which is expected as authors might attempt to identify points of improvement with the same dataset before releasing the study. This reinforces the need of our effort and even suggests that new gold standard dataset should be continuously created.

Next, we analyze the performance of 21 sentiment analysis methods in global events filtered from Twitter.

5.1.4 Polarity in Global Events

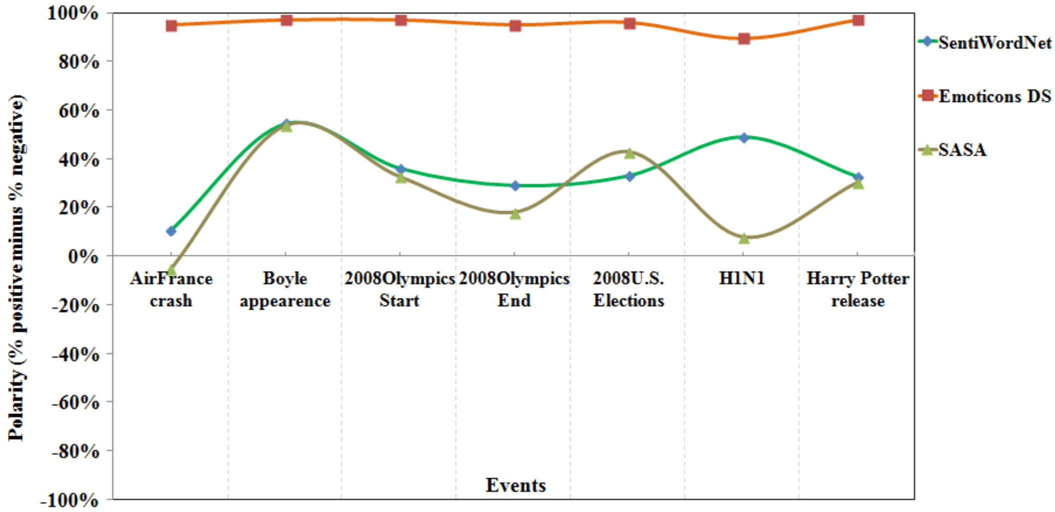
In previous sections, we started the discussion about the bias of positivity that may exists in the methods. In this section, we provide a second analysis on how polarity measured for each method varies across different global events filtered from our unlabeled Twitter dataset. As discussed before, these events, summarized in Table 3.1, span topics related to tragedies, movie releases, politics, health and sports events.

Figure 5.2 presents the polarity of each method when exposed to each dataset of a single event, grouping by methods that always give positive results independent of the nature of the event (5.2(a)), by methods that always give negative results independent of the nature of the event (5.2(b)), and by methods that achieved distinct degrees of polarity among events (5.2(c)). For each dataset and method, we computed the percentage of positive messages and the percentage of negative messages. The Y-axis shows the positive percentage minus the negative percentage.

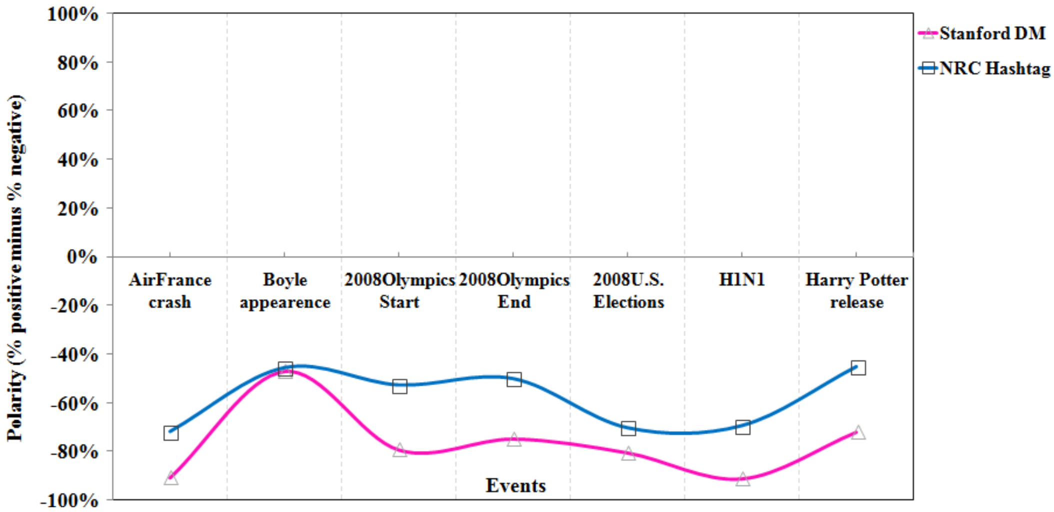
We can make several interesting observations. First, we clearly see that most methods present more positive values than the negative values, as we see few lines below 0% among all events. Second, we note that three methods obtained only positive values, even for events like H1N1 and AirFrance crash (SentiWordNet, Emoticons DS and SASA) (5.2(a)). While these event's data may contain jokes and positive tweets, it would be also reasonable to expect a large number of tweets expressing concerns and bad feelings. Similarly, Stanford DM and NRC Hashtag presented only negative values even in for events like Harry Potter release and the 2008 Olympics (5.2(b)).

This bias towards positive polarity showed by most of the methods might be trick for real time polarity detecting tools, as they might simply apply these methods in real time data, like Twitter streaming API, and account the rate of positive and negative message text. This would potentially show biased results due to the methods used.

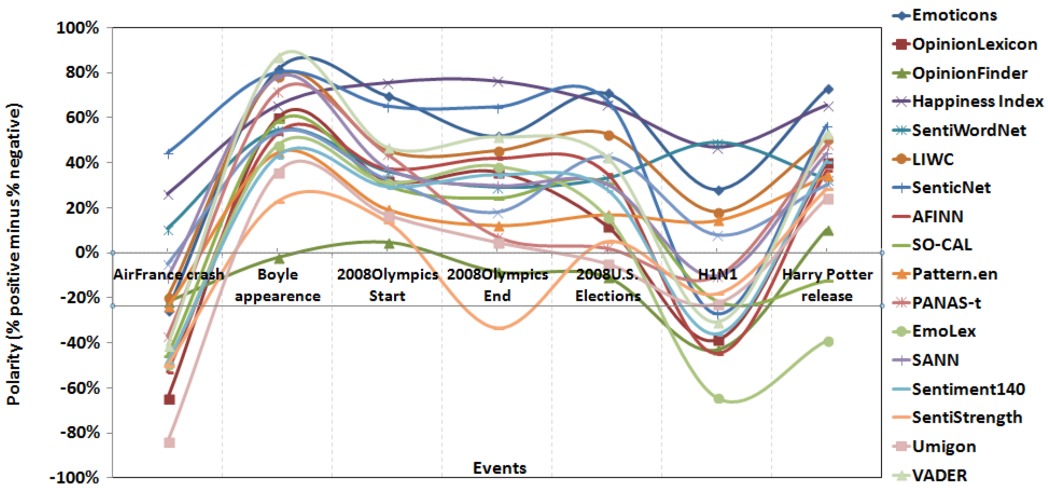
In the next section, we examine the degree to which different methods agree on the polarity of the content.



(a) Predominantly positive methods



(b) Predominantly negative methods



(c) Methods with various degrees of polarity

Figure 5.2. Polarity variation of all methods in global events filtered from the unlabeled Twitter dataset

Table 5.6. Prediction performance of the combined method on all labeled datasets.

Method	Dataset	Acc.	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
Comb. I	Tweets_RND_IV	0,62	0,60	0,64	0,62	0,52	0,63	0,57	0,69	0,61	0,65	0,61
	Tweets_DBT	0,31	0,00	0,00	0,00	0,30	0,99	0,46	0,80	0,03	0,05	0,17
	Tweets_RND_III	0,77	0,67	0,60	0,63	0,55	0,69	0,61	0,85	0,84	0,84	0,70
	Irony	0,51	0,50	0,69	0,58	0,35	0,88	0,50	0,94	0,31	0,47	0,52
	Comments_TED	0,46	0,54	0,68	0,61	0,36	0,70	0,48	0,58	0,17	0,27	0,45
	Reviews_I	0,50	0,64	0,68	0,66	0,37	0,76	0,50	0,00	0,00	0,00	0,58
	Sarcasm	0,57	0,67	0,56	0,61	0,42	0,80	0,55	0,67	0,44	0,53	0,57
	Comments_BBC	0,55	0,53	0,20	0,29	0,62	0,83	0,71	0,36	0,35	0,36	0,46
	Comments_Digg	0,52	0,47	0,41	0,44	0,48	0,79	0,60	0,62	0,37	0,47	0,50
	Myspace	0,59	0,61	0,88	0,72	0,42	0,44	0,43	0,66	0,31	0,42	0,52
	RW	0,52	0,70	0,64	0,67	0,50	0,43	0,46	0,29	0,38	0,33	0,49
	Tweets_RND_I	0,62	0,56	0,65	0,60	0,41	0,62	0,49	0,76	0,60	0,67	0,59
	Comments_YTB	0,58	0,64	0,76	0,70	0,47	0,56	0,51	0,57	0,40	0,47	0,56
	Tweets_STF	0,60	0,63	0,86	0,73	0,56	0,93	0,70	0,00	0,00	0,00	0,72
	Amazon	0,45	0,53	0,85	0,65	0,30	0,73	0,42	0,83	0,05	0,09	0,39
	Reviews_II	0,52	0,60	0,74	0,66	0,45	0,75	0,56	0,32	0,00	0,01	0,41
	Comments_NYT	0,37	0,32	0,72	0,44	0,37	0,82	0,51	0,86	0,07	0,13	0,36
	Tweets_RND_II	0,64	0,63	0,98	0,77	0,64	0,87	0,74	1,00	0,00	0,01	0,50
	YLP	0,84	0,95	0,81	0,87	0,73	0,95	0,82	0,00	0,00	0,00	0,85
	Tweets_SemEval	0,69	0,60	0,76	0,67	0,51	0,54	0,53	0,81	0,69	0,75	0,65
	Tweets_SAN	0,67	0,52	0,38	0,44	0,44	0,56	0,49	0,77	0,79	0,78	0,57

5.1.5 Agreement Among Methods

In this section we examine the degree to which different methods agree on the polarity of the content. As said before, this would strengthen the confidence in the polarity classification. In order to compute the agreement of each method, we calculated the intersections of polarity proportion given by each pair of method when they correctly classify polarity.

Some of these results is presented in Figure B.2 and B.3 (we point the reader to the complete material with the results of all methods in the 21 labeled datasets, available on Appendix B). These figures present the percentage of agreement among all methods on 2 of all labeled datasets, Tweets_RND_IV dataset and Tweets_DBT dataset. For each method in the first column, we measure, from the messages classified for each pair of methods, for what fraction of these messages they agree. Values below the diagonal (presented by a pair consisting of the same method) is the same as values above the diagonal. In order to make the visualization easier, we highlight the values sorting by the best percentage of agreement (darker cells) to the worst (lighter cells), indicating pairs of methods with more agreement percentage.

In order to summarize our findings in this analysis, we present Table 5.7 with the ranking of the three pairs of methods with highest percentage of agreement on all 21 labeled datasets. The first thing we can note is that Emoticons and PANAS-t methods appear to be the methods that have best agreement. The second thing that can be observed is that at least ten methods did not appeared in this table. This could

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtah	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	32	38	21	28	28	33	39	25	36	32	24	37	27	37	26	26	20	42	43	22
Emolex	32	100	38	16	28	26	34	38	30	29	30	25	36	22	32	21	21	15	36	36	29
Emotic.	38	38	100	24	32	37	39	42	39	55	38	35	43	30	42	27	37	27	62	51	36
Emotic. DS	21	16	24	100	19	9	13	17	5	29	16	16	20	23	25	25	25	12	29	25	2
Happ. Index	28	28	32	19	100	18	28	29	24	28	27	23	30	24	28	23	21	13	34	34	22
NRC Hashtag	28	26	37	9	18	100	26	29	23	29	22	22	29	18	27	15	25	21	36	29	21
Opin. Finder	33	34	39	13	28	26	100	37	35	28	35	26	36	20	32	19	18	16	37	38	32
Opin. Lexicon	39	38	42	17	29	29	37	100	33	33	33	27	40	24	35	23	24	17	41	42	31
PANAS	25	30	39	5	24	23	35	33	100	20	30	24	30	13	23	10	8	7	31	33	43
Pattern	36	29	55	29	28	29	28	33	20	100	29	28	38	32	41	34	41	31	57	43	15
SANN	32	30	38	16	27	22	35	33	30	29	100	23	33	22	31	20	20	16	37	37	27
SASA	24	25	35	16	23	22	26	27	24	28	23	100	28	20	27	18	20	15	35	31	22
SO-CAL	37	36	43	20	30	29	36	40	30	38	33	28	100	27	39	28	29	23	43	44	26
SWN	27	22	30	23	24	18	20	24	13	32	22	20	27	100	29	27	26	18	33	30	9,2
SentiStrength	37	32	42	25	28	27	32	35	23	41	31	27	39	29	100	29	32	23	45	42	19
SenticNet	26	21	27	25	23	15	19	23	9,6	34	20	18	28	27	29	100	27	20	33	29	5,4
Sentim.140	26	21	37	25	21	25	18	24	8	41	20	20	29	26	32	27	100	27	42	30	4
Stanford DM	20	15	27	12	13	21	16	17	7	31	16	15	23	18	23	20	27	100	30	23	4
Umigon	42	36	62	29	34	36	37	41	31	57	37	35	43	33	45	33	42	30	100	49	26
VADER	43	36	51	25	34	29	38	42	33	43	37	31	44	30	42	29	30	23	49	100	31
LIWC	22	29	36	2	22	21	32	31	43	15	27	22	26	9	19	5	4	4	26	31	100

Figure 5.3. Percentage of agreement among all methods in Tweets_RND_IV dataset.

mean that most pairs of methods do no agree in the polarity of sentences, implying that when analyzed with different tools, datasets could be interpreted very differently. In particular, for those methods that have lower than 50% agreement, the polarity will even change (e.g., from positive to negative, or negative for neutral, etc.). This results highlight that methods varies a lot about the polarity given by each sentences, probably because the variation of approach and techniques used by each one. This observation might lead us to combining sentiment analysis methods in order to group peculiarities of many methods aiming the achievement of better results of prediction performance.

After all analyzes and comparison among the 21 sentiment analysis methods, we present the table depicted in Figure 5.5. This figure present the ranking of all methods considering the average results of each them in the metrics considered in this study: execution time, memory usage, MacroF1 and accuracy, and winning number. In this overview we can confirm what was said before, that there is not a clear winner in all metrics. With this rank, we could easily choose which method to use in a hypothetical

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtah	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	28	23	11	25	21	27	34	24	20	27	20	29	18	28	19	16	18	29	32	8
Emolex	28	100	22	9	23	21	26	29	23	17	24	18	27	17	25	16	16	18	25	26	8
Emotic.	23	22	100	2	23	16	28	26	37	9	25	17	21	6	17	6	2	8	27	29	1
Emotic. DS	11	9	2	100	9	4	6	10	2	11	8	11	11	14	10	16	16	6	8	9	5
Happ. Index	25	23	23	9	100	17	23	25	24	15	23	17	24	15	23	15	11	13	24	26	5
NRC Hashtag	21	21	16	4	17	100	23	23	17	21	20	18	24	16	23	13	22	27	21	20	15
Opin. Finder	27	26	28	6	23	23	100	30	29	17	30	20	29	14	26	14	12	18	29	29	7
Opin. Lexicon	34	29	26	10	25	23	30	100	26	20	28	20	32	18	28	18	17	19	28	31	8
PANAS	24	23	37	2	24	17	29	26	100	10	26	17	23	7	19	7	4	9	27	30	2
Pattern	20	17	9	11	15	21	17	20	10	100	17	17	24	19	23	19	22	24	21	17	12
SANN	27	24	25	8	23	20	30	28	26	17	100	19	26	14	25	15	12	16	28	28	6
SASA	20	18	17	11	17	18	20	20	17	17	19	100	20	15	20	14	16	16	21	21	9
SO-CAL	29	27	21	11	24	24	29	32	23	24	26	20	100	19	31	20	20	24	29	29	10
SWN	18	17	6	14	15	16	14	18	7	19	14	15	19	100	19	19	19	18	16	15	10
SentiStrength	28	25	17	10	23	23	26	28	19	23	25	20	31	19	100	20	19	24	29	27	11
SenticNet	19	16	6	16	15	13	14	18	7	19	15	14	20	19	20	100	18	16	16	15	9
Sentim.140	16	16	2	16	11	22	12	17	4	22	12	16	20	19	19	18	100	25	14	12	16
Stanford DM	18	18	8	6	13	27	18	19	9	24	16	16	24	18	24	16	25	100	19	16	16
Umigon	29	25	27	8	24	21	29	28	27	21	28	21	29	16	29	16	14	19	100	30	8
VADER	32	26	29	9	26	20	29	31	30	17	28	21	29	15	27	15	12	16	30	100	6
LIWC	8	8	1	5	5	15	7	8	2	12	6	9	10	10	11	9	16	16	8	6	100

Figure 5.4. Percentage of agreement among all methods in Tweets_DBT dataset.

data, picking the method that better fits with our needs (e.g.: time and memory usage, prediction performance, etc.) usage. For example, SO-CAL seems to be a good choice if time performance is not so important, since this method appears among the first positions in all other metrics. Some would even avoid the highlighted methods, those that possibly have a bias towards positive or negative, respectively, as showed before by Figure 5.2.

In the next section, we describe our initial efforts in combining all these 21 sentiment analysis methods and we also show the results of this strategy.

Table 5.7. Top 3 pairs of method with highest percentage of agreement in all labeled datasets

Dataset	Top 1 pair	Top 2 pair	Top 3 pair
Tweets_SANN	Emot. - PANAS (64%)	Op.Find. - PANAS (57%)	Emot. - Op.Find. (56%)
Tweets_DBT	Emot. - PANAS (37%)	AFINN - Op. Lex. (34%)	AFINN - Op. Lex. (32%) AFINN - Op. Vader (32%)
Tweets_RDN_I	Emot. - PANAS (65%)	Op.Find. - PANAS (62%)	O.Find. - Emoticons (61%)
Tweets_RDN_II	AFINN - Vader (45%)	S.Stren. - Umig. (44%)	AFINN - Op. Lex. (43%)
Tweets_RDN_III	S.Stren. - Umig. (60%)	S.Stren. - SO-CAL (56%)	S.Stren. - Sent.140 (55%)
Tweets_RND_IV	Emot. DS - Sent.140 (59%)	S.Stren. - S.Net (58%)	Patt. - Umig. (53%)
Tweets_Semeval	AFINN - Vader (49%)	Emot. - PANAS (48%)	Op. Lex. - Vader (47%)
Tweets_STF	S.Stren. - Umig. (60%)	Sent.140 - Umig. (56%)	Patt. - Umig. (55%)
Comments_TED	S.Stren. - Stanf. (38%)	Patt. - Umig. (36%)	Op. Lex. - SO-CAL (35%)
Comments_BBC	Emot. - PANAS (87%)	NRC Hash. - Stanf. (54%)	Stanf. DM - S.Stren. (72%)
Comments_Digg	NRC Hash. - Stanf. (54%) S.Stren. - Stanf. (54%)	NRC Hash. - Stanf. (51%)	S.Stren. - SO-CAL (47%)
Comments_NYT	Emot. DS - S.Net (36%)	SWN - S.Net (34%) SWN - Sent.140 (34%)	SO-CAL - Sent.140 (32%) Stanf. - Sent.140 (32%)
Comments_YTB	Patt. - S.Stren. (46%) Umig. - S.Stren. (46%)	SO-CAL - S.Stren. (45%)	SO-CAL - Patt. (44%) Umig. - Patt. (44%)
Reviews_I	SO-CAL - Stanf. (55%)	Patt. - Stanf. (49%)	Sent.140 - Stanf. (49%)
Reviews_II	Emot. DS - Sent.140 (49%)	Patt. - SO-CAL (42%)	S.Net - SO-CAL (41%) S.Net - SWN (41%)
Myspace	Emot. DS - S.Net (55%)	S.Stren. - Stanf.DS (54%) S.Stren. - S.Net (54%)	S.Stren. - Patt. (50%)
Amazon	Emot. DS - S.Net (49%)	Patt. - SO-CAL (69%)	S.Net - SO-CAL (41%) S.Net - SWN (41%)
RW	AFINN - Vader (42%) AFINN - LIWC (42%) Emot. DS - S.Net (42%)	AFINN - Op. Lex. (38%) Happ. Index - S.Net (38%)	AFINN - S.Net (37%)
YLP	Patt. - SO-CAL (76%)	Patt. - Sent.140 (73%) Sent.140 - SO-CAL (73%)	Patt. - S.Stren. (72%) SO-CAL - S.Stren. (72%)
Irony	NRC Hash. - Stanf. (42%)	AFINN - LIWC (41%)	AFINN - SO-CAL (40%)
Sarcasm	S.Stren. - Sent.140 (40%)	S.Stren. - Umig. (39%) AFINN - Op. Lex. (39%) AFINN - S.Stren. (39%)	AFINN - Vader (37%)

Ranking	Exec. time	Memory usa.	Av. MacroF1	Av. Accuracy	Winning number (MacroF1)	Winning number (Acc.)
1°	LIWC	OpinionLexicon	AFINN	AFINN	SO-CAL	SentiStrength
2°	OpinionLexicon	Emoticons	SO-CAL	SO-CAL	SentiStrength	SO-CAL
3°	NRC Hashag	PANAS-t	Umigon	SentiStrength	Umigon	Pattern.en
4°	Sentiment140 Lexicon	SO-CAL	Pattern.en	Pattern.en	Pattern.en	Umigon
5°	Happiness Index	SentiStrength	SentiStrength	Stanford DM	Opinion Lexicon	Opinion Lexicon
6°	OpinionFinder	Happiness Index	Vader	Umigon	Vader	Sentiment140
7°	EmoLex	NRC Hashag	Stanford DM	Sentiment140 Lexicon	Stanford DM	Vader
8°	Emoticons	Sentiment140 Lexicon	OpinionLexicon	Vader	AFINN	Stanford DM
9°	AFINN	EmoLex	Sentiment140 Lexicon	OpinionLexicon	Sentiment140	AFINN
10°	PANAS-t	AFINN	SANN	SANN	SANN	SANN
11°	Emoticons DS	Vader	OpinionFinder	NRC Hashtag	Emolex	NRC Hashtag
12°	Vader	SenticNet	Emolex	OpinionFinder	Opinion Finder	Opinion Finder
13°	SentiStrength	LIWC	NRC Hashtag	Emolex	NRC Hashtag	Emolex
14°	SASA	Emoticons DS	LIWC	LIWC	LIWC	LIWC
15°	SenticNet	Pattern.en	SASA	SentiWordNet	SASA	Emoticons DS
16°	Pattern.en	SentiWordNet	Happiness Index	SenticNet	SentiWordNet	PANAS-t
17°	SentiWordNet	Stanford DM	SentiWordNet	Happiness Index	SenticNet	Emoticons
18°	Umigon	SASA	PANAS-t	Emoticons DS	Happiness Index	Happiness Index
19°	SO-CAL	SANN	Emoticons DS	SASA	PANAS-t	SASA
20°	Stanford DM	Umigon	SenticNet	PANAS-t	Emoticons	SentiWordNet
21°	SANN	OpinionFinder	Emoticons	Emoticons	Emoticons DS	SenticNet

Figure 5.5. Ranking of 21 sentiment analysis methods in relation to measures used in this study for comparing them.

Table 5.8. Winning Points Ranking for MacroF1 and Accuracy with the combined method

Ranking	MacroF1 Winning score	Accuracy Winning score
SO-CAL	379	350
SentiStrength	369	351
Combined I	**359**	**336**
Umigon	326	295
Pattern.en	322	309
Opinion Lexicon	301	287
Vader	290	263
Stanford DM	267	262
AFINN	260	241
Sentiment140	250	273
SANN	247	229
Emolex	230	213
Opinion Finder	213	214
NRC Hashtag	202	226
LIWC	196	195
SASA	156	61
SentiWordNet	149	202
SenticNet	115	108
Happinness Index	111	63
PANAS-t	109	143
Emoticons	104	137
Emoticons DS	101	183

5.2 Combining Results

In this section we present the results achieved by the combined method proposed in this study. As described before, Combined I give the polarity of a message considering the output of the 21 sentiment analysis methods using a majority vote technique. Table 5.6 show the average prediction performance of the combined method, and Table 5.8 present the winning number score achieved by it compared to the 21 methods. In order to make the analysis easier, we also present Figure 5.6 that compares the average accuracy and the average MacroF1 of the combined and all methods. As we can see, the combined method are very competitive with the single methods, appearing in the top 3 methods with best accuracy and MacroF1.

While combining all sentiment methods would yield the best prediction performance, we also analyze that there is a diminishing return effect, in that increasing the number of methods incurs only marginal gain in accuracy and MacroF1 after some point. As we combine more methods, the prediction performance increases but to a smaller extent. This indicates that combining all of the methods is not necessarily the

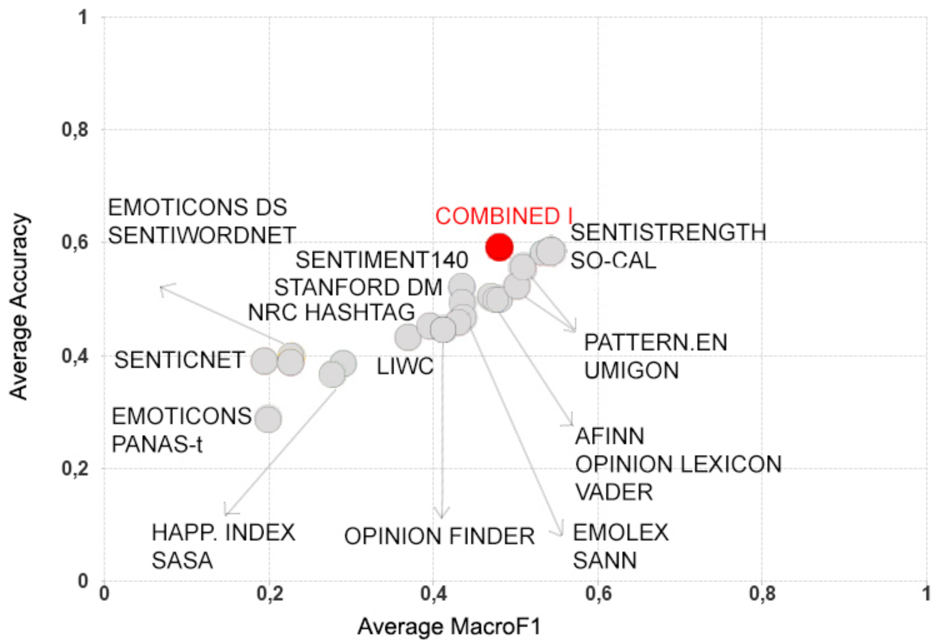


Figure 5.6. Average Accuracy vs. Average MacroF1 of the combined methods compared with the 21 methods

best strategy. The best accuracy and MacroF1 might be achieved by combining those methods best suited for a particular kind of data. For example, one might want to choose Umigon over SASA for a given data or vice versa. Reducing the amount of data needed for combined methods to obtain good results is a desirable property for a real system deployment, given that the use of fewer methods will likely require fewer resources.

Therefore Combined II could be considered as a non-practical method, since it uses the output of the 21 methods in order to tuning its own parameters (or weights), this method achieved competitive results, showing this might be an approach to be invested and investigated.

In the next chapter, we present the conclusions of this study and future works

Chapter 6

Conclusions and Future Work

Recent efforts to analyze the sentiment embedded in Web content have adapted various sentiment analysis methods and tools, 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 sentiments in the context of unstructured short messages in online social networks. Despite the vast interest on the theme and wide popularity of some tools, few is known about how they perform and even less about how they are compared to each other. In other words, it is unclear which one is better for identifying polarity. Such a comparison is key for understanding the potential limitations, advantages, and disadvantages of these methods.

In this study, we present a thorough comparison of 21 popular sentence-level sentiment analysis methods using gold standard datasets that span different types of data sources. To do it we survey the sentiment analysis area and have made significant efforts to obtain the latest working versions of the various sentiment analysis tools and datasets. The 21 methods analyzed vary from lexical to machine learning and hybrid approaches and they are: LIWC [93], Happiness Index [25], SentiWordNet [26], SASA [99], PANAS-t [32], Emoticons [31], Emoticons DS [34], SenticNet [14], SentiStrength [94], Stanford Recursive Deep Model [86], NRC Hashtag Lexicon [54], EmoLex [56], Sentiment140 Lexicon [57] , OpinionLexicon [37], VADER [38], OpinionFinder [103], AFINN [64], SO-CAL [92], Pattern.en [22], SANN [70] and Umigon [44]. The datasets cover an extensive collection of labeled: Yelp Dataset [109], Stanford Twitter Corpus [29], SentiStrength's dataset [94], a dataset from [23] with sentiment judgement of tweets from the first 2008 Presidential debate, movie-review documents [66], VADER's dataset [38], tweets with irony and sarcasm content (collected by us), comments from TED talks [70], random tweets [2, 62], tweets with specific topics [82] and tweets from the SemEval 2013 [61]) and unlabeled texts (tweets associated to global events filtered from [15] dataset). Since many not experts users and researchers are using these methods and tools without processing tasks such as parameters tuning or training, in other words, using these methods in a unsupervised way, in this work we are focused in comparing methods in this scenery.

Our comparison study focused on detecting the polarity of content (i.e., positive, negative and neutral affects) and does not yet consider other types of sentiments (e.g., psychological processes, such as anger or calmness). We adopted some measures of efficacy: execution time and memory usage performance, prediction performance (measuring the fraction of identified sentiments are in tune with the ground truth of the labeled datasets), winning number (that tries to assess the most competitive methods among a series of candidates, given a large series of pre-defined tasks they have to perform), agreement (measure the percentage of agreement between pairs of methods), and polarity in global events from Twitter (analyzing how methods behave in datasets related to real events).

Our experimental results of comparison highlighted many interesting points. First of all, we observed that almost all methods have constant memory usage when the size of the input increases, with the exception of OpinionFinder and Umigon, that showed to have a increase memory usage in this scenery. With this analysis, we also noted that SANN has problems in executing datasets with more than 10,000 sentences. In relation to the execution time, all methods increase the time of running when the size of input increased, as expected. this analysis is important since it present limitations of sentiment analysis methods to deal with big datasets due to memory and time performance, important attributes for the development of real time and mobile applications.

Regarding prediction performance, there is no clear winner in all cases. We found that the 21 methods have varying degrees of accuracy and macroF1 and no single method is always best across different text sources, suggesting that a preliminary investigation should be performed when sentiment analysis is used in a new dataset in order to guarantee a reasonable prediction performance. In this same analysis, we noted that neutral polarity showed to be harder to be detected by most of the methods. Furthermore, methods seemed to have a bias towards positive class. This observation was possible analyzing the average macroF1 of all three classes, positive, negative and neutral. These same results appeared in the analysis of global events filtered from the unlabeled dataset of Twitter. In this experiment, we showed that most methods present more positive values than negative values for all events, including those one where negative feelings are expected (e.g., tragedies). This finding might be related to characteristics of human language, which have a universal positivity bias as showed by [24]. In another experimental results we could observe that existing methods vary widely in terms of agreement about the predicted polarity, with scores ranging from 0% to 97%, implying that when analyzed with different methods, datasets could be interpreted very differently.

Finally, we present initial on showing the viability of combining 21 methods with aim at evaluate the viability of combining for maximize results of prediction performance. We proposed two simple combined methods based on majority voting (where the method choose the polarity given by most methods) and in accuracy weighting (where each method receive a weight based on previous results of prediction performance). We noted that, even built based in simple techniques, the combined method achieved competitive results of prediction performance when compared to all methods, showing that combining methods might an approach to be invested and investigated.

All methods, with the exception of LIWC due copyright restrictions, are build together in a single webpage (`www.ifeel.dcc.ufmg.br`)[4]. We release this Web system through which we would like to allow other researchers to easily compare results with the existing methods. More important, through this system one could easily test which method would be most suitable for a a particular dataset and 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 it also represents an important step towards the development of this research area as a whole.

This work has demonstrated a framework where various sentiment analysis methods can be compared in an apple-to-apple fashion. To be able to do this, we have covered a wide range of research on sentiment analysis and have made significant efforts to contact the authors of previous works to get access to their sentiment analysis tools. Unfortunately, in many cases, getting access to the tools was a nontrivial task; in this study, we were only able to compare 21 of the most widely used methods. As a natural extension of this work, we would like to continue to add more existing methods for comparison, such as the Profile of Mood States (POMS) [10]. Furthermore, we would like to expand the way we combine these methods by considering relevant machine learning techniques, such as meta learning, active learning and transfer learning.

Bibliography

[1] (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093 – 1113. ISSN 2090-4479.

[2] Aisopos, F. (2014). Manually annotated sentiment analysis twitter dataset ntua. `www.grid.ece.ntua.gr`.

[3] Amazon (2005). Amazon mechanical turk. `https://www.mturk.com/`. Accessed June 17, 2013.

[4] Araujo, M., Goncalves, P., Benevenuto, F., and Cha, M. (2014). ifeel: A system that compares and combines sentiment analysis methods. WWW.

[5] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). 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., and Tapias, D., editors, *LREC*. European Language Resources Association.

[6] Biever, C. (2010a). Twitter mood maps reveal emotional states of america. *New Scientist*, 207(2771):14 –. ISSN 0262-4079.

[7] Biever, C. (2010b). Twitter mood maps reveal emotional states of america. *The New Scientist*, 207.

[8] Bollen, J., Gonçalves, B., Ruan, G., and Mao, H. (2011). Happiness is assortative in online social networks. *CoRR*, abs/1103.0784.

[9] Bollen, J., Mao, H., and Zeng, X.-J. (2010). Twitter Mood Predicts the Stock Market. *CoRR*, abs/1010.3003.

[10] Bollen, J., Pepe, A., and Mao, H. (2009). Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. *CoRR*, abs/0911.1583.

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

[12] Brain, S. (2014). Facebook statistics. `http://www.statisticbrain.com/facebook-statistics/`. Accessed January, 2015.

[13] Cambria, E., Hussain, A., Havasi, C., Eckl, C., and Munro, J. (2010a). Towards crowd validation of the uk national health service. In *ACM Web Science Conference (WebSci)*.

[14] Cambria, E., Speer, R., Havasi, C., and Hussain, A. (2010b). Senticnet: A publicly available semantic resource for opinion mining. In *AAAI Fall Symposium Series*.

[15] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010). Measuring user influence in twitter: The million follower fallacy. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[16] Chauhan Ashish P., K. M. P. (2015). Sentiment analysis using hybrid approach: A survey. *Journal of Engineering Research and Applications*, 5:73--77.

[17] Choudhury, M. D., , and Counts, S. (2012a). The nature of emotional expression in social media: Measurement, inference, and utility.

[18] Choudhury, M. D., Gamon, M., and Counts, S. (2012b). Happy, nervous or surprised? classification of human affective states in social media.

[19] Company, T. (2015). Our mission: To give everyone the power to create and share ideas and information instantly, without barriers. `https://about.twitter.com/company`. Accessed January 26, 2015.

[20] Cool-Smileys (2010). List of text emoticons: The ultimate resource. `www.cool-smileys.com/text-emoticons`.

[21] De Choudhury, M., Counts, S., Horvitz, E. J., and Hoff, A. (2014). Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 626--638, New York, NY, USA. ACM.

[22] De Smedt, T. and Daelemans, W. (2012). Pattern for python. *The Journal of Machine Learning Research*, 13(1):2063--2067.

[23] Diakopoulos, N. and Shamma, D. (2010). Characterizing debate performance via aggregated twitter sentiment. In *Proc. CHI*.

[24] 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., and Danforth, C. M. (2015). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394.

[25] Dodds, P. S. and Danforth, C. M. (2009). Measuring the happiness of large-scale written expression: songs, blogs, and presidents. *J. of Happiness Studies*, 11.

[26] Esuli and Sebastiani (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proc. LREC*.

[27] Ewen, M. (2009). Omg! oxford english dictionary grows a heart: Graphic symbol for love (and that exclamation) are added as words. `tinyurl.com/klv36p`.

[28] Feldman, R. (2013). Techniques and applications for sentiment analysis. *Commun. ACM*, 56.

[29] Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *Processing*.

[30] Gomide, J., Veloso, A., Jr., W. M., Almeida, V., Benevenuto, F., Ferraz, F., and Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *ACM Web Science Conference (WebSci)*.

[31] Goncalves, P., Benevenuto, F., and Almeida, V. (2013). O que tweets contendo emoticons podem revelar sobre sentimentos coletivos? In *Proceedings of the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM13)*.

[32] Gonçalves, P., Benevenuto, F., and Cha, M. (2013). PANAS-t: A Pychometric Scale for Measuring Sentiments on Twitter. abs/1308.1857v1.

[33] Hai, Z., Chang, K., and Kim, J.-j. (2011). Implicit feature identification via co-occurrence association rule mining. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*, CICLing'11, pages 393--404, Berlin, Heidelberg. Springer-Verlag.

[34] Hannak, A., Anderson, E., Barrett, L. F., Lehmann, S., Mislove, A., and Riedewald, M. (2012). Tweetin' in the rain: Exploring societal-scale effects of weather on mood. In *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*.

[35] He, Y. and Zhou, D. (2011). Self-training from labeled features for sentiment analysis. *Information Processing and Management*, 47(4):606--616. ISSN 0306-4573.

[36] Hendler, J. (2009). Web 3.0 emerging. *Computer*, 42(1):111--113. ISSN 0018-9162.

[37] Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. Proc. KDD'04, pages 168--177.

[38] Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*.

[39] Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4):620--630.

[40] Kang, H., Yoo, S. J., and Han, D. (2012). Senti-lexicon and improved naive bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, 39(5):6000 – 6010. ISSN 0957-4174.

[41] Kaplan, R. and Saccuzzo, D. (2008). *Psychological Testing: Principles, Applications, and Issues*. Wadsworth Cengage Learning. ISBN 9780495095552.

[42] Kaufmann, M. (2012). Jmaxalign: A maximum entropy parallel sentence alignment tool. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Demonstration Papers, 8-15 December 2012, Mumbai, India*, pages 277--288.

[43] Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *41st Annual Meeting of the Association for Computational Linguistics*.

[44] Levallois, C. (2013). 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)*, pages 414--417, Atlanta, Georgia, USA. Association for Computational Linguistics.

[45] Li, Y.-M. and Li, T.-Y. (2013). Deriving market intelligence from microblogs. *Decision Support Systems*, 55(1):206--217.

[46] Liu, B. (2010). Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca*.

[47] Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, 28(2):203--208.

[48] MacAskill, E. (2009). US confirms it asked Twitter to stay open to help Iran protesters. `tinyurl.com/klv36p`.

[49] Manmatha, R., Rath, T., and Feng, F. (2001). Modeling score distributions for combining the outputs of search engines. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267--275. ACM.

[50] Messenger, Y. (2014). Yahoo messenger emoticons. `http://messenger.yahoo.com/features/emoticons`.

[51] Miller, G. A. (1995a). Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38.

[52] Miller, G. A. (1995b). Wordnet: a lexical database for english. *Communications of the ACM*, 38.

[53] min Kim, S. and Hovy, E. (2007). Crystal: Analyzing predictive opinions on the web. In *In EMNLPCoNLL 2007*.

[54] Mohammad, S. (2012). #emotional tweets. In *SEM*.

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

[56] Mohammad, S. and Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29.

[57] Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proc. SemEval-2013*.

[58] Moraes, F., Vasconcelos, M. A., Prado, P., Dalip, D. H., Almeida, J. M., and Goncalves, M. A. (2013). Polarity detection of foursquare tips. SOCINFO.

[59] MSGWeb (2006). List of emoticons in msn messenger. `http://messenger.msn.com/Resource/Emoticons.aspx`.

[60] Mudinas, A., Zhang, D., and Levene, M. (2012). Combining lexicon and learning based approaches for concept-level sentiment analysis. In *WISDOM*.

[61] Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., and Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in twitter.

[62] Narr, S., Hulfenhaus, M., and Albayrak, S. (2012). Language-independent twitter sentiment analysis. *Knowledge Discovery and Machine Learning (KDML)*, pages 12–14.

[63] Nielsen, F. Å. (2010). Afinn-96.txt.

[64] Nielsen, F. Å. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

[65] Oliveira, N., Cortez, P., and Areal, N. (2013). On the predictability of stock market behavior using stocktwits sentiment and posting volume. In Correia, L., Reis, L. P., and Cascalho, J., editors, *EPIA*, volume 8154 of *Lecture Notes in Computer Science*, pages 355–365. Springer.

[66] Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. Annual meeting of ACL Conference*.

[67] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Found. and Trends in IR*, 2.

[68] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP*.

[69] Pappas, N., Katsimpras, G., and Stamatatos, E. (2013). Distinguishing the popularity between topics: A system for up-to-date opinion retrieval and mining in the web. In *14th International Conference on Intelligent Text Processing and Computational Linguistics*. LNCS, ACM.

[70] Pappas, N. and Popescu-Belis, A. (2013). 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*, pages 773--776. ACM.

[71] Park, J., Barash, V., Fink, C., and Cha, M. (2013). Emoticon style: Interpreting differences in emoticons across cultures. In *Int'L AAAI Conference on Weblogs and Social Media (ICWSM)*.

[72] Plutchik, R. (1980). *A general psychoevolutionary theory of emotion*, pages 3--33. Academic press, New York.

[73] Popescu, A.-M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 339--346, Stroudsburg, PA, USA. Association for Computational Linguistics.

[74] Qin, T., Liu, T.-Y., Xu, J., and Li, H. (2010). Letor: A benchmark collection for research on learning to rank for information retrieval. *Inf. Retr.*, 13.

[75] Qiu, G., He, X., Zhang, F., Shi, Y., Bu, J., and Chen, C. (2010). Dasa: Dissatisfaction-oriented advertising based on sentiment analysis. *Expert Systems with Applications*, 37(9):6182 – 6191. ISSN 0957-4174.

[76] Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn.*, 1(1):81--106. ISSN 0885-6125.

[77] Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *ACL Student Research Workshop*.

[78] Read, J. and Carroll, J. (2009). Weakly supervised techniques for domain-independent sentiment classification. In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, pages 45--52, New York, NY, USA. ACM.

[79] Richardson, M. and Domingos, P. (2001). The intelligent surfer: Probabilistic combination of link and content information in pagerank. In *NIPS*, pages 1441--1448.

[80] Saif, H., He, Y., and Alani, H. (2012). Semantic sentiment analysis of twitter. In *Proceedings of the 11th International Conference on The Semantic Web - Volume Part I*, ISWC'12, pages 508--524, Berlin, Heidelberg. Springer-Verlag.

[81] Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Int'l Conference on World wide web (WWW)*.

[82] Sanders, N. (2011). Twitter sentiment corpus by niek sanders. `http://www.sananalytics.com/lab/twitter-sentiment/`.

[83] Schapire, R. E. and Singer, Y. (2000). Boostexter: A boosting-based system for text categorization. In *Machine Learning*, pages 135--168.

[84] Schwartz, A. H., Eichstaedt, J., Kern, L. M., Park, G., Sap, M., Stillwell, D., Kosinski, M., and Ungar, L. (2014). *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, chapter Towards Assessing Changes in Degree of Depression through Facebook, pages 118--125. Association for Computational Linguistics.

[85] Seewald, A. K. (2002). Meta-learning for stacked classification. *Austrian Research Institute for Artificial Intelligence*, 4.

[86] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Conf. on Empirical Methods in NLP*.

[87] Somasundaran, S., Wiebe, J., and Ruppenhofer, J. (2008). Discourse level opinion interpretation. In *Int'l Conference on Computational Linguistics (COLING)*.

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

[89] Strapparava, C. and Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 70--74, Stroudsburg, PA, USA. Association for Computational Linguistics.

[90] Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor. ISBN 0385721706.

[91] Taboada, M., Anthony, C., and Voll, K. (2006). Methods for creating semantic orientation dictionaries. In *Conference on Language Resources and Evaluation (LREC)*, pages 427--432.

[92] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267--307. ISSN 0891-2017.

[93] Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *J. of Lang. and Soc. Psych.*, 29.

[94] Thelwall, M. (2013). Heart and soul: Sentiment strength detection in the social web with sentistrength. `http://sentistrength.wlv.ac.uk/documentation/SentiStrengthChapter.pdf`.

[95] Tsytsarau, M. and Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Min. Knowl. Discov.*, 24(3):478--514. ISSN 1384-5810.

[96] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *ICWSM*.

[97] Valitutti, R. (2004). Wordnet-affect: an affective extension of wordnet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*.

[98] van de Camp, M. and van den Bosch, A. (2012). The socialist network. *Decision Support Systems*, 53(4):761--769. ISSN 0167-9236. 1) Computational Approaches to Subjectivity and Sentiment Analysis 2) Service Science in Information Systems Research : Special Issue on {PACIS} 2010.

[99] Wang, H., Can, D., Kazemzadeh, A., Bar, F., and Narayanan, S. (2012). A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In *ACL System Demonstrations*.

[100] Watson, D. and Clark, L. (1985). Development and validation of brief measures of positive and negative affect: the panas scales. *J. of Pers. and So. Psych.*, 54.

[101] Watson, D. and Clark, L. A. (1994). The PANAS-X: Manual for the Positive and Negative Affect Schedule - Expanded Form.

[102] Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2):0.

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

[104] Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005b). Opinionfinder: a system for subjectivity analysis. In *HLT/EMNLP on Interactive Demonstrations*.

[105] Wilson, T., Wiebe, J., and Hoffmann, P. (2005c). Recognizing contextual polarity in phrase-level sentiment analysis. In *ACL Conference on Empirical Methods in Natural Language Processing*.

[106] Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Trans. On Evol. Comp.*, 1.

[107] Wu, Y., Zhang, Q., Huang, X., and Wu, L. (2009). Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1533--1541, Stroudsburg, PA, USA. Association for Computational Linguistics.

[108] Xianghua, F., Guo, L., Yanyan, G., and Zhiqiang, W. (2013). Multi-aspect sentiment analysis for chinese online social reviews based on topic modeling and hownet lexicon. *Knowledge-Based Systems*, 37(0):186--195. ISSN 0950-7051.

[109] Yelp (2014). Yelp dataset challenge. `http://www.yelp.com/dataset_challenge`. Accessed April 23, 2014.

[110] Yi, H. and Li, W. (2011). Document sentiment classification by exploring description model of topical terms. *Comput Speech Lang*, 25:386--403.

[111] Yu, H. and Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 129--136, Stroudsburg, PA, USA. Association for Computational Linguistics.

[112] Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., and Liu, B. (2011). Combining lexicon-based and learning-based methods for twitter sentiment analysis. Technical report, HP.

Appendix A

Complete Results of Prediction Performance

Table A.1. Prediction performance of all methods in Tweets_SAN dataset, including the combined method

Dataset	Method	Acc.	Posit. sentiment			Negat. sentiment			Neut. sentiment			MacroF1
			P	R	F1	P	R	F1	P	R	F1	
Tweets_SAN	AFINN	0.60	0.27	0.52	0.36	0.48	0.37	0.42	0.78	0.67	0.72	0.50
	Emolex	0.60	0.23	0.30	0.26	0.43	0.30	0.35	0.74	0.75	0.74	0.45
	Emotic.	0.67	0.30	0.07	0.11	0.51	0.03	0.06	0.69	0.96	0.80	0.32
	Emot. DS	0.18	0.15	0.95	0.26	0.40	0.01	0.03	0.70	0.05	0.08	0.12
	H. Index	0.54	0.20	0.48	0.28	0.00	0.00	0.00	0.73	0.68	0.70	0.33
	LIWC	0.58	0.62	0.28	0.39	0.37	0.49	0.42	0.62	0.79	0.69	0.50
	NRC H.	0.47	0.17	0.14	0.15	0.26	0.68	0.38	0.74	0.49	0.59	0.37
	O.Finder	0.65	0.31	0.24	0.27	0.40	0.24	0.30	0.74	0.85	0.79	0.45
	Opin. Lex.	0.63	0.31	0.38	0.35	0.45	0.33	0.38	0.75	0.77	0.76	0.49
	PANAS-t	0.68	0.47	0.06	0.10	0.19	0.02	0.03	0.69	0.97	0.81	0.31
	Pattern.en	0.52	0.26	0.66	0.38	0.40	0.51	0.45	0.81	0.49	0.61	0.48
	SANN	0.60	0.28	0.41	0.33	0.40	0.25	0.31	0.75	0.74	0.74	0.46
	SASA	0.50	0.19	0.50	0.28	0.00	0.00	0.00	0.70	0.63	0.66	0.31
	SO-CAL	0.59	0.28	0.56	0.37	0.48	0.53	0.51	0.80	0.61	0.69	0.52
	SWN	0.16	0.15	0.99	0.26	0.00	0.00	0.00	0.73	0.01	0.02	0.09
	S.Strength	0.56	0.30	0.68	0.42	0.41	0.58	0.48	0.85	0.52	0.65	0.51
	SenticNet	0.15	0.15	1.00	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.09
	Sentim.140	0.25	0.18	0.63	0.28	0.29	0.71	0.41	0.77	0.05	0.10	0.26
	Stanf.DM	0.25	0.43	0.20	0.28	0.18	0.93	0.31	0.74	0.10	0.17	0.25
	Umigon	0.60	0.36	0.57	0.44	0.40	0.55	0.46	0.81	0.63	0.70	0.54
	Vader	0.66	0.37	0.48	0.42	0.55	0.27	0.36	0.75	0.80	0.77	0.52
	Combined I	0.67	0.52	0.38	0.44	0.44	0.56	0.49	0.77	0.79	0.78	0.57

Table A.2. Prediction performance of all methods in Tweets_RND_IV and Tweets_DBT datasets, including the combined method

Dataset	**Method**	**Acc.**	**Posit. sentiment**			**Negat. sentiment**			**Neut. sentiment**			**MacroF1**
			P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	
	AFINN	0.50	0.48	0.57	0.52	0.52	0.38	0.44	0.50	0.50	0.50	0.49
	Emolex	0.48	0.49	0.41	0.44	0.39	0.26	0.31	0.50	0.65	0.57	0.44
	Emotic.	0.77	0.76	0.68	0.72	0.82	0.82	0.82	0.75	0.81	0.78	0.77
	Emot. DS	0.34	0.33	0.98	0.49	0.83	0.04	0.08	0.61	0.05	0.09	0.22
	H. Index	0.40	0.33	0.57	0.42	0.00	0.00	0.00	0.48	0.49	0.48	0.30
	LIWC	0.44	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	NRC H.	0.44	0.51	0.20	0.29	0.34	0.72	0.46	0.57	0.47	0.51	0.42
Tweets_	O.Finder	0.49	0.54	0.32	0.40	0.38	0.27	0.32	0.50	0.73	0.60	0.44
RND_IV	Opin. Lex.	0.53	0.52	0.44	0.48	0.48	0.34	0.40	0.54	0.69	0.61	0.49
	PANAS-t	0.48	0.70	0.09	0.16	0.46	0.11	0.18	0.47	0.96	0.64	0.32
	Pattern.en	0.64	0.58	0.89	0.70	0.61	0.87	0.72	0.87	0.34	0.49	0.64
	SANN	0.47	0.44	0.42	0.43	0.43	0.25	0.32	0.49	0.61	0.54	0.43
	SASA	0.42	0.35	0.62	0.45	0.00	0.00	0.00	0.51	0.50	0.51	0.32
	SO-CAL	0.55	0.49	0.54	0.52	0.54	0.47	0.50	0.59	0.59	0.59	0.54
	SWN	0.32	0.32	0.99	0.48	0.00	0.00	0.00	0.44	0.02	0.03	0.17
	S.Strength	0.54	0.54	0.72	0.62	0.41	0.50	0.45	0.67	0.43	0.52	0.53
	SenticNet	0.32	0.32	1.00	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.16
	Sentim.140	0.47	0.45	0.70	0.55	0.46	0.89	0.61	0.78	0.08	0.15	0.43
	Stanf. DM	0.35	0.68	0.36	0.47	0.27	0.82	0.40	0.45	0.09	0.16	0.34
	Umigon	0.74	0.70	0.83	0.76	0.66	0.91	0.76	0.90	0.59	0.72	0.75
	Vader	0.62	0.63	0.71	0.67	0.73	0.37	0.49	0.58	0.69	0.63	0.60
	Combined I	0.62	0.60	0.64	0.62	0.52	0.63	0.57	0.69	0.61	0.65	0.61
	AFINN	0.43	0.34	0.42	0.37	0.60	0.25	0.36	0.42	0.61	0.50	0.41
	Emolex	0.41	0.27	0.32	0.29	0.55	0.29	0.38	0.42	0.58	0.49	0.39
	Emotic.	0.39	0.29	0.01	0.03	0.47	0.01	0.01	0.39	0.98	0.55	0.20
	Emot. DS	0.24	0.23	0.97	0.37	0.73	0.01	0.01	0.50	0.04	0.07	0.15
	H. Index	0.33	0.22	0.40	0.28	0.00	0.00	0.00	0.40	0.61	0.49	0.26
	LIWC	0.39	0.52	0.67	0.59	0.29	0.82	0.43	0.00	0.00	0.00	0.34
	NRC H.	0.45	0.30	0.13	0.18	0.48	0.67	0.56	0.45	0.42	0.43	0.39
	O.Finder	0.43	0.35	0.18	0.24	0.57	0.28	0.38	0.41	0.73	0.53	0.38
Tweets_	Opin. Lex.	0.45	0.37	0.37	0.37	0.61	0.29	0.39	0.44	0.67	0.53	0.43
_DBT	PANAS-t	0.39	0.29	0.02	0.04	0.64	0.04	0.07	0.39	0.96	0.55	0.22
	Pattern.en	0.41	0.33	0.47	0.38	0.47	0.57	0.52	0.38	0.21	0.27	0.39
	SANN	0.42	0.29	0.28	0.29	0.59	0.25	0.35	0.41	0.66	0.51	0.38
	SASA	0.31	0.24	0.61	0.34	0.00	0.00	0.00	0.41	0.43	0.42	0.25
	SO-CAL	0.47	0.39	0.44	0.41	0.59	0.41	0.48	0.45	0.56	0.50	0.46
	SWN	0.23	0.22	0.95	0.36	0.00	0.00	0.00	0.42	0.04	0.07	0.14
	S.Strength	0.45	0.32	0.39	0.35	0.54	0.48	0.51	0.46	0.45	0.45	0.44
	SenticNet	0.23	0.23	1.00	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.12
	Sentim.140	0.40	0.30	0.65	0.41	0.50	0.61	0.55	0.57	0.05	0.09	0.35
	Stanf. DM	0.44	0.47	0.23	0.31	0.44	0.82	0.57	0.45	0.20	0.27	0.39
	Umigon	0.45	0.40	0.28	0.33	0.58	0.31	0.40	0.41	0.68	0.52	0.42
	Vader	0.44	0.40	0.32	0.35	0.67	0.19	0.29	0.41	0.76	0.54	0.39
	Combined I	0.31	0.00	0.00	0.00	0.30	0.99	0.46	0.80	0.03	0.05	0.17

Table A.3. Prediction performance of all methods in Tweets_RND_III and Irony datasets, including the combined method

Dataset	Method	Acc.	Posit. sentiment			Negat. sentiment			Neut. sentiment			MacroF1
			P	R	F1	P	R	F1	P	R	F1	
Tweets _RDN_III	AFINN	0.64	0.41	0.65	0.50	0.49	0.48	0.48	0.81	0.67	0.73	0.57
	Emolex	0.64	0.38	0.41	0.40	0.43	0.41	0.42	0.76	0.75	0.75	0.52
	Emotic.	0.70	0.70	0.17	0.27	0.66	0.09	0.15	0.70	0.98	0.82	0.41
	Emot. DS	0.21	0.20	0.98	0.33	0.90	0.04	0.07	0.60	0.02	0.04	0.15
	H. Index	0.53	0.27	0.65	0.38	0.00	0.00	0.00	0.77	0.59	0.67	0.35
	LIWC	0.47	0.38	0.22	0.28	0.18	0.19	0.18	0.55	0.70	0.62	0.36
	NRC H.	0.51	0.39	0.30	0.34	0.25	0.80	0.39	0.78	0.52	0.62	0.45
	O.Finder	0.72	0.57	0.33	0.42	0.50	0.35	0.41	0.76	0.90	0.82	0.55
	Opin. Lex.	0.70	0.48	0.50	0.49	0.56	0.43	0.48	0.78	0.81	0.80	0.59
	PANAS-t	0.70	0.77	0.13	0.22	0.55	0.06	0.11	0.70	0.98	0.82	0.38
	Pattern.en	0.54	0.36	0.77	0.49	0.35	0.59	0.44	0.84	0.46	0.59	0.51
	SANN	0.67	0.43	0.49	0.46	0.46	0.36	0.40	0.78	0.78	0.78	0.55
	SASA	0.52	0.26	0.67	0.37	0.00	0.00	0.00	0.78	0.57	0.66	0.34
	SO-CAL	0.67	0.43	0.69	0.53	0.52	0.61	0.56	0.84	0.67	0.75	0.61
	SWN	0.32	0.24	0.71	0.36	0.23	0.49	0.32	0.75	0.17	0.28	0.32
	S.Strength	0.65	0.45	0.80	0.58	0.42	0.73	0.54	0.92	0.58	0.71	0.61
	SenticNet	0.20	0.20	1.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.11
	Sentim.140	0.29	0.24	0.72	0.36	0.28	0.76	0.41	0.81	0.07	0.13	0.30
	Stanf. DM	0.32	0.64	0.39	0.48	0.16	0.85	0.26	0.76	0.20	0.31	0.35
	Umigon	0.74	0.58	0.70	0.63	0.49	0.68	0.57	0.89	0.76	0.82	0.67
	Vader	0.73	0.54	0.65	0.59	0.68	0.41	0.51	0.81	0.82	0.81	0.64
	Combined I	0.77	0.67	0.60	0.63	0.55	0.69	0.61	0.85	0.84	0.84	0.70
Irony	AFINN	0.56	0.65	0.59	0.62	0.86	0.42	0.56	0.35	0.88	0.50	0.56
	Emolex	0.47	0.53	0.45	0.49	0.86	0.42	0.56	0.24	0.63	0.35	0.47
	Emotic.	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.20	1.00	0.33	0.11
	Emot. DS	0.32	0.29	1.00	0.45	1.00	0.02	0.05	0.75	0.19	0.30	0.26
	H. Index	0.33	0.41	0.64	0.50	0.00	0.00	0.00	0.28	0.81	0.41	0.30
	LIWC	0.49	0.64	0.50	0.56	0.30	0.87	0.45	0.81	0.34	0.48	0.50
	NRC H.	0.59	0.44	0.18	0.26	0.71	0.84	0.77	0.38	0.50	0.43	0.49
	O.Finder	0.38	0.70	0.32	0.44	0.89	0.19	0.31	0.26	1.00	0.41	0.39
	Opin. Lex.	0.44	0.53	0.36	0.43	0.88	0.33	0.47	0.28	0.88	0.42	0.44
	PANAS-t	0.21	0.00	0.00	0.00	0.50	0.02	0.04	0.20	1.00	0.34	0.13
	Pattern.en	0.53	0.63	0.77	0.69	0.76	0.30	0.43	0.35	0.81	0.49	0.54
	SANN	0.41	0.41	0.41	0.41	1.00	0.23	0.38	0.29	0.88	0.43	0.41
	SASA	0.25	0.31	0.55	0.39	0.00	0.00	0.00	0.19	0.50	0.28	0.22
	SO-CAL	0.56	0.59	0.59	0.59	0.83	0.47	0.60	0.34	0.75	0.47	0.55
	SWN	0.27	0.28	1.00	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.14
	S.Strength	0.56	0.53	0.45	0.49	0.65	0.60	0.63	0.41	0.56	0.47	0.53
	SenticNet	0.27	0.27	1.00	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.14
	Sentim.140	0.53	0.42	0.68	0.52	0.63	0.56	0.59	0.57	0.25	0.35	0.49
	Stanf.DM	0.63	0.77	0.45	0.57	0.64	0.84	0.73	0.42	0.31	0.36	0.55
	Umigon	0.42	0.53	0.41	0.46	0.64	0.33	0.43	0.26	0.69	0.38	0.42
	Vader	0.42	0.71	0.45	0.56	0.89	0.19	0.31	0.28	1.00	0.43	0.43
	Combined I	0.51	0.50	0.69	0.58	0.35	0.88	0.50	0.94	0.31	0.47	0.52

Table A.4. Prediction performance of all methods in Comments_TED and Reviews_I datasets, including the combined method

Dataset	Method	Acc.	Posit. sentiment			Negat. sentiment			Neut. sentiment			MacroF1
			P	R	F1	P	R	F1	P	R	F1	
Comments_TED	AFINN	0.45	0.58	0.64	0.61	0.74	0.27	0.39	0.18	0.54	0.27	0.42
	Emolex	0.43	0.53	0.47	0.50	0.70	0.39	0.50	0.16	0.46	0.24	0.41
	Emotic.	0.15	1.00	0.01	0.02	0.89	0.02	0.04	0.14	1.00	0.24	0.10
	Emot. DS	0.37	0.37	0.96	0.54	0.00	0.00	0.00	0.12	0.03	0.04	0.19
	H. Index	0.31	0.44	0.66	0.53	0.00	0.00	0.00	0.15	0.47	0.22	0.25
	LIWC	0.32	0.38	0.41	0.39	0.24	0.54	0.33	0.45	0.14	0.21	0.31
	NRC H.	0.45	0.52	0.27	0.36	0.61	0.63	0.62	0.15	0.34	0.21	0.39
	O.Finder	0.42	0.59	0.38	0.46	0.65	0.44	0.52	0.15	0.48	0.23	0.41
	Opin. Lex.	0.43	0.54	0.52	0.53	0.72	0.34	0.46	0.18	0.54	0.27	0.42
	PANAS-t	0.17	0.74	0.07	0.13	0.72	0.03	0.06	0.14	0.96	0.24	0.14
	Pattern.en	0.52	0.57	0.70	0.63	0.62	0.46	0.53	0.20	0.26	0.22	0.46
	SANN	0.49	0.61	0.56	0.58	0.68	0.45	0.54	0.17	0.43	0.25	0.46
	SASA	0.41	0.52	0.51	0.51	0.64	0.34	0.45	0.12	0.34	0.18	0.38
	SO-CAL	0.53	0.65	0.70	0.67	0.65	0.45	0.53	0.17	0.31	0.22	0.47
	SWN	0.37	0.38	0.97	0.54	0.00	0.00	0.00	0.12	0.02	0.03	0.19
	S.Strength	0.54	0.71	0.54	0.61	0.66	0.57	0.61	0.20	0.44	0.27	0.50
	SenticNet	0.38	0.38	1.00	0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.18
	Sentim.140	0.52	0.45	0.68	0.54	0.65	0.53	0.58	0.10	0.03	0.04	0.39
	Stanf. DM	0.56	0.76	0.51	0.61	0.60	0.70	0.65	0.12	0.16	0.14	0.47
	Umigon	0.41	0.68	0.51	0.59	0.59	0.25	0.35	0.18	0.67	0.28	0.40
	Vader	0.40	0.75	0.48	0.58	0.74	0.25	0.38	0.16	0.71	0.26	0.41
	Combined I	0.46	0.54	0.68	0.61	0.36	0.70	0.48	0.58	0.17	0.27	0.45
Reviews_I	AFINN	0.42	0.63	0.55	0.59	0.71	0.29	0.41	0.00	0.00	0.00	0.50
	Emolex	0.47	0.61	0.53	0.57	0.63	0.40	0.49	0.00	0.00	0.00	0.53
	Emotic.	0.00	0.00	0.00	0.00	0.60	0.00	0.00	0.00	0.00	0.00	0.00
	Emot. DS	0.50	0.50	1.00	0.67	1.00	0.00	0.00	0.00	0.00	0.00	0.34
	H. Index	0.32	0.51	0.65	0.57	0.00	0.00	0.00	0.00	0.00	0.00	0.29
	LIWC	0.42	0.58	0.61	0.59	0.25	0.67	0.36	0.00	0.00	0.00	0.48
	NRC H.	0.38	0.67	0.21	0.32	0.58	0.56	0.57	0.00	0.00	0.00	0.45
	O.Finder	0.38	0.70	0.22	0.34	0.58	0.54	0.56	0.00	0.00	0.00	0.45
	Opin. Lex.	0.51	0.69	0.57	0.62	0.70	0.45	0.55	0.00	0.00	0.00	0.59
	PANAS-t	0.05	0.69	0.07	0.12	0.55	0.04	0.07	0.00	0.00	0.00	0.10
	Pattern.en	0.58	0.65	0.62	0.64	0.66	0.55	0.60	0.00	0.00	0.00	0.62
	SANN	0.42	0.62	0.49	0.55	0.63	0.35	0.45	0.00	0.00	0.00	0.50
	SASA	0.28	0.49	0.57	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.27
	SO-CAL	0.64	0.72	0.66	0.69	0.71	0.61	0.66	0.00	0.00	0.00	0.68
	SWN	0.49	0.50	0.99	0.66	0.00	0.00	0.00	0.00	0.00	0.00	0.33
	S.Strength	0.50	0.64	0.43	0.52	0.61	0.56	0.58	0.00	0.00	0.00	0.55
	SenticNet	0.50	0.50	1.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.34
	Sentim.140	0.61	0.58	0.85	0.69	0.72	0.37	0.49	0.00	0.00	0.00	0.59
	Stanf.DM	0.76	0.88	0.70	0.78	0.78	0.82	0.80	0.00	0.00	0.00	0.79
	Umigon	0.34	0.66	0.30	0.42	0.61	0.38	0.47	0.00	0.00	0.00	0.45
	Vader	0.30	0.71	0.43	0.53	0.73	0.17	0.28	0.00	0.00	0.00	0.41
	Combined I	0.50	0.64	0.68	0.66	0.37	0.76	0.50	0.00	0.00	0.00	0.58

Table A.5. Prediction performance of all methods in Sarcasm and Comments_BBC datasets, including the combined method

Dataset	Method	Acc.	Posit. sentiment			Negat. sentiment			Neut. sentiment			MacroF1
			P	R	F1	P	R	F1	P	R	F1	
	AFINN	0.51	0.49	0.64	0.55	0.65	0.29	0.40	0.46	0.67	0.54	0.50
	Emolex	0.46	0.45	0.45	0.45	0.71	0.32	0.44	0.38	0.71	0.49	0.46
	Emotic.	0.25	0.50	0.03	0.06	0.00	0.00	0.00	0.25	0.96	0.39	0.15
	Emot. DS	0.36	0.35	1.00	0.52	0.00	0.00	0.00	1.00	0.04	0.08	0.20
	H. Index	0.32	0.34	0.48	0.40	0.00	0.00	0.00	0.29	0.58	0.39	0.26
	LIWC	0.49	0.73	0.49	0.59	0.34	0.72	0.46	0.42	0.36	0.39	0.48
	NRC H.	0.53	0.50	0.33	0.40	0.58	0.82	0.68	0.40	0.33	0.36	0.48
	Opin.Finder	0.48	0.55	0.48	0.52	0.69	0.24	0.35	0.40	0.88	0.55	0.47
	Opin. Lex.	0.48	0.56	0.58	0.57	0.61	0.29	0.39	0.37	0.67	0.48	0.48
Sarcasm	PANAS-t	0.27	0.00	0.00	0.00	0.67	0.05	0.10	0.26	1.00	0.42	0.17
	Pattern.en	0.49	0.46	0.76	0.57	0.52	0.34	0.41	0.56	0.38	0.45	0.48
	SANN	0.42	0.43	0.58	0.49	0.73	0.21	0.33	0.33	0.54	0.41	0.41
	SASA	0.31	0.31	0.61	0.41	0.00	0.00	0.00	0.29	0.38	0.33	0.25
	SO-CAL	0.53	0.45	0.58	0.51	0.84	0.42	0.56	0.44	0.63	0.52	0.53
	SWN	0.34	0.33	0.91	0.49	0.00	0.00	0.00	0.40	0.08	0.14	0.21
	S.Strength	0.58	0.56	0.76	0.64	0.66	0.61	0.63	0.47	0.29	0.36	0.54
	SenticNet	0.35	0.35	1.00	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.17
	Sentim.140	0.58	0.53	0.76	0.63	0.61	0.74	0.67	1.00	0.08	0.15	0.48
	Stanf. DM	0.49	0.70	0.42	0.53	0.47	0.74	0.57	0.33	0.21	0.26	0.45
	Umigon	0.51	0.60	0.55	0.57	0.55	0.42	0.48	0.39	0.58	0.47	0.51
	Vader	0.45	0.59	0.61	0.60	1.00	0.11	0.19	0.33	0.79	0.47	0.42
	Combined I	0.57	0.67	0.56	0.61	0.42	0.80	0.55	0.67	0.44	0.53	0.57
	AFINN	0.45	0.15	0.49	0.23	0.84	0.45	0.59	0.34	0.44	0.38	0.40
	Emolex	0.48	0.16	0.54	0.25	0.82	0.53	0.64	0.34	0.36	0.35	0.41
	Emotic.	0.25	0.00	0.00	0.00	0.33	0.00	0.00	0.25	0.99	0.39	0.13
	Emot. DS	0.10	0.10	0.96	0.18	0.50	0.00	0.00	0.31	0.02	0.03	0.07
	Happ. Index	0.15	0.09	0.68	0.16	0.00	0.00	0.00	0.36	0.34	0.35	0.17
	LIWC	0.33	0.42	0.11	0.17	0.34	0.68	0.45	0.28	0.23	0.25	0.29
	NRC H.	0.64	0.19	0.12	0.15	0.71	0.87	0.78	0.42	0.23	0.29	0.41
Comments	O.Finder	0.52	0.15	0.35	0.21	0.79	0.60	0.68	0.34	0.36	0.35	0.41
_BBC	Opin. Lex.	0.51	0.19	0.48	0.28	0.87	0.52	0.65	0.35	0.50	0.41	0.45
	PANAS-t	0.28	0.14	0.08	0.10	0.72	0.08	0.14	0.25	0.88	0.39	0.21
	Pattern.en	0.46	0.14	0.59	0.23	0.77	0.53	0.63	0.38	0.23	0.29	0.38
	SANN	0.40	0.15	0.60	0.23	0.80	0.38	0.51	0.33	0.38	0.36	0.37
	SASA	0.17	0.12	0.72	0.20	0.00	0.00	0.00	0.25	0.40	0.31	0.17
	SO-CAL	0.56	0.21	0.58	0.31	0.80	0.63	0.71	0.40	0.35	0.37	0.46
	SWN	0.10	0.10	0.96	0.17	0.00	0.00	0.00	0.50	0.02	0.05	0.07
	S.Strength	0.65	0.32	0.55	0.41	0.76	0.81	0.79	0.48	0.26	0.34	0.51
	SenticNet	0.10	0.10	1.00	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.06
	Sentim.140	0.43	0.13	0.63	0.21	0.74	0.56	0.64	0.44	0.02	0.03	0.29
	Stanf. DM	0.66	0.43	0.36	0.40	0.71	0.89	0.79	0.38	0.15	0.21	0.47
	Umigon	0.46	0.28	0.36	0.32	0.76	0.41	0.53	0.29	0.62	0.40	0.42
	Vader	0.45	0.18	0.44	0.26	0.89	0.39	0.54	0.31	0.59	0.41	0.40
	Combined I	0.55	0.53	0.20	0.29	0.62	0.83	0.71	0.36	0.35	0.36	0.46

Table A.6. Prediction performance of all methods in Comments_Digg and Myspace datasets, including the combined method

Dataset	Method	Acc.	Posit. sentiment			Negat. sentiment			Neut. sentiment			MacroF1
			P	R	F1	P	R	F1	P	R	F1	
Comments_Digg	AFINN	0.47	0.33	0.51	0.40	0.82	0.38	0.52	0.37	0.61	0.46	0.46
	Emolex	0.43	0.22	0.29	0.25	0.74	0.41	0.53	0.35	0.58	0.43	0.40
	Emotic.	0.29	0.65	0.06	0.11	0.60	0.01	0.02	0.28	0.98	0.43	0.19
	Emot. DS	0.21	0.19	0.91	0.32	0.60	0.01	0.01	0.39	0.09	0.15	0.16
	H. Index	0.25	0.18	0.56	0.27	0.00	0.00	0.00	0.36	0.53	0.43	0.24
	LIWC	0.31	0.35	0.20	0.25	0.24	0.51	0.33	0.42	0.27	0.33	0.30
	NRC H.	0.52	0.28	0.10	0.15	0.63	0.76	0.69	0.33	0.35	0.34	0.39
	O.Finder	0.46	0.33	0.32	0.33	0.71	0.43	0.53	0.35	0.62	0.45	0.44
	Opin. Lex.	0.44	0.31	0.43	0.36	0.77	0.37	0.50	0.33	0.57	0.42	0.43
	PANAS-t	0.29	0.20	0.02	0.04	0.82	0.05	0.09	0.28	0.96	0.43	0.19
	Pattern.en	0.49	0.33	0.60	0.43	0.70	0.49	0.57	0.41	0.42	0.41	0.47
	SANN	0.42	0.27	0.46	0.34	0.76	0.33	0.46	0.35	0.56	0.43	0.41
	SASA	0.24	0.20	0.66	0.30	0.00	0.00	0.00	0.32	0.40	0.35	0.22
	SO-CAL	0.54	0.39	0.53	0.45	0.75	0.56	0.64	0.40	0.49	0.44	0.51
	SWN	0.21	0.19	0.92	0.32	0.00	0.00	0.00	0.51	0.13	0.20	0.17
	S.Strength	0.58	0.43	0.52	0.47	0.72	0.66	0.69	0.45	0.45	0.45	0.54
	SenticNet	0.19	0.19	1.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.11
	Sentim.140	0.48	0.28	0.60	0.38	0.64	0.63	0.64	0.44	0.09	0.15	0.39
	Stanf. DM	0.57	0.46	0.30	0.36	0.64	0.80	0.71	0.42	0.32	0.36	0.48
	Umigon	0.54	0.53	0.47	0.50	0.72	0.55	0.62	0.36	0.56	0.44	0.52
	Vader	0.44	0.37	0.41	0.39	0.86	0.31	0.46	0.33	0.72	0.45	0.43
	Combined I	0.52	0.47	0.41	0.44	0.48	0.79	0.60	0.62	0.37	0.47	0.50
Myspace	AFINN	0.40	0.82	0.35	0.49	0.30	0.32	0.31	0.22	0.65	0.33	0.38
	Emolex	0.26	0.87	0.10	0.18	0.38	0.04	0.07	0.21	0.95	0.34	0.20
	Emotic.	0.67	0.68	0.97	0.80	0.43	0.02	0.04	0.32	0.06	0.10	0.31
	Emot. DS	0.56	0.73	0.68	0.70	0.00	0.00	0.00	0.27	0.51	0.35	0.35
	H. Index	0.55	0.64	0.77	0.70	0.22	0.27	0.24	0.46	0.27	0.34	0.43
	LIWC	0.33	0.83	0.24	0.37	0.21	0.80	0.33	0.21	0.34	0.26	0.32
	NRC H.	0.39	0.86	0.30	0.45	0.23	0.24	0.24	0.24	0.76	0.37	0.35
	O.Finder	0.41	0.81	0.37	0.51	0.32	0.34	0.33	0.21	0.60	0.32	0.39
	Opin. Lex.	0.28	0.95	0.12	0.21	0.41	0.05	0.09	0.21	0.97	0.35	0.22
	PANAS-t	0.60	0.83	0.69	0.75	0.31	0.40	0.35	0.32	0.45	0.37	0.49
	Pattern.en	0.46	0.81	0.45	0.58	0.29	0.27	0.28	0.24	0.63	0.35	0.40
	SANN	0.49	0.67	0.57	0.62	0.00	0.00	0.00	0.24	0.50	0.32	0.31
	SASA	0.54	0.85	0.55	0.67	0.36	0.48	0.41	0.27	0.55	0.36	0.48
	SO-CAL	0.56	0.76	0.68	0.72	0.20	0.36	0.26	0.31	0.27	0.28	0.42
	SWN	0.69	0.84	0.81	0.83	0.36	0.60	0.45	0.45	0.31	0.37	0.55
	S.Strength	0.67	0.67	1.00	0.81	0.00	0.00	0.00	0.00	0.00	0.00	0.27
	SenticNet	0.58	0.75	0.73	0.74	0.28	0.58	0.37	0.20	0.08	0.12	0.41
	Sentim.140	0.35	0.89	0.28	0.43	0.17	0.77	0.27	0.33	0.34	0.34	0.35
	Stanf. DM	0.56	0.89	0.59	0.71	0.26	0.52	0.34	0.34	0.51	0.40	0.49
	Umigon	0.60	0.88	0.63	0.73	0.62	0.29	0.39	0.31	0.71	0.43	0.52
	Vader	0.60	0.88	0.63	0.73	0.62	0.29	0.39	0.31	0.71	0.43	0.52
	Combined I	0.59	0.61	0.88	0.72	0.42	0.44	0.43	0.66	0.31	0.42	0.52

Table A.7. Prediction performance of all methods in RW and Tweets_RND_I datasets, including the combined method

Dataset	Method	Acc.	Posit. sentiment			Negat. sentiment			Neut. sentiment			MacroF1
			P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	
RW	AFINN	0.52	0.60	0.71	0.65	0.51	0.32	0.39	0.40	0.39	0.40	0.48
	Emolex	0.44	0.56	0.49	0.52	0.37	0.38	0.38	0.36	0.42	0.39	0.43
	Emotic.	0.37	0.65	0.17	0.27	0.43	0.05	0.08	0.33	0.87	0.48	0.28
	Emot. DS	0.47	0.47	0.99	0.64	0.67	0.01	0.02	0.54	0.02	0.04	0.23
	H. Index	0.45	0.48	0.80	0.60	0.00	0.00	0.00	0.36	0.26	0.30	0.30
	LIWC	0.54	0.80	0.59	0.68	0.30	0.55	0.39	0.32	0.41	0.36	0.48
	NRC H.	0.29	0.70	0.13	0.23	0.24	0.83	0.37	0.31	0.17	0.22	0.27
	O.Finder	0.39	0.57	0.39	0.46	0.28	0.43	0.34	0.35	0.38	0.36	0.39
	Opin. Lex.	0.49	0.60	0.62	0.61	0.44	0.36	0.40	0.36	0.39	0.37	0.46
	PANAS-t	0.35	0.54	0.10	0.17	0.27	0.08	0.13	0.33	0.87	0.48	0.26
	Pattern.en	0.48	0.60	0.67	0.63	0.33	0.53	0.41	0.39	0.16	0.23	0.42
	SANN	0.47	0.57	0.63	0.60	0.38	0.34	0.36	0.35	0.32	0.33	0.43
	SASA	0.42	0.48	0.58	0.53	0.00	0.00	0.00	0.35	0.48	0.41	0.31
	SO-CAL	0.49	0.63	0.63	0.63	0.37	0.55	0.44	0.38	0.26	0.31	0.46
	SWN	0.46	0.46	0.98	0.63	0.00	0.00	0.00	0.45	0.03	0.06	0.23
	S.Strength	0.49	0.67	0.63	0.65	0.32	0.61	0.42	0.44	0.21	0.29	0.45
	SenticNet	0.46	0.46	1.00	0.63	0.00	0.00	0.00	0.00	0.00	0.00	0.21
	Sentim.140	0.46	0.60	0.64	0.62	0.34	0.77	0.47	0.33	0.02	0.03	0.37
	Stanf. DM	0.32	0.79	0.16	0.26	0.24	0.91	0.38	0.47	0.16	0.24	0.29
	Umigon	0.42	0.71	0.40	0.51	0.28	0.68	0.39	0.41	0.28	0.33	0.41
	Vader	0.53	0.63	0.69	0.66	0.63	0.25	0.36	0.39	0.49	0.44	0.48
	Combined I	0.52	0.70	0.64	0.67	0.50	0.43	0.46	0.29	0.38	0.33	0.49
Tweets _RND_I	AFINN	0.51	0.48	0.35	0.41	0.44	0.33	0.38	0.53	0.70	0.61	0.46
	Emolex	0.50	0.69	0.15	0.25	0.49	0.06	0.10	0.48	0.95	0.64	0.33
	Emotic.	0.33	0.32	0.99	0.49	0.87	0.01	0.03	0.67	0.04	0.07	0.19
	Emot. DS	0.45	0.37	0.62	0.46	0.00	0.00	0.00	0.55	0.56	0.55	0.34
	H. Index	0.41	0.43	0.35	0.39	0.18	0.30	0.23	0.50	0.49	0.49	0.37
	LIWC	0.45	0.51	0.26	0.35	0.34	0.74	0.46	0.59	0.45	0.51	0.44
	NRC H.	0.55	0.62	0.32	0.42	0.50	0.31	0.38	0.54	0.83	0.65	0.48
	O.Finder	0.56	0.56	0.43	0.48	0.56	0.35	0.43	0.56	0.75	0.64	0.52
	Opin. Lex.	0.48	0.69	0.07	0.12	0.48	0.06	0.11	0.47	0.96	0.63	0.29
	PANAS-t	0.53	0.50	0.69	0.58	0.46	0.49	0.47	0.63	0.45	0.53	0.53
	Pattern.en	0.53	0.50	0.44	0.47	0.53	0.29	0.37	0.55	0.72	0.62	0.49
	SANN	0.46	0.40	0.41	0.41	0.39	0.35	0.37	0.53	0.55	0.54	0.44
	SASA	0.58	0.55	0.55	0.55	0.55	0.48	0.51	0.61	0.64	0.62	0.56
	SO-CAL	0.32	0.32	0.98	0.48	0.00	0.00	0.00	0.59	0.03	0.06	0.18
	SWN	0.58	0.55	0.70	0.62	0.49	0.59	0.54	0.69	0.49	0.57	0.58
	S.Strength	0.32	0.32	1.00	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.16
	SenticNet	0.41	0.40	0.70	0.51	0.39	0.70	0.50	0.76	0.08	0.14	0.38
	Sentim.140	0.31	0.65	0.24	0.35	0.24	0.88	0.38	0.49	0.08	0.14	0.29
	Stanf. DM	0.61	0.64	0.57	0.61	0.50	0.52	0.51	0.64	0.67	0.65	0.59
	Umigon	0.58	0.62	0.54	0.58	0.65	0.24	0.35	0.56	0.78	0.65	0.53
	Vader	0.58	0.62	0.54	0.58	0.65	0.24	0.35	0.56	0.78	0.65	0.53
	Combined I	0.62	0.56	0.65	0.60	0.41	0.62	0.49	0.76	0.60	0.67	0.59

Table A.8. Prediction performance of all methods in Comments_YTB and Tweets_STF datasets, including the combined method

Dataset	**Method**	**Acc.**	**Posit. sentiment**			**Negat. sentiment**			**Neut. sentiment**			**MacroF1**
			P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	
Comments_YTB	AFINN	0.54	0.69	0.61	0.64	0.59	0.34	0.43	0.38	0.58	0.46	0.51
	Emolex	0.43	0.64	0.37	0.47	0.47	0.34	0.40	0.32	0.61	0.42	0.43
	Emotic.	0.33	0.75	0.11	0.19	0.37	0.02	0.04	0.29	0.94	0.45	0.23
	Emot.DS	0.48	0.49	0.93	0.64	0.68	0.02	0.05	0.28	0.07	0.11	0.27
	H.Index	0.43	0.51	0.58	0.55	0.00	0.00	0.00	0.33	0.51	0.40	0.32
	LIWC	0.41	0.53	0.53	0.53	0.17	0.27	0.21	0.40	0.31	0.35	0.36
	NRC H.	0.37	0.70	0.21	0.33	0.34	0.72	0.46	0.27	0.35	0.30	0.36
	O.Finder	0.42	0.70	0.31	0.43	0.42	0.32	0.36	0.33	0.70	0.45	0.41
	Opin. Lex.	0.48	0.69	0.46	0.55	0.54	0.36	0.43	0.34	0.62	0.44	0.47
	PANAS-t	0.31	0.70	0.05	0.09	0.49	0.04	0.08	0.29	0.96	0.45	0.20
	Pattern.en	0.58	0.71	0.73	0.72	0.48	0.48	0.48	0.42	0.39	0.40	0.53
	SANN	0.49	0.67	0.52	0.59	0.48	0.29	0.36	0.36	0.62	0.46	0.47
	SASA	0.47	0.52	0.73	0.60	0.00	0.00	0.00	0.36	0.39	0.37	0.33
	SO-CAL	0.57	0.74	0.62	0.68	0.54	0.52	0.53	0.40	0.53	0.46	0.55
	SWN	0.47	0.49	0.89	0.63	0.00	0.00	0.00	0.31	0.12	0.17	0.27
	S.Strength	0.61	0.75	0.75	0.75	0.49	0.61	0.54	0.47	0.38	0.42	0.57
	SenticNet	0.49	0.49	1.00	0.66	0.00	0.00	0.00	0.00	0.00	0.00	0.22
	Sentim.140	0.47	0.59	0.65	0.62	0.35	0.59	0.44	0.25	0.07	0.11	0.39
	Stanf. DM	0.47	0.82	0.47	0.60	0.33	0.72	0.45	0.35	0.27	0.30	0.45
	Umigon	0.57	0.79	0.62	0.70	0.44	0.51	0.47	0.43	0.54	0.48	0.55
	Vader	0.56	0.78	0.59	0.67	0.68	0.30	0.41	0.39	0.72	0.51	0.53
	Combined I	0.58	0.64	0.76	0.70	0.47	0.56	0.51	0.57	0.40	0.47	0.56
Tweets_STF	AFINN	0.53	0.76	0.62	0.68	0.88	0.45	0.59	0.00	0.00	0.00	0.64
	Emolex	0.37	0.70	0.41	0.51	0.82	0.33	0.47	0.00	0.00	0.00	0.49
	Emotic.	0.11	0.83	0.14	0.24	0.94	0.09	0.16	0.00	0.00	0.00	0.20
	Emot. DS	0.52	0.53	1.00	0.69	1.00	0.03	0.05	0.00	0.00	0.00	0.37
	H. Index	0.42	0.59	0.55	0.57	0.89	0.27	0.42	0.00	0.00	0.00	0.50
	LIWC	0.46	0.57	0.64	0.60	0.34	0.65	0.45	0.00	0.00	0.00	0.53
	NRC H.	0.50	0.81	0.24	0.37	0.76	0.77	0.77	0.00	0.00	0.00	0.57
	O.Finder	0.35	0.81	0.31	0.45	0.80	0.40	0.53	0.00	0.00	0.00	0.49
	Opin. Lex.	0.46	0.77	0.50	0.61	0.93	0.42	0.58	0.00	0.00	0.00	0.60
	PANAS-t	0.07	0.80	0.07	0.12	0.86	0.07	0.13	0.00	0.00	0.00	0.13
	Pattern.en	0.67	0.76	0.75	0.75	0.81	0.58	0.67	0.00	0.00	0.00	0.71
	SANN	0.43	0.69	0.47	0.56	0.78	0.39	0.52	0.00	0.00	0.00	0.54
	SASA	0.30	0.50	0.60	0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.28
	SO-CAL	0.67	0.83	0.69	0.75	0.93	0.66	0.77	0.00	0.00	0.00	0.76
	SWN	0.49	0.51	0.97	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.34
	S.Strength	0.69	0.82	0.68	0.74	0.84	0.69	0.76	0.00	0.00	0.00	0.75
	SenticNet	0.51	0.51	1.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.34
	Sentim.140	0.75	0.77	0.71	0.74	0.75	0.78	0.76	0.00	0.00	0.00	0.75
	Stanf. DM	0.60	0.88	0.31	0.46	0.61	0.89	0.73	0.00	0.00	0.00	0.60
	Umigon	0.71	0.92	0.67	0.77	0.83	0.75	0.79	0.00	0.00	0.00	0.78
	Vader	0.45	0.88	0.54	0.67	0.91	0.36	0.51	0.00	0.00	0.00	0.59
	Combined I	0.60	0.63	0.86	0.73	0.56	0.93	0.70	0.00	0.00	0.00	0.72

Table A.9. Prediction performance of all methods in Amazon and Reviews_II datasets, including the combined method

Dataset	Method	Acc.	Posit. sentiment			Negat. sentiment			Neut. sentiment			MacroF1
			P	R	F1	P	R	F1	P	R	F1	
Amazon	AFINN	0.38	0.76	0.48	0.59	0.68	0.20	0.31	0.04	0.80	0.08	0.33
	Emolex	0.33	0.68	0.38	0.48	0.52	0.24	0.33	0.04	0.68	0.07	0.29
	Emotic.	0.03	0.00	0.00	0.00	1.00	0.00	0.00	0.03	1.00	0.05	0.02
	Emot. DS	0.57	0.58	0.97	0.73	0.74	0.01	0.02	0.10	0.13	0.11	0.29
	H. Index	0.30	0.61	0.49	0.54	0.00	0.00	0.00	0.04	0.73	0.07	0.20
	LIWC	0.34	0.47	0.63	0.54	0.14	0.48	0.22	0.47	0.03	0.06	0.27
	NRC H.	0.40	0.77	0.22	0.35	0.51	0.66	0.57	0.04	0.46	0.07	0.33
	O.Finder	0.27	0.78	0.28	0.41	0.54	0.22	0.31	0.04	0.85	0.07	0.26
	Opin. Lex.	0.42	0.78	0.53	0.63	0.67	0.25	0.37	0.05	0.80	0.09	0.36
	PANAS-t	0.05	0.89	0.04	0.07	0.54	0.00	0.01	0.03	1.00	0.05	0.04
	Pattern.en	0.55	0.76	0.62	0.68	0.62	0.44	0.51	0.06	0.53	0.10	0.43
	SANN	0.33	0.73	0.43	0.54	0.65	0.17	0.27	0.04	0.82	0.07	0.29
	SASA	0.40	0.59	0.68	0.63	0.00	0.00	0.00	0.04	0.50	0.07	0.23
	SO-CAL	0.56	0.80	0.65	0.72	0.71	0.44	0.54	0.06	0.66	0.11	0.46
	SWN	0.56	0.57	0.97	0.72	0.00	0.00	0.00	0.08	0.09	0.08	0.27
	S.Strength	0.45	0.82	0.47	0.60	0.64	0.38	0.48	0.05	0.81	0.09	0.39
	SenticNet	0.57	0.57	1.00	0.73	0.00	0.00	0.00	0.00	0.00	0.00	0.24
	Sentim.140	0.59	0.72	0.52	0.60	0.50	0.72	0.59	0.18	0.08	0.11	0.44
	Stanf. DM	0.55	0.88	0.45	0.59	0.55	0.70	0.62	0.07	0.51	0.12	0.44
	Umigon	0.38	0.85	0.39	0.53	0.57	0.35	0.43	0.04	0.80	0.08	0.35
	Vader	0.29	0.88	0.38	0.53	0.74	0.10	0.18	0.04	0.95	0.07	0.26
	Combined I	0.45	0.53	0.85	0.65	0.30	0.73	0.42	0.83	0.05	0.09	0.39
Reviews_II	AFINN	0.38	0.62	0.50	0.55	0.69	0.26	0.38	0.00	0.38	0.01	0.31
	Emolex	0.41	0.60	0.49	0.54	0.63	0.33	0.44	0.00	0.35	0.01	0.33
	Emotic.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.01	0.00
	Emot.DS	0.49	0.50	0.99	0.66	0.96	0.00	0.01	0.00	0.00	0.00	0.22
	H. Index	0.34	0.51	0.68	0.58	0.00	0.00	0.00	0.00	0.24	0.00	0.19
	LIWC	0.41	0.58	0.60	0.59	0.24	0.66	0.35	0.19	0.00	0.00	0.31
	NRC H.	0.39	0.66	0.22	0.33	0.57	0.56	0.57	0.00	0.32	0.01	0.30
	Opin.Finder	0.38	0.69	0.21	0.33	0.58	0.54	0.56	0.00	0.41	0.01	0.30
	Opin. Lex.	0.46	0.67	0.53	0.59	0.68	0.39	0.49	0.00	0.27	0.01	0.36
	PANAS-t	0.06	0.71	0.07	0.13	0.57	0.04	0.08	0.00	0.95	0.01	0.07
	Pattern.en	0.60	0.65	0.63	0.64	0.66	0.56	0.61	0.00	0.14	0.01	0.42
	SANN	0.43	0.62	0.51	0.56	0.63	0.36	0.46	0.00	0.27	0.01	0.34
	SASA	0.29	0.49	0.58	0.53	0.00	0.00	0.00	0.00	0.49	0.01	0.18
	SO-CAL	0.64	0.72	0.68	0.70	0.72	0.61	0.66	0.00	0.14	0.01	0.46
	SWN	0.49	0.49	0.99	0.66	0.00	0.00	0.00	0.00	0.00	0.00	0.22
	S.Strength	0.51	0.65	0.44	0.53	0.62	0.57	0.59	0.00	0.14	0.00	0.38
	SenticNet	0.49	0.49	1.00	0.66	0.00	0.00	0.00	0.00	0.00	0.00	0.22
	Sentim.140	0.61	0.59	0.77	0.67	0.67	0.47	0.55	0.00	0.00	0.00	0.41
	Stanf. DM	0.79	0.89	0.76	0.82	0.83	0.82	0.82	0.00	0.08	0.01	0.55
	Umigon	0.36	0.67	0.32	0.44	0.61	0.40	0.48	0.00	0.38	0.01	0.31
	Vader	0.31	0.71	0.44	0.55	0.74	0.18	0.29	0.00	0.51	0.01	0.28
	Combined I	0.52	0.60	0.74	0.66	0.45	0.75	0.56	0.32	0.00	0.01	0.41

Table A.10. Prediction performance of all methods in Comments_NYT and Tweets_RND_II datasets, including the combined method

Dataset	Method	Acc.	Posit. sentiment			Negat. sentiment			Neut. sentiment			MacroF1
			P	R	F1	P	R	F1	P	R	F1	
Comments_NYT	AFINN	0.35	0.63	0.35	0.45	0.81	0.30	0.44	0.07	0.82	0.13	0.34
	Emolex	0.37	0.53	0.42	0.47	0.72	0.32	0.44	0.07	0.62	0.12	0.34
	Emotic.	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.05	1.00	0.09	0.03
	Emot. DS	0.42	0.43	0.98	0.60	1.00	0.01	0.01	0.13	0.07	0.09	0.23
	H. Index	0.27	0.44	0.57	0.50	0.00	0.00	0.00	0.06	0.54	0.10	0.20
	LIWC	0.24	0.28	0.43	0.34	0.19	0.54	0.28	0.53	0.05	0.09	0.24
	NRC H.	0.45	0.56	0.18	0.27	0.59	0.68	0.63	0.08	0.42	0.13	0.34
	O.Finder	0.29	0.69	0.19	0.30	0.77	0.33	0.46	0.06	0.88	0.12	0.29
	Opin. Lex.	0.38	0.65	0.37	0.47	0.80	0.35	0.49	0.07	0.82	0.13	0.36
	PANAS-t	0.07	0.59	0.03	0.06	0.62	0.02	0.05	0.05	0.99	0.09	0.07
	Pattern.en	0.45	0.55	0.45	0.49	0.64	0.46	0.53	0.08	0.46	0.13	0.39
	SANN	0.28	0.57	0.29	0.39	0.78	0.22	0.34	0.06	0.79	0.11	0.28
	SASA	0.25	0.43	0.51	0.47	0.00	0.00	0.00	0.06	0.61	0.10	0.19
	SO-CAL	0.51	0.64	0.51	0.57	0.77	0.49	0.60	0.10	0.66	0.17	0.45
	SWN	0.42	0.43	0.99	0.60	0.00	0.00	0.00	0.16	0.06	0.08	0.23
	S.Strength	0.43	0.69	0.27	0.39	0.72	0.53	0.61	0.08	0.77	0.15	0.38
	SenticNet	0.42	0.42	1.00	0.60	0.00	0.00	0.00	0.00	0.00	0.00	0.20
	Sentim.140	0.59	0.54	0.63	0.58	0.65	0.60	0.62	0.20	0.06	0.09	0.43
	Stanf. DM	0.52	0.73	0.21	0.33	0.59	0.78	0.67	0.10	0.38	0.15	0.39
	Umigon	0.24	0.69	0.16	0.26	0.69	0.25	0.36	0.06	0.89	0.11	0.25
	Vader	0.23	0.74	0.22	0.34	0.87	0.17	0.29	0.06	0.93	0.11	0.24
	Combined I	0.37	0.32	0.72	0.44	0.37	0.82	0.51	0.86	0.07	0.13	0.36
Tweets_RND_II	AFINN	0.37	0.87	0.34	0.49	0.71	0.41	0.52	0.00	0.75	0.00	0.34
	Emolex	0.15	0.98	0.18	0.30	0.97	0.07	0.14	0.00	1.00	0.00	0.15
	Emotic.	0.69	0.71	0.96	0.82	0.98	0.07	0.13	0.00	0.00	0.00	0.32
	Emot.DS	0.40	0.69	0.58	0.63	0.00	0.00	0.00	0.00	0.75	0.00	0.21
	H. Index	0.49	0.59	0.74	0.66	0.28	0.44	0.34	0.25	0.00	0.00	0.33
	LIWC	0.38	0.89	0.23	0.37	0.44	0.70	0.54	0.00	0.50	0.00	0.31
	NRC H.	0.32	0.94	0.27	0.42	0.63	0.43	0.51	0.00	1.00	0.00	0.31
	O.Finder	0.43	0.94	0.42	0.58	0.81	0.45	0.58	0.00	1.00	0.00	0.39
	Opin. Lex.	0.08	0.96	0.08	0.14	0.76	0.09	0.16	0.00	1.00	0.00	0.10
	PANAS-t	0.70	0.93	0.73	0.82	0.74	0.62	0.68	0.00	1.00	0.01	0.50
	Pattern.en	0.44	0.90	0.46	0.61	0.72	0.40	0.51	0.00	1.00	0.00	0.38
	SANN	0.45	0.71	0.65	0.68	0.00	0.00	0.00	0.00	0.75	0.00	0.23
	SASA	0.59	0.94	0.57	0.71	0.77	0.64	0.70	0.00	0.75	0.00	0.47
	SO-CAL	0.68	0.69	0.99	0.81	0.00	0.00	0.00	0.00	0.00	0.00	0.27
	SWN	0.73	0.97	0.70	0.82	0.78	0.77	0.78	0.00	0.75	0.01	0.53
	S.Strength	0.69	0.69	1.00	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.27
	SenticNet	0.69	0.85	0.68	0.76	0.54	0.70	0.61	0.00	0.00	0.00	0.46
	Sentim.140	0.54	0.94	0.40	0.56	0.44	0.85	0.58	0.00	0.25	0.00	0.38
	Stanf. DM	0.63	0.98	0.62	0.76	0.74	0.64	0.68	0.00	1.00	0.01	0.48
	Umigon	0.60	0.99	0.63	0.77	0.99	0.52	0.68	0.00	1.00	0.00	0.49
	Vader	0.60	0.99	0.63	0.77	0.99	0.52	0.68	0.00	1.00	0.00	0.49
	Combined I	0.64	0.63	0.98	0.77	0.64	0.87	0.74	1.00	0.00	0.01	0.50

Table A.11. Prediction performance of all methods in YLP and Tweets_SemEval datasets, including the combined method

Dataset	Method	Acc.	Posit. sentiment			Negat. sentiment			Neut. sentiment			MacroF1
			P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	
	AFINN	0.66	0.68	0.94	0.79	0.95	0.38	0.55	0.00	0.00	0.00	0.67
	Emolex	0.62	0.62	0.85	0.72	0.84	0.38	0.53	0.00	0.00	0.00	0.63
	Emotic.	0.04	0.74	0.06	0.11	0.76	0.01	0.03	0.00	0.00	0.00	0.07
	Emot. DS	0.50	0.50	1.00	0.67	0.83	0.00	0.00	0.00	0.00	0.00	0.34
	Happ. Index	0.52	0.53	0.94	0.67	0.86	0.10	0.18	0.00	0.00	0.00	0.43
	LIWC	0.58	0.86	0.61	0.71	0.29	0.79	0.42	0.00	0.00	0.00	0.57
	NRC H.	0.55	0.95	0.15	0.26	0.59	0.95	0.73	0.00	0.00	0.00	0.50
	O.Finder	0.50	0.61	0.45	0.52	0.55	0.55	0.55	0.00	0.00	0.00	0.54
YLP	Opin. Lex.	0.68	0.71	0.90	0.80	0.94	0.46	0.62	0.00	0.00	0.00	0.71
	PANAS-t	0.21	0.70	0.31	0.43	0.80	0.12	0.21	0.00	0.00	0.00	0.32
	Pattern.en	0.84	0.80	0.93	0.86	0.92	0.75	0.83	0.00	0.00	0.00	0.85
	SANN	0.68	0.67	0.90	0.77	0.88	0.46	0.60	0.00	0.00	0.00	0.69
	SASA	0.28	0.50	0.57	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.27
	SO-CAL	0.84	0.83	0.93	0.88	0.94	0.75	0.84	0.00	0.00	0.00	0.86
	SWN	0.50	0.50	1.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.34
	S.Strength	0.81	0.84	0.81	0.82	0.81	0.80	0.81	0.00	0.00	0.00	0.82
	SenticNet	0.50	0.50	1.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.34
	Sentim.140	0.83	0.83	0.84	0.84	0.84	0.82	0.83	0.00	0.00	0.00	0.84
	Stanf. DM	0.67	0.95	0.41	0.57	0.64	0.93	0.76	0.00	0.00	0.00	0.67
	Umigon	0.63	0.87	0.45	0.59	0.62	0.82	0.70	0.00	0.00	0.00	0.65
	Vader	0.65	0.74	0.94	0.83	0.98	0.36	0.53	0.00	0.00	0.00	0.68
	Combined I	0.84	0.95	0.81	0.87	0.73	0.95	0.82	0.00	0.00	0.00	0.85
	AFINN	0.60	0.60	0.61	0.60	0.44	0.41	0.42	0.64	0.65	0.65	0.56
	Emolex	0.51	0.50	0.36	0.42	0.34	0.37	0.35	0.56	0.65	0.60	0.46
	Emotic.	0.53	0.73	0.11	0.19	0.56	0.05	0.10	0.52	0.97	0.67	0.32
	Emot. DS	0.37	0.37	1.00	0.54	0.50	0.00	0.00	0.50	0.00	0.00	0.18
	H. Index	0.47	0.41	0.68	0.51	0.00	0.00	0.00	0.58	0.45	0.51	0.34
	LIWC	0.40	0.42	0.37	0.39	0.13	0.15	0.14	0.45	0.49	0.47	0.33
	NRC H.	0.38	0.53	0.22	0.31	0.21	0.72	0.32	0.55	0.41	0.47	0.36
Tweets	O.Finder	0.58	0.68	0.28	0.40	0.41	0.34	0.37	0.58	0.86	0.69	0.49
_SemEval	Opin. Lex.	0.60	0.63	0.47	0.54	0.44	0.40	0.42	0.62	0.75	0.68	0.55
	PANAS-t	0.54	0.85	0.12	0.21	0.46	0.06	0.10	0.52	0.98	0.68	0.33
	Pattern.en	0.50	0.58	0.68	0.63	0.25	0.56	0.34	0.68	0.35	0.46	0.48
	SANN	0.55	0.53	0.47	0.50	0.39	0.30	0.34	0.59	0.67	0.63	0.49
	SASA	0.50	0.43	0.55	0.48	0.00	0.00	0.00	0.56	0.61	0.59	0.36
	SO-CAL	0.59	0.59	0.59	0.59	0.40	0.54	0.46	0.66	0.60	0.63	0.56
	SWN	0.37	0.37	1.00	0.53	0.00	0.00	0.00	0.33	0.00	0.00	0.18
	S.Strength	0.58	0.61	0.68	0.65	0.31	0.64	0.42	0.77	0.49	0.60	0.55
	SenticNet	0.37	0.37	1.00	0.54	0.00	0.00	0.00	0.00	0.00	0.00	0.18
	Sentim.140	0.36	0.43	0.71	0.54	0.25	0.72	0.37	0.00	0.00	0.00	0.30
	Stanf. DM	0.23	0.72	0.18	0.29	0.15	0.91	0.26	0.47	0.07	0.12	0.22
	Umigon	0.66	0.75	0.56	0.64	0.40	0.56	0.46	0.71	0.76	0.73	0.61
	Vader	0.64	0.70	0.57	0.63	0.52	0.27	0.36	0.62	0.79	0.70	0.56
	Combined I	0.69	0.60	0.76	0.67	0.51	0.54	0.53	0.81	0.69	0.75	0.65

Appendix B

Complete Results of Percentage of Agreement

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtah	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	45	45	11	41	31	48	49	46	38	46	35	43	22	43	20	14	14	45	52	37
Emolex	45	100	49	8	42	34	50	50	50	35	44	36	44	19	37	16	12	12	40	49	37
Emotic.	45	49	100	4	45	33	56	51	64	34	49	42	41	15	36	10	4	7	43	54	42
Emotic. DS	11	8	4	100	9	5	6	9	4	13	9	7	11	12	13	15	11	4	11	10	11
Happ. Index	41	42	45	9	100	28	45	43	46	33	42	34	38	19	34	17	10	9	39	45	35
NRC Hashtag	31	34	33	5	28	100	34	34	33	25	30	24	30	16	29	13	14	16	31	33	26
Opin. Finder	48	50	56	6	45	34	100	52	57	37	51	39	45	19	40	15	9	11	45	54	40
Opin. Lexicon	49	50	51	9	43	34	52	100	52	38	48	38	47	21	41	18	13	13	44	52	39
PANAS	46	50	64	4	46	33	57	52	100	34	49	42	42	15	36	10	4	7	43	54	42
Pattern	38	35	34	13	33	25	37	38	34	100	36	28	40	23	37	21	16	15	39	40	31
SANN	46	44	49	9	42	30	51	48	49	36	100	36	42	20	41	17	11	11	43	51	37
SASA	35	36	42	7	34	24	39	38	42	28	36	100	32	14	31	11	8	7	34	40	31
SO-CAL	43	44	41	11	38	30	45	47	42	40	42	32	100	23	40	21	17	16	42	47	36
SWN	22	19	15	12	19	16	19	21	15	23	20	14	23	100	22	19	14	12	21	21	18
SentiStrength	43	37	36	13	34	29	40	41	36	37	41	31	40	22	100	21	17	16	43	45	33
SenticNet	20	16	10	15	17	13	15	18	10	21	17	11	21	19	21	100	15	11	20	19	17
Sentim.140	14	12	4	11	10	14	9	13	4	16	11	8	17	14	17	15	100	15	16	13	13
Stanford DM	14	12	7	4	9	16	11	13	7	15	11	7	16	12	16	11	15	100	16	12	12
Umigon	45	40	43	11	39	31	45	44	43	39	43	34	42	21	43	20	16	16	100	48	37
VADER	52	49	54	10	45	33	54	52	54	40	51	40	47	21	45	19	13	12	48	100	41
LIWC	37	37	42	11	35	26	40	39	42	31	37	31	36	18	33	17	13	12	37	41	100

Figure B.1. Percentage of agreement among all methods in two labeled datasets: Tweets_SAN.

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtag	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	32	38	21	28	28	33	39	25	36	32	24	37	27	37	26	26	20	42	43	22
Emolex	32	100	38	16	28	26	34	38	30	29	30	25	36	22	32	21	21	15	36	36	29
Emotic.	38	38	100	24	32	37	39	42	39	55	38	35	43	30	42	27	37	27	62	51	36
Emotic. DS	21	16	24	100	19	9	13	17	5	29	16	16	20	23	25	25	25	12	29	25	2
Happ. Index	28	28	32	19	100	18	28	29	24	28	27	23	30	24	28	23	21	13	34	34	22
NRC Hashtag	28	26	37	9	18	100	26	29	23	29	22	22	29	18	27	15	25	21	36	29	21
Opin. Finder	33	34	39	13	28	26	100	37	35	28	35	26	36	20	32	19	18	16	37	38	32
Opin. Lexicon	39	38	42	17	29	29	37	100	33	33	33	27	40	24	35	23	24	17	41	42	31
PANAS	25	30	39	5	24	23	35	33	100	20	30	24	30	13	23	10	8	7	31	33	43
Pattern	36	29	55	29	28	29	28	33	20	100	29	28	38	32	41	34	41	31	57	43	15
SANN	32	30	38	16	27	22	35	33	30	29	100	23	33	22	31	20	20	16	37	37	27
SASA	24	25	35	16	23	22	26	27	24	28	23	100	28	20	27	18	20	15	35	31	22
SO-CAL	37	36	43	20	30	29	36	40	30	38	33	28	100	27	39	28	29	23	43	44	26
SWN	27	22	30	23	24	18	20	24	13	32	22	20	27	100	29	27	26	18	33	30	9,2
SentiStrength	37	32	42	25	28	27	32	35	23	41	31	27	39	29	100	29	32	23	45	42	19
SenticNet	26	21	27	25	23	15	19	23	9,6	34	20	18	28	27	29	100	27	20	33	29	5,4
Sentim.140	26	21	37	25	21	25	18	24	8	41	20	20	29	26	32	27	100	27	42	30	4
Stanford DM	20	15	27	12	13	21	16	17	7	31	16	15	23	18	23	20	27	100	30	23	4
Umigon	42	36	62	29	34	36	37	41	31	57	37	35	43	33	45	33	42	30	100	49	26
VADER	43	36	51	25	34	29	38	42	33	43	37	31	44	30	42	29	30	23	49	100	31
LIWC	22	29	36	2	22	21	32	31	43	15	27	22	26	9	19	5	4	4	26	31	100

(a) Percentage of agreement on Tweets_RND_IV dataset

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtag	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	28	23	11	25	21	27	34	24	20	27	20	29	18	28	19	16	18	29	32	8
Emolex	28	100	22	9	23	21	26	29	23	17	24	18	27	17	25	16	16	18	25	26	8
Emotic.	23	22	100	2	23	16	28	26	37	9	25	17	21	6	17	6	2	8	27	29	1
Emotic. DS	11	9	2	100	9	4	6	10	2	11	8	11	11	14	10	16	16	6	8	9	5
Happ. Index	25	23	23	9	100	17	23	25	24	15	23	17	24	15	23	15	11	13	24	26	5
NRC Hashtag	21	21	16	4	17	100	23	23	17	21	20	18	24	16	23	13	22	27	21	20	15
Opin. Finder	27	26	28	6	23	23	100	30	29	17	30	20	29	14	26	14	12	18	29	29	7
Opin. Lexicon	34	29	26	10	25	23	30	100	26	20	28	20	32	18	28	18	17	19	28	31	8
PANAS	24	23	37	2	24	17	29	26	100	10	26	17	23	7	19	7	4	9	27	30	2
Pattern	20	17	9	11	15	21	17	20	10	100	17	17	24	19	23	19	22	24	21	17	12
SANN	27	24	25	8	23	20	30	28	26	17	100	19	26	14	25	15	12	16	28	28	6
SASA	20	18	17	11	17	18	20	20	17	17	19	100	20	15	20	14	16	16	21	21	9
SO-CAL	29	27	21	11	24	24	29	32	23	24	26	20	100	19	31	20	20	24	29	29	10
SWN	18	17	6	14	15	16	14	18	7	19	14	15	19	100	19	19	19	18	16	15	10
SentiStrength	28	25	17	10	23	23	26	28	19	23	25	20	31	19	100	20	19	24	29	27	11
SenticNet	19	16	6	16	15	13	14	18	7	19	15	14	20	19	20	100	18	16	16	15	9
Sentim.140	16	16	2	16	11	22	12	17	4	22	12	16	20	19	19	18	100	25	14	12	16
Stanford DM	18	18	8	6	13	27	18	19	9	24	16	16	24	18	24	16	25	100	19	16	16
Umigon	29	25	27	8	24	21	29	28	27	21	28	21	29	16	29	16	14	19	100	30	8
VADER	32	26	29	9	26	20	29	31	30	17	28	21	29	15	27	15	12	16	30	100	6
LIWC	8	8	1	5	5	15	7	8	2	12	6	9	10	10	11	9	16	16	8	6	100

(b) Percentage of agreement on Tweets_DBT dataset

Figure B.2. Percentage of agreement among all methods in two labeled datasets: Tweets_RND_IV and Tweets_DBT.

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtag	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	46	47	14	41	34	52	54	47	41	50	38	48	24	49	25	19	21	54	58	32
Emolex	46	100	51	10	40	37	54	54	52	35	49	36	49	21	43	20	16	20	50	51	32
Emotic.	47	51	100	5	41	36	61	55	65	35	54	40	47	14	42	13	8	15	55	57	38
Emotic. DS	14	10	5	100	14	7	8	11	4	16	11	11	15	15	17	18	15	9	15	14	8
Happ. Index	41	40	41	14	100	28	43	43	42	34	42	32	40	21	38	22	15	17	43	46	27
NRC Hashtag	34	37	36	7	28	100	40	40	36	26	36	28	36	17	34	16	18	20	40	37	23
Opin. Finder	52	54	61	8	43	40	100	60	62	38	58	42	52	20	47	18	14	20	57	59	36
Opin. Lexicon	54	54	55	11	43	40	60	100	56	41	55	40	54	23	49	22	18	22	56	58	35
PANAS	47	52	65	4	42	36	62	56	100	34	54	40	47	13	42	13	8	16	54	57	38
Pattern	41	35	35	16	34	26	38	41	34	100	38	31	43	25	41	26	20	21	44	44	25
SANN	50	49	54	11	42	36	58	55	54	38	100	38	48	22	47	21	15	19	53	56	34
SASA	38	36	40	11	32	28	42	40	40	31	38	100	38	18	37	18	15	16	42	43	27
SO-CAL	48	49	47	15	40	36	52	54	47	43	48	38	100	25	48	26	21	24	52	54	32
SWN	24	21	14	15	21	17	20	23	13	25	22	18	25	100	25	22	18	15	25	25	13
SentiStrength	49	43	42	17	38	34	47	49	42	41	47	37	48	25	100	28	22	25	53	53	31
SenticNet	25	20	13	18	22	16	18	22	13	26	21	18	26	22	28	100	19	16	26	25	14
Sentim.140	19	16	8	15	15	18	14	18	8	20	15	15	21	18	22	19	100	17	22	18	10
Stanford DM	21	20	15	9	17	20	20	22	16	21	19	16	24	15	25	16	17	100	25	22	12
Umigon	54	50	55	15	43	40	57	56	54	44	53	42	52	25	53	26	22	25	100	61	36
VADER	58	51	57	14	46	37	59	58	57	44	56	43	54	25	53	25	18	22	61	100	37
LIWC	32	32	38	8	27	23	36	35	38	25	34	27	32	13	31	14	10	12	36	37	100

(a) Percentage of agreement on Tweets_RDN_III dataset

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtag	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	36	17	21	28	32	32	40	19	40	31	26	40	32	37	31	32	35	30	35	41
Emolex	36	100	12	16	22	28	22	30	14	28	25	20	33	25	30	25	30	31	23	25	32
Emotic.	17	12	100	4	16	10	20	17	20	16	17	10	15	10	11	9	5	6	14	20	16
Emotic. DS	21	16	4	100	20	9	12	15	4	25	16	16	21	26	16	27	22	15	15	16	22
Happ. Index	28	22	16	20	100	14	25	22	17	27	22	19	25	23	20	26	19	19	22	26	27
NRC Hashtag	32	28	10	9	14	100	21	30	11	27	21	25	36	25	37	17	37	42	26	20	28
Opin. Finder	32	22	20	12	25	21	100	31	21	28	28	20	28	23	26	21	16	21	26	33	30
Opin. Lexicon	40	30	17	15	22	30	31	100	17	32	30	22	37	26	31	22	26	27	27	31	36
PANAS	19	14	20	4	17	11	21	17	100	17	17	11	16	10	12	10	6	7	15	21	17
Pattern	40	28	16	25	27	27	28	32	17	100	32	28	38	33	36	35	32	31	30	35	38
SANN	31	25	17	16	22	21	28	30	17	32	100	17	31	25	30	22	20	22	23	33	31
SASA	26	20	10	16	19	25	20	22	11	28	17	100	27	25	25	23	23	27	22	20	23
SO-CAL	40	33	15	21	25	36	28	37	16	38	31	27	100	28	37	28	37	37	30	30	38
SWN	32	25	10	26	23	25	23	26	10	33	25	25	28	100	35	33	26	27	23	27	30
SentiStrength	37	30	11	16	20	37	26	31	12	36	30	25	37	35	100	25	28	41	28	28	33
SenticNet	31	25	8,6	27	26	17	21	22	9,9	35	22	23	28	33	25	100	23	23	25	26	31
Sentim.140	32	30	5	22	19	37	16	26	6	32	20	23	37	26	28	23	100	37	20	20	28
Stanford DM	35	31	6	15	19	42	21	27	7	31	22	27	37	27	41	23	37	100	26	21	31
Umigon	30	23	14	15	22	26	26	27	15	30	23	22	30	23	28	25	20	26	100	26	33
VADER	35	25	20	16	26	20	33	31	21	35	33	20	30	27	28	26	20	21	26	100	33
LIWC	41	32	16	22	27	28	30	36	17	38	31	23	38	30	33	31	28	31	33	33	100

(b) Percentage of agreement on Irony dataset

Figure B.3. Percentage of agreement among all methods in two labeled datasets: Tweets_RDN_III and Irony.

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtah	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	27	8	25	25	21	24	31	9	29	28	21	32	27	33	29	28	27	29	30	15
Emolex	27	100	7	18	23	25	23	30	9	25	23	17	28	26	31	24	28	27	23	25	13
Emotic.	8	7	100	1	7	5	7	8	13	4	6	5	5	1	7	1	2	3	10	11	6
Emotic. DS	25	18	1	100	23	11	14	20	3	25	20	18	25	26	19	32	26	18	18	18	14
Happ. Index	25	23	7	23	100	18	18	24	8	24	23	20	26	24	25	27	24	21	23	21	14
NRC Hashtag	21	25	5	11	18	100	23	24	7	24	26	21	28	21	31	19	30	32	20	21	13
Opin. Finder	24	23	7	14	18	23	100	27	9	28	30	19	29	23	31	21	23	28	20	23	13
Opin. Lexicon	31	30	8	20	24	24	27	100	9	28	26	18	35	26	32	25	27	29	24	27	14
PANAS	9	9	13	3	8	7	9	9	100	7	7	7	8	3	9	4	4	5	11	12	7
Pattern	29	25	4	25	24	24	28	28	7	100	33	23	36	31	34	32	32	35	27	27	17
SANN	28	23	6	20	23	26	30	26	7	33	100	26	32	28	34	27	27	33	26	26	16
SASA	21	17	5	18	20	21	19	18	7	23	26	100	24	24	24	23	25	25	21	19	12
SO-CAL	32	28	5	25	26	28	29	35	8	36	32	24	100	31	39	32	32	35	29	31	17
SWN	27	26	1	26	24	21	23	26	3	31	28	24	31	100	30	32	32	31	21	22	14
SentiStrength	33	31	7	19	25	31	31	32	9	34	34	24	39	30	100	29	32	38	30	31	16
SenticNet	29	24	0,8	32	27	19	21	25	3,6	32	27	23	32	32	29	100	33	28	25	24	15
Sentim.140	28	28	2	26	24	30	23	27	4	32	27	25	32	32	32	33	100	36	22	23	15
Stanford DM	27	27	3	18	21	32	28	29	5	35	33	25	35	31	38	28	36	100	24	25	16
Umigon	29	23	10	18	23	20	20	24	11	27	26	21	29	21	30	25	22	24	100	27	14
VADER	30	25	11	18	21	21	23	27	12	27	26	19	31	22	31	24	23	25	27	100	13
LIWC	15	13	6	14	14	13	13	14	7	17	16	12	17	14	16	15	15	16	14	13	100

(a) Percentage of agreement on Comments_TED dataset

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtah	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	28	0	28	22	17	13	33	4	32	25	16	35	30	30	32	34	35	21	25	30
Emolex	28	100	0	27	22	20	16	33	3	33	25	18	37	32	31	32	34	39	20	21	27
Emotic.	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Emotic. DS	28	27	0	100	28	10	10	28	3	31	24	18	33	33	22	42	43	35	15	21	29
Happ. Index	22	22	0	28	100	11	9	22	3	23	18	13	25	24	20	29	28	26	13	18	22
NRC Hashtag	17	20	0	10	11	100	17	22	2	24	17	14	27	21	24	17	22	32	16	12	16
Opin. Finder	13	16	0	10	9	17	100	17	2	21	14	12	23	18	19	14	18	29	13	9	12
Opin. Lexicon	33	33	0	28	22	22	17	100	4	38	30	19	43	35	35	34	37	44	23	24	31
PANAS	4	3	0	3	3	2	2	4	100	4	3	2	4	4	4	4	4	5	3	3	4
Pattern	32	33	0	31	23	24	21	38	4	100	30	21	45	38	35	35	40	49	26	24	31
SANN	25	25	0	24	18	17	14	30	3	30	100	15	33	28	27	28	30	35	19	20	25
SASA	16	18	0	18	13	14	12	19	2	21	15	100	23	20	18	19	22	26	13	12	16
SO-CAL	35	37	0	33	25	27	23	43	4	45	33	23	100	40	40	39	44	55	27	26	35
SWN	30	32	0	33	24	21	18	35	4	38	28	20	40	100	32	36	39	44	22	22	30
SentiStrength	30	31	0	22	20	24	19	35	4	35	27	18	40	32	100	29	33	43	24	23	32
SenticNet	32	32	0	42	29	17	14	34	3,9	35	28	19	39	36	29	100	43	40	20	24	33
Sentim.140	34	34	0	43	28	22	18	37	4	40	30	22	44	39	33	43	100	48	23	25	34
Stanford DM	35	39	0	35	26	32	29	44	5	49	35	26	55	44	43	40	48	100	29	26	35
Umigon	21	20	0	15	13	16	13	23	3	26	19	13	27	22	24	20	23	29	100	16	20
VADER	25	21	0	21	18	12	9	24	3	24	20	12	26	22	23	24	25	26	16	100	25
LIWC	30	27	0	29	22	16	12	31	4	31	25	16	35	30	32	33	34	35	20	25	100

(b) Percentage of agreement on Reviews_I dataset

Figure B.4. Percentage of agreement among all methods in two labeled datasets: Comments_TED and Reviews_I.

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtag	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	32	16	23	22	28	33	39	19	34	31	22	37	28	39	28	32	27	37	35	34
Emolex	32	100	17	17	26	25	36	32	19	25	25	18	35	25	32	23	25	24	29	28	28
Emotic.	16	17	100	2	15	8	22	16	24	11	13	8	15	6	8	4	3	4	15	20	9
Emotic. DS	23	17	2	100	16	12	18	21	1	26	21	15	20	24	26	31	26	15	19	22	25
Happ. Index	22	26	15	16	100	17	20	22	16	18	15	13	21	17	20	19	15	14	21	22	18
NRC Hashtag	28	25	8	12	17	100	23	28	11	25	19	20	32	25	37	19	37	35	29	21	27
Opin. Finder	33	36	22	18	20	23	100	35	22	31	28	19	36	27	32	23	26	23	29	34	31
Opin. Lexicon	39	32	16	21	22	28	35	100	18	31	31	21	36	28	34	26	31	26	28	34	32
PANAS	19	19	24	1	16	11	22	18	100	11	14	11	17	6	9	5	4	7	17	21	12
Pattern	34	25	11	26	18	25	31	31	11	100	32	19	35	28	35	29	32	29	35	27	32
SANN	31	25	13	21	15	19	28	31	14	32	100	17	32	22	27	20	23	19	26	26	29
SASA	22	18	8	15	13	20	19	21	11	19	17	100	22	19	26	20	22	21	20	18	22
SO-CAL	37	35	15	20	21	32	36	36	17	35	32	22	100	32	37	28	33	31	34	29	38
SWN	28	25	6	24	17	25	27	28	6	28	22	19	32	100	35	29	31	28	28	26	27
SentiStrength	39	32	8	26	20	37	32	34	9	35	27	26	37	35	100	33	40	35	39	29	41
SenticNet	28	23	4,2	31	19	19	23	26	5,3	29	20	20	28	29	33	100	33	22	26	25	29
Sentim.140	32	25	3	26	15	37	26	31	4	32	23	22	33	31	40	33	100	37	32	23	31
Stanford DM	27	24	4	15	14	35	23	26	7	29	19	21	31	28	35	22	37	100	29	18	28
Umigon	37	29	15	19	21	29	29	28	17	35	26	20	34	28	39	26	32	29	100	32	31
VADER	35	28	20	22	22	21	34	34	21	27	26	18	29	26	29	25	23	18	32	100	29
LIWC	34	28	9	25	18	27	31	32	12	32	29	22	38	27	41	29	31	28	31	29	100

(a) Percentage of agreement on Sarcasm dataset

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtag	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	31	11	5	22	31	30	35	14	28	25	18	34	24	36	18	24	32	27	33	17
Emolex	31	100	9	6	24	36	31	35	13	28	26	17	36	26	41	20	27	37	25	30	18
Emotic.	11	9	100	0	9	6	9	13	22	6	9	10	9	2	7	2	1	4	15	15	7
Emotic. DS	5	6	0	100	7	1	4	5	1	6	6	5	6	8	6	9	6	4	4	5	4
Happ. Index	22	24	9	7	100	20	22	23	11	20	19	13	25	18	25	16	17	23	19	23	12
NRC Hashtag	31	36	6	1	20	100	38	36	10	34	27	24	42	28	51	19	36	54	28	29	21
Opin. Finder	30	31	9	4	22	38	100	35	13	30	27	20	37	25	42	18	27	40	27	29	18
Opin. Lexicon	35	35	13	5	23	36	35	100	16	30	29	20	37	25	41	19	26	36	30	33	18
PANAS	14	13	22	1	11	10	13	16	100	10	12	11	13	5	12	4	5	9	17	17	9
Pattern	28	28	6	6	20	34	30	30	10	100	25	19	35	26	38	19	26	37	26	25	17
SANN	25	26	9	6	19	27	27	29	12	25	100	17	29	21	32	16	19	28	23	25	14
SASA	18	17	10	5	13	24	20	20	11	19	17	100	23	16	25	12	16	25	19	18	12
SO-CAL	34	36	9	6	25	42	37	37	13	35	29	23	100	27	47	21	31	44	30	32	20
SWN	24	26	2	8	18	28	25	25	5	26	21	16	27	100	32	20	24	31	19	22	16
SentiStrength	36	41	7	6	25	51	42	41	12	38	32	25	47	32	100	24	35	54	33	35	23
SenticNet	18	20	1,8	8,8	16	19	18	19	3,9	19	16	12	21	20	24	100	18	21	15	17	12
Sentim.140	24	27	1	6	17	36	27	26	5	26	19	16	31	24	35	18	100	37	19	22	15
Stanford DM	32	37	4	4	23	54	40	36	9	37	28	25	44	31	54	21	37	100	30	29	23
Umigon	27	25	15	4	19	28	27	30	17	26	23	19	30	19	33	15	19	30	100	28	14
VADER	33	30	15	5	23	29	29	33	17	25	25	18	32	22	35	17	22	29	28	100	16
LIWC	17	18	7	4	12	21	18	18	9	17	14	12	20	16	23	12	15	23	14	16	100

(b) Percentage of agreement on Coments_BBC dataset

Figure B.5. Percentage of agreement among all methods in two labeled datasets: Sarcasm and Coments_BBC.

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtah	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	31	17	13	24	26	30	35	18	32	30	23	33	26	36	23	25	29	34	33	16
Emolex	31	100	16	8	23	26	28	30	17	27	26	21	32	24	33	19	23	29	31	27	15
Emotic.	17	16	100	4	15	10	17	16	26	13	16	11	14	7	13	7	4	9	17	20	12
Emotic. DS	13	8	4	100	12	4	9	11	3	13	11	9	12	13	12	17	14	7	11	10	7
Happ. Index	24	23	15	12	100	16	22	24	16	23	21	16	25	20	25	19	17	18	24	24	12
NRC Hashtag	26	26	10	4	16	100	27	26	12	28	22	25	33	26	36	15	32	39	32	23	14
Opin. Finder	30	28	17	9	22	27	100	31	19	30	29	22	34	23	34	19	22	30	32	29	14
Opin. Lexicon	35	30	16	11	24	26	31	100	17	29	28	21	34	24	34	21	25	28	33	30	15
PANAS	18	17	26	3	16	12	19	17	100	14	17	12	16	8	15	7	5	11	18	21	12
Pattern	32	27	13	13	23	28	30	29	14	100	28	25	36	28	36	24	29	33	37	30	16
SANN	30	26	16	11	21	22	29	28	17	28	100	21	31	22	32	19	21	26	31	28	14
SASA	23	21	11	9	16	25	22	21	12	25	21	100	26	20	29	17	21	26	27	22	15
SO-CAL	33	32	14	12	25	33	34	34	16	36	31	26	100	30	42	24	30	37	40	32	16
SWN	26	24	7	13	20	26	23	24	8	28	22	20	30	100	31	24	28	30	28	22	13
SentiStrength	36	33	13	12	25	36	34	34	15	36	32	29	42	31	100	25	32	40	41	34	17
SenticNet	23	19	6,7	17	19	15	19	21	6,9	24	19	17	24	24	25	100	21	19	23	20	11
Sentim.140	25	23	4	14	17	32	22	25	5	29	21	21	30	28	32	21	100	34	29	20	13
Stanford DM	29	29	9	7	18	39	30	28	11	33	26	26	37	30	40	19	34	100	37	25	16
Umigon	34	31	17	11	24	32	32	33	18	37	31	27	40	28	41	23	29	37	100	34	17
VADER	33	27	20	10	24	23	29	30	21	30	28	22	32	22	34	20	20	25	34	100	16
LIWC	16	15	12	7	12	14	14	15	12	16	14	15	16	13	17	11	13	16	17	16	100

(a) Percentage of agreement on Comments_Digg dataset

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtah	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	32	17	42	42	22	30	36	20	48	37	27	41	39	47	46	37	26	44	47	39
Emolex	32	100	15	24	31	20	25	31	16	30	27	18	30	26	29	29	25	19	26	32	25
Emotic.	17	15	100	8	15	8	16	14	19	15	16	12	15	9	12	11	8	9	15	19	14
Emotic. DS	42	24	8	100	45	17	22	26	9	47	30	22	38	46	54	55	49	19	40	43	43
Happ. Index	42	31	15	45	100	19	28	30	18	43	34	25	39	40	44	47	36	21	37	44	37
NRC Hashtag	22	20	8	17	19	100	16	21	10	21	17	17	21	19	23	21	23	19	22	21	18
Opin. Finder	30	25	16	22	28	16	100	26	18	28	29	18	29	25	27	25	20	16	26	30	23
Opin. Lexicon	36	31	14	26	30	21	26	100	15	32	29	19	32	28	31	31	27	20	30	34	26
PANAS	20	16	19	9	18	10	18	15	100	17	17	16	17	10	14	12	7	12	17	22	15
Pattern	48	30	15	47	43	21	28	32	17	100	36	27	42	42	50	49	40	26	46	48	40
SANN	37	27	16	30	34	17	29	29	17	36	100	21	34	32	35	33	27	19	32	38	29
SASA	27	18	12	22	25	17	18	19	16	27	21	100	24	21	28	25	21	20	27	27	22
SO-CAL	41	30	15	38	39	21	29	32	17	42	34	24	100	37	43	43	34	24	39	43	34
SWN	39	26	9	46	40	19	25	28	10	42	32	21	37	100	45	47	39	21	37	39	36
SentiStrength	47	29	12	54	44	23	27	31	14	50	35	28	43	45	100	54	47	27	48	47	43
SenticNet	46	29	11	55	47	21	25	31	12	49	33	25	43	47	54	100	46	24	43	45	42
Sentim.140	37	25	8	49	36	23	20	27	7	40	27	21	34	39	47	46	100	23	37	36	35
Stanford DM	26	19	9	19	21	19	16	20	12	26	19	20	24	21	27	24	23	100	26	24	19
Umigon	44	26	15	40	37	22	26	30	17	46	32	27	39	37	48	43	37	26	100	43	37
VADER	47	32	19	43	44	21	30	34	22	48	38	27	43	39	47	45	36	24	43	100	39
LIWC	39	25	14	43	37	18	23	26	15	40	29	22	34	36	43	42	35	19	37	39	100

(b) Percentage of agreement on Myspace dataset

Figure B.6. Percentage of agreement among all methods in two labeled datasets: Comments_Digg and Myspace.

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtag	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	32	18	34	36	15	26	38	17	35	35	24	36	34	35	37	29	16	26	42	42
Emolex	32	100	16	23	29	14	24	32	17	27	28	18	31	26	28	27	24	16	21	31	31
Emotic.	18	16	100	8	14	7	15	17	26	12	15	17	13	9	12	9	7	6	13	22	16
Emotic. DS	34	23	8	100	36	7	19	29	6	32	30	19	30	36	30	42	30	8	19	33	38
Happ. Index	36	29	14	36	100	11	24	34	14	32	32	21	34	34	31	38	28	13	22	36	37
NRC Hashtag	15	14	7	7	11	100	14	16	8	17	14	12	18	14	18	12	20	20	19	13	14
Opin. Finder	26	24	15	19	24	14	100	27	15	24	26	16	26	23	24	23	20	16	21	26	26
Opin. Lexicon	38	32	17	29	34	16	27	100	17	31	33	23	35	31	33	34	28	16	25	37	38
PANAS	17	17	26	6	14	8	15	17	100	10	15	17	14	8	11	8	5	7	12	20	16
Pattern	35	27	12	32	32	17	24	31	10	100	31	21	35	34	35	34	31	20	28	33	36
SANN	35	28	15	30	32	14	26	33	15	31	100	21	34	30	30	32	26	15	25	36	36
SASA	24	18	17	19	21	12	16	23	17	21	21	100	21	19	21	21	18	13	17	25	23
SO-CAL	36	31	13	30	34	18	26	35	14	35	34	21	100	34	35	34	31	21	29	36	37
SWN	34	26	9	36	34	14	23	31	8	34	30	19	34	100	33	37	32	16	24	32	37
SentiStrength	35	28	12	30	31	18	24	33	11	35	30	21	35	33	100	33	31	21	30	33	37
SenticNet	37	27	9,5	42	38	12	23	34	8,2	34	32	21	34	37	33	100	32	13	23	35	40
Sentim.140	29	24	7	30	28	20	20	28	5	31	26	18	31	32	31	32	100	22	26	28	31
Stanford DM	16	16	6	8	13	20	16	16	7	20	15	13	21	16	21	13	22	100	22	14	16
Umigon	26	21	13	19	22	19	21	25	12	28	25	17	29	24	30	23	26	22	100	26	28
VADER	42	31	22	33	36	13	26	37	20	33	36	25	36	32	33	35	28	14	26	100	42
LIWC	42	31	16	38	37	14	26	38	16	36	36	23	37	37	37	40	31	16	28	42	100

(a) Percentage of agreement on RW dataset

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtag	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	36	30	21	34	27	38	43	30	38	39	29	40	28	40	28	24	16	43	45	26
Emolex	36	100	33	13	31	27	38	40	33	30	35	27	38	23	33	21	19	14	36	37	23
Emotic.	30	33	100	6	27	22	38	35	42	26	34	26	31	13	26	11	8	6	35	38	24
Emotic. DS	21	13	6	100	20	10	12	15	4	23	16	14	19	24	24	28	24	8	20	19	15
Happ. Index	34	31	27	20	100	21	31	33	27	31	31	24	34	25	32	26	19	11	34	36	22
NRC Hashtag	27	27	22	10	21	100	28	30	23	26	26	23	30	21	29	18	24	20	31	27	17
Opin. Finder	38	38	38	12	31	28	100	42	39	32	41	30	40	21	35	20	16	13	40	41	26
Opin. Lexicon	43	40	35	15	33	30	42	100	36	35	40	29	42	25	38	24	21	15	41	43	26
PANAS	30	33	42	4	27	23	39	36	100	23	34	26	32	12	25	10	6	6	33	37	24
Pattern	38	30	26	23	31	26	32	35	23	100	34	26	39	30	39	31	27	19	41	38	23
SANN	39	35	34	16	31	26	41	40	34	34	100	28	38	24	37	23	19	13	39	41	25
SASA	29	27	26	14	24	23	30	29	26	26	28	100	29	20	28	19	18	13	31	31	20
SO-CAL	40	38	31	19	34	30	40	42	32	39	38	29	100	28	41	28	25	18	42	42	26
SWN	28	23	13	24	25	21	21	25	12	30	24	20	28	100	30	30	26	16	27	26	17
SentiStrength	40	33	26	24	32	29	35	38	25	39	37	28	41	30	100	31	29	21	44	41	25
SenticNet	28	21	11	28	26	18	20	24	9,7	31	23	19	28	30	31	100	27	15	27	26	18
Sentim.140	24	19	8	24	19	24	16	21	6	27	19	18	25	26	29	27	100	21	27	21	15
Stanford DM	16	14	6	8	11	20	13	15	6	19	13	13	18	16	21	15	21	100	19	14	9
Umigon	43	36	35	20	34	31	40	41	33	41	39	31	42	27	44	27	27	19	100	45	27
VADER	45	37	38	19	36	27	41	43	37	38	41	31	42	26	41	26	21	14	45	100	28
LIWC	26	23	24	15	22	17	26	26	24	23	25	20	26	17	25	18	15	9	27	28	100

(b) Percentage of agreement on Tweets_RND_I dataset

Figure B.7. Percentage of agreement among all methods in two labeled datasets: RW and Tweets_RND_I.

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtag	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	33	19	32	34	22	31	40	19	41	37	28	41	34	42	37	30	29	41	43	25
Emolex	33	100	18	20	29	21	28	33	19	30	29	21	33	27	30	27	23	24	30	32	19
Emotic.	19	18	100	7	16	11	20	19	26	16	20	13	17	10	15	11	6	10	19	24	14
Emotic. DS	32	20	7	100	28	12	17	25	4	35	27	23	31	32	36	39	33	24	30	30	25
Happ. Index	34	29	16	28	100	17	26	30	17	33	31	22	34	30	33	34	24	24	31	34	22
NRC Hashtag	22	21	11	12	17	100	19	23	11	23	20	19	25	20	24	18	23	24	23	21	13
Opin. Finder	31	28	20	17	26	19	100	31	22	29	31	21	32	24	29	24	19	22	29	32	18
Opin. Lexicon	40	33	19	25	30	23	31	100	19	36	35	25	39	30	36	32	27	28	35	37	22
PANAS	19	19	26	4	17	11	22	19	100	14	19	12	18	10	14	9	4	10	17	23	13
Pattern	41	30	16	35	33	23	29	36	14	100	36	30	44	37	46	41	34	34	44	41	27
SANN	37	29	20	27	31	20	31	35	19	36	100	25	38	30	36	33	25	27	35	38	23
SASA	28	21	13	23	22	19	21	25	12	30	25	100	30	26	32	28	24	25	31	29	20
SO-CAL	41	33	17	31	34	25	32	39	18	44	38	30	100	36	45	40	32	34	42	42	25
SWN	34	27	10	32	30	20	24	30	10	37	30	26	36	100	37	37	31	28	33	33	22
SentiStrength	42	30	15	36	33	24	29	36	14	46	36	32	45	37	100	42	35	36	46	43	28
SenticNet	37	27	11	39	34	18	24	32	9,1	41	33	28	40	37	42	100	33	30	36	37	26
Sentim.140	30	23	6	33	24	23	19	27	4	34	25	24	32	31	35	33	100	29	32	28	20
Stanford DM	29	24	10	24	24	24	22	28	10	34	27	25	34	28	36	30	29	100	33	30	19
Umigon	41	30	19	30	31	23	29	35	17	44	35	31	42	33	46	36	32	33	100	42	25
VADER	43	32	24	30	34	21	32	37	23	41	38	29	42	33	43	37	28	30	42	100	26
LIWC	25	19	14	25	22	13	18	22	13	27	23	20	25	22	28	26	20	19	25	26	100

(a) Percentage of agreement on Comments_YTB dataset

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtag	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	30	8	44	35	26	26	38	6	48	34	29	43	35	51	51	45	36	44	44	32
Emolex	30	100	5	25	23	20	18	27	4	28	20	17	29	24	31	31	30	24	26	26	18
Emotic.	8	5	100	12	7	5	4	6	1	14	6	7	8	6	10	10	10	7	12	11	8
Emotic. DS	44	25	12	100	39	18	20	31	6	51	32	32	40	30	49	59	48	28	44	44	40
Happ. Index	35	23	7	39	100	17	20	26	6	37	26	24	35	28	39	44	35	26	33	34	27
NRC Hashtag	26	20	5	18	17	100	17	23	4	27	19	18	27	23	30	26	31	28	26	24	16
Opin. Finder	26	18	4	20	20	17	100	23	4	27	22	17	26	21	29	27	25	22	25	24	16
Opin. Lexicon	38	27	6	31	26	23	23	100	4	35	26	21	35	27	37	37	35	28	32	31	22
PANAS	6	4	1	6	6	4	4	4	100	7	5	4	7	5	7	7	6	6	7	7	4
Pattern	48	28	14	51	37	27	27	35	7	100	35	34	48	38	56	56	50	41	53	49	36
SANN	34	20	6	32	26	19	22	26	5	35	100	23	32	26	38	38	33	26	32	33	24
SASA	29	17	7	32	24	18	17	21	4	34	23	100	29	24	35	35	32	25	31	31	24
SO-CAL	43	29	8	40	35	27	26	35	7	48	32	29	100	35	50	49	44	38	43	43	30
SWN	35	24	6	30	28	23	21	27	5	38	26	24	35	100	40	40	36	31	34	32	24
SentiStrength	51	31	10	49	39	30	29	37	7	56	38	35	50	40	100	58	52	44	53	52	38
SenticNet	51	31	10	59	44	26	27	37	7,3	56	38	35	49	40	58	100	53	39	50	50	41
Sentim.140	45	30	10	48	35	31	25	35	6	50	33	32	44	36	52	53	100	40	46	43	34
Stanford DM	36	24	7	28	26	28	22	28	6	41	26	25	38	31	44	39	40	100	39	35	24
Umigon	44	26	12	44	33	26	25	32	7	53	32	31	43	34	53	50	46	39	100	46	33
VADER	44	26	11	44	34	24	24	31	7	49	33	31	43	32	52	50	43	35	46	100	32
LIWC	32	18	8	40	27	16	16	22	4	36	24	24	30	24	38	41	34	24	33	32	100

(b) Percentage of agreement on Tweets_STF dataset

Figure B.8. Percentage of agreement among all methods in two labeled datasets: Comments_YTB and Tweets_STF.

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtag	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	22	2	28	20	16	17	33	4	30	22	20	31	28	28	31	26	24	23	22	15
Emolex	22	100	2	22	16	14	15	23	3	23	15	16	25	23	21	24	23	20	17	15	13
Emotic.	2	2	100	0	2	1	2	2	3	1	2	1	2	1	2	1	0	1	2	3	1
Emotic. DS	28	22	0	100	27	13	16	31	2	36	24	28	37	39	27	49	30	26	22	22	27
Happ. Index	20	16	2	27	100	10	13	20	4	23	15	16	24	23	19	28	18	17	17	15	14
NRC Hashtag	16	14	1	13	10	100	12	18	2	23	13	17	24	21	20	18	30	28	18	12	10
Opin. Finder	17	15	2	16	13	12	100	19	3	21	17	14	22	19	20	19	17	19	15	14	10
Opin. Lexicon	33	23	2	31	20	18	19	100	4	33	23	21	35	30	30	33	28	28	25	22	17
PANAS	4	3	3	2	4	2	3	4	100	3	3	3	4	2	4	3	2	3	4	4	2
Pattern	30	23	1	36	23	23	21	33	3	100	26	26	42	38	34	39	35	36	29	24	20
SANN	22	15	2	24	15	13	17	23	3	26	100	18	26	23	24	26	20	21	19	19	14
SASA	20	16	1	28	16	17	14	21	3	26	18	100	26	26	22	28	25	24	18	16	16
SO-CAL	31	25	2	37	24	24	22	35	4	42	26	26	100	38	37	41	36	38	30	25	22
SWN	28	23	1	39	23	21	19	30	2	38	23	26	38	100	30	41	34	33	26	21	21
SentiStrength	28	21	2	27	19	20	20	30	4	34	24	22	37	30	100	32	29	32	28	24	17
SenticNet	31	24	0,6	49	28	18	19	33	2,7	39	26	28	41	41	32	100	33	31	26	24	25
Sentim.140	26	23	0	30	18	30	17	28	2	35	20	25	36	34	29	33	100	38	25	18	19
Stanford DM	24	20	1	26	17	28	19	28	3	36	21	24	38	33	32	31	38	100	27	20	17
Umigon	23	17	2	22	17	18	15	25	4	29	19	18	30	26	28	26	25	27	100	20	14
VADER	22	15	3	22	15	12	14	22	4	24	19	16	25	21	24	24	18	20	20	100	12
LIWC	15	13	1	27	14	10	10	17	2	20	14	16	22	21	17	25	19	17	14	12	100

(a) Percentage of agreement on Amazon dataset

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtag	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	24	0	25	20	16	18	28	3	29	23	14	31	27	27	29	30	33	20	23	18
Emolex	24	100	0	24	20	18	19	28	3	29	23	16	33	28	27	29	30	35	18	20	19
Emotic.	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Emotic. DS	25	24	0	100	29	11	11	26	3	31	25	18	33	33	22	42	38	38	16	22	29
Happ. Index	20	20	0	29	100	12	12	21	3	24	20	14	26	25	21	30	27	29	14	18	19
NRC Hashtag	16	18	0	11	12	100	20	20	2	25	17	14	27	22	24	17	25	33	17	13	13
Opin. Finder	18	19	0	11	12	20	100	24	3	27	22	13	30	23	27	18	23	33	19	14	13
Opin. Lexicon	28	28	0	26	21	20	24	100	4	34	28	17	38	31	31	31	34	40	21	23	20
PANAS	3	3	0	3	3	2	3	4	100	4	3	2	4	4	4	4	4	5	3	3	3
Pattern	29	29	0	31	24	25	27	34	4	100	31	21	46	39	36	36	41	51	27	25	25
SANN	23	23	0	25	20	17	22	28	3	31	100	15	34	29	28	29	30	36	20	21	19
SASA	14	16	0	18	14	14	13	17	2	21	15	100	23	20	18	20	23	27	14	12	14
SO-CAL	31	33	0	33	26	27	30	38	4	46	34	23	100	41	41	39	44	56	28	27	27
SWN	27	28	0	33	25	22	23	31	4	39	29	20	41	100	33	37	38	46	23	23	25
SentiStrength	27	27	0	22	21	24	27	31	4	36	28	18	41	33	100	30	34	44	25	24	20
SenticNet	29	29	0	42	30	17	18	31	4,1	36	29	20	39	37	30	100	40	43	21	25	27
Sentim.140	30	30	0	38	27	25	23	34	4	41	30	23	44	38	34	40	100	51	24	25	28
Stanford DM	33	35	0	38	29	33	33	40	5	51	36	27	56	46	44	43	51	100	31	27	32
Umigon	20	18	0	16	14	17	19	21	3	27	20	14	28	23	25	21	24	31	100	17	14
VADER	23	20	0	22	18	13	14	23	3	25	21	12	27	23	24	25	25	27	17	100	15
LIWC	18	19	0	29	19	13	13	20	3	25	19	14	27	25	20	27	28	32	14	15	100

(b) Percentage of agreement on Reviews_II dataset

Figure B.9. Percentage of agreement among all methods in two labeled datasets: Amazon and Reviews_II.

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtah	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	21	4	15	16	18	17	25	5	21	16	12	25	22	24	22	23	20	15	18	9
Emolex	21	100	3	18	18	19	16	23	4	20	15	11	26	24	23	25	25	21	12	14	10
Emotic.	4	3	100	0	3	2	4	4	5	2	4	3	3	1	4	1	0	2	4	4	3
Emotic. DS	15	18	0	100	22	8	8	16	2	19	13	13	22	28	12	36	27	9	7	10	12
Happ. Index	16	18	3	22	100	12	11	17	4	17	12	10	21	21	16	25	20	13	10	12	9
NRC Hashtag	18	19	2	8	12	100	17	20	3	23	13	14	26	23	26	17	30	33	13	12	10
Opin. Finder	17	16	4	8	11	17	100	19	5	17	15	10	22	17	22	15	17	19	13	13	8
Opin. Lexicon	25	23	4	16	17	20	19	100	5	22	18	12	28	24	26	23	25	23	15	17	10
PANAS	5	4	5	2	4	3	5	5	100	4	5	4	5	3	5	3	2	3	5	5	3
Pattern	21	20	2	19	17	23	17	22	4	100	16	14	29	29	24	26	29	27	15	14	11
SANN	16	15	4	13	12	13	15	18	5	16	100	9	19	17	19	17	16	15	11	13	8
SASA	12	11	3	13	10	14	10	12	4	14	9	100	16	16	13	15	17	15	9	9	8
SO-CAL	25	26	3	22	21	26	22	28	5	29	19	16	100	32	32	30	33	30	17	18	13
SWN	22	24	1	28	21	23	17	24	3	29	17	16	32	100	25	34	34	27	14	14	13
SentiStrength	24	23	4	12	16	26	22	26	5	24	19	13	32	25	100	23	27	30	17	19	11
SenticNet	22	25	0,8	36	25	17	15	23	2,6	26	17	15	30	34	23	100	33	20	13	14	13
Sentim.140	23	25	0	27	20	30	17	25	2	29	16	17	33	34	27	33	100	33	14	14	14
Stanford DM	20	21	2	9	13	33	19	23	3	27	15	15	30	27	30	20	33	100	15	14	11
Umigon	15	12	4	7	10	13	13	15	5	15	11	9	17	14	17	13	14	15	100	12	7
VADER	18	14	4	10	12	12	13	17	5	14	13	9	18	14	19	14	14	14	12	100	7
LIWC	9	10	3	12	9	10	8	10	3	11	8	8	13	13	11	13	14	11	7	7	100

(a) Percentage of agreement on Comments_NYT dataset

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtah	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	32	8	46	0	28	28	40	7	49	35	0	44	0	53	52	48	38	45	44	34
Emolex	32	100	51	32	28	46	59	69	54	44	52	21	56	1	48	38	38	35	46	51	36
Emotic.	8	51	100	16	34	32	53	49	79	34	47	32	36	1	25	13	14	17	40	47	30
Emotic. DS	46	32	16	100	3	22	25	36	10	56	39	1	45	0	52	38	58	30	46	46	54
Happ. Index	0	28	34	3	100	16	29	27	39	12	24	17	20	1	11	55	3	6	15	21	12
NRC Hashtag	28	46	32	22	16	100	43	47	35	38	40	12	44	1	42	28	49	49	41	39	30
Opin. Finder	28	59	53	25	29	43	100	63	60	45	61	24	56	1	48	25	32	35	49	52	35
Opin. Lexicon	40	69	49	36	27	47	63	100	52	51	58	21	62	1	54	45	42	39	53	57	38
PANAS	7	54	79	10	39	35	60	52	100	27	49	34	38	1	26	15	11	17	35	45	29
Pattern	49	44	34	56	12	38	45	51	27	100	51	9	63	0	67	59	56	49	68	62	46
SANN	35	52	47	39	24	40	61	58	49	51	100	21	56	1	54	37	39	36	53	59	40
SASA	0	21	32	1	17	12	24	21	34	9	21	100	14	0	9	33	1	4	14	18	10
SO-CAL	44	56	36	45	20	44	56	62	38	63	56	14	100	0	65	52	51	49	61	62	42
SWN	0	1	1	0	1	1	1	1	1	0	1	0	0	100	0	59	0	0	0	1	0
SentiStrength	53	48	25	52	11	42	48	54	26	67	54	9	65	0	100	19	56	53	68	67	47
SenticNet	57	51	29	38	16	46	50	56	31	58	54	13	62	1	19	100	49	49	60	57	43
Sentim.140	48	38	14	58	3	49	32	42	11	56	39	1	51	0	56	56	100	54	51	45	42
Stanford DM	38	35	17	30	6	49	35	39	17	49	36	4	49	0	53	30	54	100	48	41	33
Umigon	45	46	40	46	15	41	49	53	35	68	53	14	61	0	68	33	51	48	100	65	43
VADER	44	51	47	46	21	39	52	57	45	62	59	18	62	1	67	54	45	41	65	100	45
LIWC	34	36	30	54	12	30	35	38	29	46	40	10	42	0	47	38	42	33	43	45	100

(b) Percentage of agreement on Tweets_RND_II dataset

Figure B.10. Percentage of agreement among all methods in two labeled datasets: Comments_NYT and Tweets_RND_II.

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtah	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	52	3	47	48	26	35	58	18	62	56	29	62	57	57	52	58	39	39	58	47
Emolex	52	100	3	42	44	26	33	54	18	57	51	27	58	53	53	48	55	36	36	51	43
Emotic.	3	3	100	3	3	1	2	3	1	3	3	1	3	3	3	3	3	2	1	3	3
Emotic. DS	47	42	3	100	47	8	22	45	15	46	45	21	46	47	40	49	42	21	22	47	43
Happ. Index	48	44	3	47	100	12	25	47	16	49	46	22	49	48	43	49	45	24	25	48	42
NRC Hashtag	26	26	1	8	12	100	30	30	9	44	29	23	44	29	46	15	48	50	45	25	21
Opin. Finder	35	33	2	22	25	30	100	37	11	43	36	22	44	36	42	28	43	35	33	34	27
Opin. Lexicon	58	54	3	45	47	30	37	100	19	64	57	29	65	58	59	51	61	41	41	57	46
PANAS	18	18	1	15	16	9	11	19	100	20	18	8	20	18	19	16	19	12	11	18	15
Pattern	62	57	3	46	49	44	43	64	20	100	64	36	76	65	72	53	73	56	54	62	51
SANN	56	51	3	45	46	29	36	57	18	64	100	29	63	56	58	50	59	41	41	56	46
SASA	29	27	1	21	22	23	22	29	8	36	29	100	35	30	34	25	35	29	27	28	24
SO-CAL	62	58	3	46	49	44	44	65	20	76	63	35	100	64	72	53	73	56	54	61	51
SWN	57	53	3	47	48	29	36	58	18	65	56	30	64	100	59	52	60	42	41	56	47
SentiStrength	57	53	3	40	43	46	42	59	19	72	58	34	72	59	100	47	70	57	56	56	47
SenticNet	52	48	3	49	49	15	28	51	16	53	50	25	53	52	47	100	49	28	29	52	45
Sentim.140	58	55	3	42	45	48	43	61	19	73	59	35	73	60	70	49	100	58	55	57	49
Stanford DM	39	36	2	21	24	50	35	41	12	56	41	29	56	42	57	28	58	100	52	38	31
Umigon	39	36	1	22	25	45	33	41	11	54	41	27	54	41	56	29	55	52	100	38	31
VADER	58	51	3	47	48	25	34	57	18	62	56	28	61	56	56	52	57	38	38	100	46
LIWC	47	43	3	43	42	21	27	46	15	51	46	24	51	47	47	45	49	31	31	46	100

(a) Percentage of agreement on YLP dataset

	AFINN	Emolex	Emotic.	Emotic. DS	Happ. Index	NRC Hashtah	Opin. Finder	Opin. Lexicon	PANAS	Pattern	SANN	SASA	SO-CAL	SWN	SentiStrength	SenticNet	Sentim.140	Stanford DM	Umigon	VADER	LIWC
AFINN	100	37	33	22	34	24	41	46	36	35	40	33	42	23	42	26	22	13	46	49	24
Emolex	37	100	33	13	27	21	37	39	34	26	32	27	36	15	31	17	15	10	38	38	21
Emotic.	33	33	100	4	25	21	42	38	48	21	34	31	31	5	27	6	4	5	41	41	23
Emotic. DS	22	13	4	100	24	8	10	17	5	25	17	17	22	26	25	32	26	7	21	21	15
Happ. Index	34	27	25	24	100	17	29	32	26	28	29	27	32	21	32	25	19	9	34	36	21
NRC Hashtag	24	21	21	8	17	100	25	25	22	18	22	19	24	12	23	12	15	13	27	25	14
Opin. Finder	41	37	42	10	29	25	100	44	45	27	41	33	38	13	35	14	11	10	45	44	24
Opin. Lexicon	46	39	38	17	32	25	44	100	40	32	40	33	43	19	39	21	19	13	45	47	24
PANAS	36	34	48	5	26	22	45	40	100	22	35	33	33	5	29	7	4	5	41	42	24
Pattern	35	26	21	25	28	18	27	32	22	100	29	25	35	25	34	28	25	14	38	36	20
SANN	40	32	34	17	29	22	41	40	35	29	100	31	36	18	36	20	16	10	41	42	23
SASA	33	27	31	17	27	19	33	33	33	25	31	100	32	16	30	19	15	8	36	36	22
SO-CAL	42	36	31	22	32	24	38	43	33	35	36	32	100	23	39	26	22	14	44	43	23
SWN	23	15	5	26	21	12	13	19	5	25	18	16	23	100	25	28	25	12	22	21	13
SentiStrength	42	31	27	25	32	23	35	39	29	34	36	30	39	25	100	29	25	16	44	43	23
SenticNet	26	17	6,4	32	25	12	14	21	7,1	28	20	19	26	28	29	100	27	11	25	24	15
Sentim.140	22	15	4	26	19	15	11	19	4	25	16	15	22	25	25	27	100	15	22	19	13
Stanford DM	13	10	5	7	9	13	10	13	5	14	10	8	14	12	16	11	15	100	15	11	6
Umigon	46	38	41	21	34	27	45	45	41	38	41	36	44	22	44	25	22	15	100	50	27
VADER	49	38	41	21	36	25	44	47	42	36	42	36	43	21	43	24	19	11	50	100	27
LIWC	24	21	23	15	21	14	24	24	24	20	23	22	23	13	23	15	13	6	27	27	100

(b) Percentage of agreement on Tweets_SemEval dataset

Figure B.11. Percentage of agreement among all methods in two labeled datasets: YLP and Tweets_SemEval.