On the Combination of "Off-The-Shelf" Sentiment Analysis Methods

Pollyanna Gonçalves
Federal University of Minas Gerais
Belo Horizonte, Brazil
pollyannaog@dcc.ufmg.br

Daniel Hasan Dalip
Federal University of Minas Gerais
Belo Horizonte, Brazil
hasan@dcc.ufmg.br

Helen Costa
Federal University of Ouro Preto
Ouro Preto, Brazil
helen@decsi.ufop.br

Marcos André Gonçalves
Federal University of Minas Gerais
Belo Horizonte, Brazil
mgoncalv@dcc.ufmg.br

Fabrício Benevenuto
Federal University of Minas Gerais
Belo Horizonte, Brazil
fabricio@dcc.ufmg.br

ABSTRACT

Sentiment analysis has become an important topic on the Web, especially in social media, with applications in many domains such as the monitoring of businesses and products as well as the analysis of the repercussion of important events. Several methods and techniques have been independently developed for this purpose in the literature. However, recent work has showed that all of them have varying degrees of coverage and prediction accuracy, with no "silver bullet" for all cases and scenarios. In this paper, we tackle this issue by proposing ensemble combination methods aimed at combining the outputs of several state-of-the-art proposals in order to maximize both goals, which sometimes can be conflicting. We focus on combining "off-the-shelf" methods, increasing enormously the applicability of our strategy. We tested our solutions in a very rich experimentation environment, covering thirteen widely used methods and fourteen labeled datasets from many domains, including messages from social networks, movie and product reviews, opinions and comments in news articles. Our experimental results demonstrate that we can be very successful in our goal, meaning that our proposal can produce a real and important impact in the area of sentiment classification research.

Keywords

Sentiment analysis; machine learning; social networks

1. INTRODUCTION

Given the recent popularity of Online Social Networks (OSNs) and other Web 2.0 applications (e.g., micro-blogs), sentiment analysis has become an important research topic, mainly when considering short and informal texts, a challenging scenario. Applications of sentiment analysis include the monitoring of reviews or opinions about a company, product or a brand, and political analysis, including the tracking of sentiments expressed by voters about candidates for an election, to cite a few. Due to its applicability, specially in the Web context, there are many researchers and companies currently developing tools and strategies to extract sentiments from text [10].

Several methods exploit a variant of the problem, namely polarity detection – the degree to which a given message express a positive or negative opinion about a topic [24]. In practice, it is usually used to measure the sentiment of small sets of sentences in which the topic is known *a priori*. This is the focus of our work.

There is a number of methods for sentiment analysis that rely in 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 [25]. 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 [31] was originally proposed to analyze sentiment patterns in formally written English texts, whereas PANAS-t [13] was proposed as psychometric scale adapted to the Web context. Other techniques include deep-learning based methods [29] and natural language processing approaches [4].

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, recent efforts have showed that there is no single method that always achieves the best prediction performance for different datasets [12, 27]. This is similar to the well-known "no-free lunch theorem" in machine learning (ML) [36]. Furthermore, it has also been realized that the methods have varying degrees of *coverage*, i.e., the fraction of messages whose sentiment can be, in fact, identified. Many methods may have a high accuracy when making a valid prediction, but cannot really identify if a text is positive or negative for the majority of the entries (e.g, a neutral score or "no-score" is predicted for most text inputs even when they have in fact a manually identified polarity). This is an important aspect that should be considered while evaluating sentiment analysis methods.

SAC 2016,April 04-08, 2016, Pisa, Italy

http://dx.doi.org/10.1145/2851613.2851820

In this paper, we tackle all these issues by proposing ensemble combination methods aimed at combining the outputs of several of the most popular state-of-the-art methods available in the literature in order to maximize both goals (i.e., accuracy and coverage), which is sometimes a conflicting goal. For instance, maximizing prediction accuracy may mean that many messages will not be classified, while focusing on improving coverage may come with losses in the correctness of many predictions due to issues such as uncertainty, ambiguity, etc. Furthermore, as some proposed methods were trained in specific domains, they may not perform well in different ones. Our method overcomes this issue by combining all available methods in order to identify the best methods (or the best combination of them) for a given domain. The main hypothesis is that the ensembles can "take the best of each method" in a given situation, adapting itself to the particularities and idiosyncrasies of each situation. We believe this is an essential component of a robust sentiment classification method, which can be seriously affected by the subjectivity and particularity of each domain. Moreover, the exploited methods in our ensemble are all open-source and freely available. Although some of them rely on ML techniques, they are *pre-trained* and can be used as an "off-the-shelf" method, increasing enormously the applicability of our solution for many situations.

Our experimental results demonstrate that, while, by construction, our ensembles can overcome the coverage problem by relying on a "multiple expert" approach (i.e., at least one base method provides a non-neutral output in the majority of the cases), this comes with no loss in accuracy, in fact, in some datasets the ensembles accuracy is better than any single base method. Our ensembles also outperform strong baselines which either propose some alternative type of combination of the base methods or rely in supervised text classification, especially when training is scarce. Overall, we demonstrate that it is possible to maximize in a very effective way both goals with our proposed solutions.

Next, we describe the state-of-the-art sentiment analysis methods that are used for comparison and combination. Then, we describe the proposed ensemble methods and the ground truth data used for evaluation. Finally, we present comparison results, briefly survey related efforts and concludes the paper.

2. SENTIMENT ANALYSIS METHODS

In this work, we are focused in combining popular "off-the-shelf" sentiment analysis methods that are freely available for use. As discussed before, the methods we use in this paper are based on different paradigms. We should notice however, that the employed ML methods have been *pre-trained* with selected datasets chosen by the developers of the methods, meaning that these methods are able to produce a polarity score based on their learned models for any given input text.

SentiStrength[1][32]: SentiStrength was built with the use of supervised and unsupervised classification methods. It classifies positive (from 1 to 5) and negative (from 1 to 5) polarity strength separately as the default setup of the method.

SASA[2][33]: SailAil Sentiment Analyzer (SASA) uses a supervised machine learning classifier to identify sentiments. It uses the statistical classifier Naïve Bayes on unigram features.

Stanford Recursive Deep Model[3][29]: It is based on a model called Recursive Neural Tensor Network that processes all sentences taking their structures into account and compute the interactions among them.

Emoticons [12]: These are primarily face-based and represent happy or sad feelings. To extract polarity from emoticons, a set of common emoticons was used. This set also includes the popular variations that express the primary polarities of positive and negative.

EmoLex [19]: Also called NRC Emotion Lexicon, it is lexical method with up 10,000 word-sense pairs. Each entry lists the association of a word-sense pair with 8 basic emotions: joy, sadness, anger, fear, trust, disgust, anticipation and surprise, defined by [26]. EmoLex version 0.92 was shared by the authors.

NRC Hashtag [18]: It is a lexicon dictionary of Twitter's hashtags associated with eight sentiments: joy, sadness, anger, fear, trust, disgust, anticipation and surprise, just like EmoLex. The dictionary of up to 32,000 hashtags was created from a collection of 775,310 tweets posted in 2012. In this paper, we used the NRC Hashtag Sentiment Lexicon version 0.2, which was shared by the authors.

Sentiment140 Lexicon[4][20]: It consists of a dictionary of words associated with positive and negative sentiments. The dictionary of Sentiment140 Lexicon contains 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.

OpinionLexicon[5][14]: Also known as Sentiment Lexicon, it is a lexical method consisting of two lists with 2,006 positive words and 4,783 negative words. It includes slang, misspellings, morphological variants, and social-media markups.

VADER[6][15]: Valence Aware Dictionary for sEntiment Reasoning (VADER) is a human-validated sentiment analysis method developed for micro-blogging and social media, requiring no training data.

Happiness Index [8]: It is a sentiment scale that measures the level of happiness in a given text. It uses the popular Affective Norms for English Words (ANEW) [2], a collection of 1,034 words commonly used associated with their affective dimensions of valence, arousal, and dominance.

SentiWordNet[7][9]: It is a tool widely used in opinion mining, based on the English lexical dictionary WordNet [17]. This dictionary groups adjectives, nouns, verbs and other grammatical classes into synonym sets called synsets. SentiWordNet associates three scores with synset from the WordNet dictionary to indicate the sentiment of the text: positive, negative, and objective (neutral).

SenticNet[8][4]: It is a method of opinion mining and sentiment analysis that explores machine learning and semantic Web 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 con-

[1]http://sentistrength.wlv.ac.uk/Download
[2]https://pypi.python.org/pypi/sasa/0.1.3
[3]http://nlp.stanford.edu/software/corenlp.shtml
[4]http://www.saifmohammad.com/WebPages/ResearchInterests.html
[5]http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
[6]http://comp.social.gatech.edu/papers/
[7]http://sentiwordnet.isti.cnr.it/
[8]http://sentic.net/

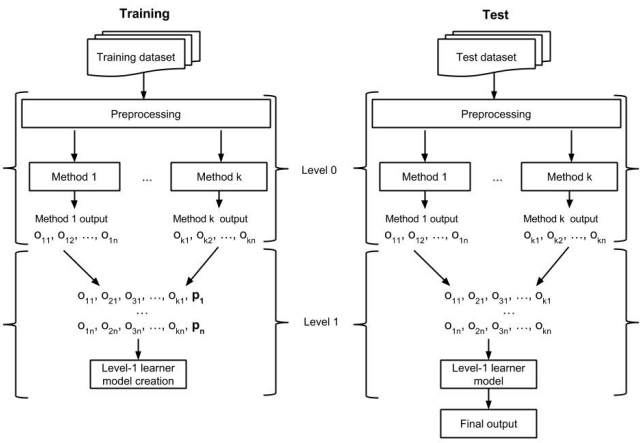


Figure 1: Ensemble method. o_{ij} is the output for the "off-the-shelf" method i for the message j. p_j represents the polarity of the labeled message j

cepts.

PANAS-t [13]: The last considered method is a psychometric scale proposed for detecting mood fluctuations of Twitter users. The method consists of an adapted version of the Positive Affect Negative Affect Scale (PANAS) [34], a well-known method in psychology. The PANAS-t is based on a large set of words associated with eleven moods: joviality, assurance, serenity, surprise, fear, sadness, guilt, hostility, shyness, fatigue, and attentiveness. The method is designed to track any increase or decrease in sentiments over time.

3. PROPOSED ENSEMBLE METHODS

Next, we discuss the two techniques we propose for combining the previously described base methods. Figure 1 shows the implemented combination proposal. The figure is divided into two parts: training and test.

The training is divided into two levels. In the first level (or learning Level 0), for each message j, an "off-the-shelf" method i is run to produce an estimate of the polarity output o_{ij} based on its own criteria. In some cases, a preprocessing of the original message j may be required (e.g., for stopwords removal) before running the base method, depending on the underlying technique exploited by each of these methods. After that, in the second level (or learning Level 1) the output o_{ij} of each method is then combined into a single vector $\{o_{1j}, o_{2j}, ..., o_{kj}\}$ representing all the k methods output for the message j, along with the corresponding polarity label p_j. This vector is then given as input for the creation of the ensemble learner, whose goal is to learn a model that better combines the output of each method in order to maximize some goal, for instance, accuracy, based on the known labeled information. This technique is similar to a stacking of multiple classifiers [35].

Similarly to the training step, the test step receives as input a set containing short messages, but differently from the first step, without the polarities. After preprocessing messages and getting the results of each method, the message j is represented as a single vector $\{o_{1j}, o_{2j}, ..., o_{kj}\}$. Then, we apply the Level-1 learner model in the messages vector in order to obtain the final output (polarity) for each message.

Table 1: Labeled datasets. For all datasets, half of messages are positive and the other half negatives. Emot stands for emoticons.

Dataset	Nomeclature	# Msgs	Number of words			% Msgs
			min	max	avg	w/ emot.
Stanford Twitter Corpus	**Stanford**	359	2	30	14	6.13
Yelp Dataset	**Yelp**	4,997	1	934	129	1.64
Bo & Pang movie-reviews	**Reviews I**	4,993	1	115	21	4.67
2008 Presidential Debate	**Debate**	1,486	1	30	13	1.01
SentiStrength's Tweets	**Twitter I**	308	2	33	18	6.49
SentiStrength's Myspace	**Myspace**	264	1	389	119	13.26
SentiStrength's Youtube comments	**YouTube**	162	1	90	30	5.55
SentiStrength's Digg	**Digg**	414	1	225	18	0.97
SentiStrength's Runners World	**RW**	442	1	536	72	14.48
SentiStrength's BBC comments	**BBC**	199	1	1,217	62	2.51
VADER's Amazon reviews	**Amazon**	1,248	1	69	15	0.40
VADER's NYT opinions	**NYT**	1,250	1	77	17	0.00
VADER's Tweets	**Twitter II**	1,250	1	31	12	4.72
VADER's movie-reviews	**Reviews II**	1,244	3	51	19	0.08

All thirteen base methods previously discussed exploit varying techniques and approaches for estimating the polarity of the texts. As such, it is common that the output of these methods also vary. We have chosen to preserve the original characteristics of the methods, thus in our experiments we consider the output of each method "as it is".

For the ensemble learners we use two state-of-the-art machine learning techniques that have reproduced "top-notch" performance in a number for different applications, namely Support Vector Machines (SVM) and Random Forests (RF). Particularly, we used LibSVM[9] implementation of SVM with a RBF kernel, as it is known to work better on smaller and denser feature spaces. In case of RF, we use the implementation available in the Weka software[10]. We have used GridSearch, also available in Weka, to optimize the best parameter choices for this algorithm. In the remainder of this paper, we named the SVM as Ensemble I, and RF as Ensemble II.

4. EXPERIMENTAL METHODOLOGY

Having introduced our Ensemble methods, we now describe the datasets and techniques used for evaluation.

4.1 Datasets

The datasets considered in this paper consist of seven sets of messages labeled as positive or negative from many domains, including messages from social networks, movie and product reviews, opinions and comments in news articles. The first dataset, Yelp [37], consists of a set of business reviews from the greater Phoenix, AZ metropolitan area. We used more two datasets of reviews: movie and product reviews [23, 15]. We also consider datasets of messages from many OSNs: the human labeled messages used in the Stanford Twitter Corpus [11], tweets from the SentiStrength and VADER research research [32, 15], tweets about the 2008 U.S. Presidential debate [7], messages from Myspace [32] and comments in YouTube and Digg [32]. Another set of datasets consists of messages from comments in the BBC and Runners World forum

[9]http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[10]http://www.cs.waikato.ac.nz/ml/weka/

from SentiStrength research [32], and also sentence-level snippets from New York Times opinion news editorials/articles [15]. All these datasets are available as part of a benchmark for sentence-level sentiment analysis [27].

Overall, we have fourteen datasets. Table 1 summarizes the characteristics of each dataset and also provides a simpler nomenclature for each one that will be used in the remainder of this paper. We removed some of the original messages to balance them in terms of the number of positive and negative messages. This was to avoid an extra factor to be accounted for in our evaluation (class imbalance) and to simplify the evaluation metrics, described next.

4.2 Evaluation Metrics

Next, we describe the metrics used to evaluate and compare the proposed methods.

Coverage is the fraction of messages that a method is able to classify as either positive or negative in a given dataset. Ideally, polarity detection methods should retain high coverage to avoid bias in the results, due to the unidentified messages. For instance, suppose that a sentiment method has classified only 10% of a given set of tweets. The remaining 90% consisting of unidentified tweets may completely change the result, that is, whether the context drawn from tweets should be positive or negative. Therefore, having high coverage in data is essential in analyzing Web data. In addition to high coverage, it is also desirable that sentiment analysis methods have a high prediction performance as we will discuss next.

Although the coverage measure may be interpreted as similar to the well-known *recall* measure, it is quite different. While recall measures the percentage of items of a given class that have been correctly classified, usually misclassified items of this class are assigned to *some* other class provided to the classifier as training. On the contrary, coverage captures the phenomenon that some of the methods we study simply *abstain* to assign an item to any known class, a behavior usually not found in supervised classifiers.

Accuracy (A) represents the rate at which the method predicts results correctly. Since all datasets used in this paper are intentionally balanced (the same number of positive and negative messages), we can analyze the prediction performance of a method only looking at this metric.

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 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 $S_i(M) = \sum_{j=1}^{14} \sum_{k=1}^{18} \mathbf{1}_{M_i(j)>M_k(j)}$, where j is the dataset index (14 datasets) , i and k are the methods' index (18 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:

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

Thus, the larger $S_i(M)$ is, the better the $i-th$ method performs compared to the others. Notice that '$>$' means *statistical superiority* according to our statistical tests (see Section 4.4 for details); statistical ties are not counted.

4.3 Baselines

Next, we briefly describe three baseline methods we use to compare our proposed approach. The first two consists of intuitive strategies of combining methods and the third is a supervised learning method that explore information from the training data, but without combining any of the combined methods.

Majority Voting (Baseline I): An intuitive way to combine the base methods for sentiment analysis is to assign as the polarity of a message, the most frequent polarity detected by all base methods. More specifically, this combination works as follows: (1) take the result of each method applied to a single message; (2) check the most frequent polarity given by all methods; and (3) assign the most frequent polarity as the final polarity of this message.

Accuracy-Based Weighting Method (Baseline II): This baseline was proposed in a recent work [12, 1]. In this technique, a combination method can be built by given different weights to the methods based on their prediction performance previously calculated in some of the datasets. In that work, these values were obtained from the results of each method in the labeled datasets considered. Each method received a different weight based on the its average accuracy in these datasets. So, the method with the highest overall accuracy receives the highest weight and consequently would have more impact in the final polarity calculation of a single message. We can simply describe this process in three steps: (1) compute average accuracy of each method in all labeled datasets and create a ranking of methods; (2) take the result of each method applied to a single message; and (3) give the significance for each method based in the ranking built in the Step 1.

BoW Supervised Approach (Baseline III): Since we are using supervised methods to train our ensembles to combine the output of the sentiment classification base methods, for completeness, we also include as baseline a supervised method that works in the "bag-of-words" (BoW) representation of the messages. In this case, each message is represented as a vector of word weights, following the traditional "term frequency-inverted document frequency" (TF-IDF) weighting scheme. It should be notice that before calculating the weights we have applied standard preprocessing procedures such as stemming and the removal of stopwords. For learning we use again SVM, previously described, as it is known to be one of the best text classifiers found in the literature [16]. Particularly, in this case, we use a linear kernel as it is known to produce the best results for highly dimensional spaces as is the case for text classification. We left as future work the use of recently proposed improvements to BoW, such as the use of specific meta-features for sentiment analysis [5], as baseline or even to improve our approach.

4.4 Experimental Setup

Our experiments were run using a 5-fold cross validation setup, with best parameters for the learning methods found also using cross-validation within the training set. In case of SVM, the parameters include the cost (for the linear and RBF kernels) and *gamma* for the RBF kernel function. In case of RF, the parameters are the number of features ($numFeatures$), used in random selection of attributes, and the number of trees to be generated ($numTrees$). This procedure was repeated 10 times, therefore all results correspond to the average of 50 runs (test). To compare the average results on our cross-validation experiments, we assess the statisti-

cal significance of our results by means of a paired t-test with 95% confidence.

In the case of the base methods, the original output values (i.e. any positive, negative or "zero" values) were considered as positive, negative our neutral polarities. In particular, an output equals to zero was considered as a neutral polarity or "absence of opinion". Only in case of the Happiness Index, that output sentiments in the range from 1 to 9, we considered any message classified in the range of $[1..5)$ to be negative and in the range of $[5..9]$ to be positive. Particularly, this method also outputs a value of "undefined" when it cannot identify any polarity; we consider such value as 0 or neutral.

5. RESULTS

We start the analysis of our experiments by comparing the results of coverage *vs.* accuracy of all base methods along with our proposed ensembles and all the described baselines in each dataset. It is important to remind that in all datasets, every message has a positive or negative polarity. As mentioned before, methods such as SentiStrength and VADER were trained or evaluated with some of the datasets used in our study. For reasons of completeness, we chose to show the results of all methods in all datasets. However, we will identify the method in the figures with an asterisk ($*$) when a specific dataset was used as a training dataset of it. These results are shown in Figures 2 ((a)-(d), Figures 3 ((a)-(d)) and Figures 4 ((a)-(d)). The results for two of the evaluated datasets were omitted due to space constraints. In the Figures, we can see some general trends and behaviors, including:

• Regarding accuracy, as [12] showed, we can see that there is no a clear winner, although our ensembles are always among the best methods in all datasets.
• By construction, our ensembles and baselines II and III always produce full coverage. Baseline I does not have full coverage as it is based on majority voting – if the majority of the methods vote for a "neutral" position, this will be the final decision.
• In 11 and 10 out of 13 datasets, our ensembles have a higher accuracy when compared to Baselines II and III, respectively, while producing the same (full) coverage. This demonstrates the capability of the ensembles to adapt to the specific characteristics and idiosyncrasies of each dataset and to explore well the "opinions of the committee of expert base methods".
• Baseline III has a good performance in a few datasets such as Yelp and Reviews I, but in smaller datasets such as BBC, Myspace and Twitter I, it losses by a large margins (e.g., in BBC the accuracy is under 50%). This shows that our ensembles are also more robust to work in smaller datasets or with less training data.
• A few base methods have a high accuracy in several datasets but with a very poor (e.g., VADER) to medium coverage (e.g., SentiStrength).
• The performance of our ensembles is very similar in most datasets with a slight advantage for Ensemble I only in the Yelp dataset.
• The accuracy of Emoticons is good when the dataset has a higher proportion of messages with this type of characteristic, but its coverage is usually very low.
• As discussed before, some methods were executed in datasets used for their training or evaluation. However, as experimentally observed, these methods did not achieve the best accuracy and/or coverage rates even in these datasets. For instance, SentiStrength was trained using the following datasets: Twitter I, Myspace, Youtube, Digg, RW and BBC. In none of these datasets it obtained the best

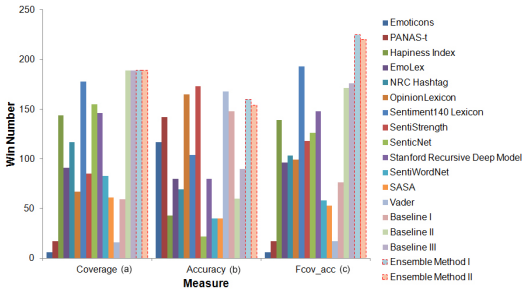


Figure 5: Winning for Coverage, Accuracy and Fcov_acc

results when compared with other base methods, baselines or our ensembles, even in similar dataset such as Twitter II and Stanford Twitter Corpus. So, we can say that combining methods trained in different domains, and also methods with no training (e.g., based on lexicons), can help to add useful knowledge to our ensembles in order to improve coverage and accuracy.

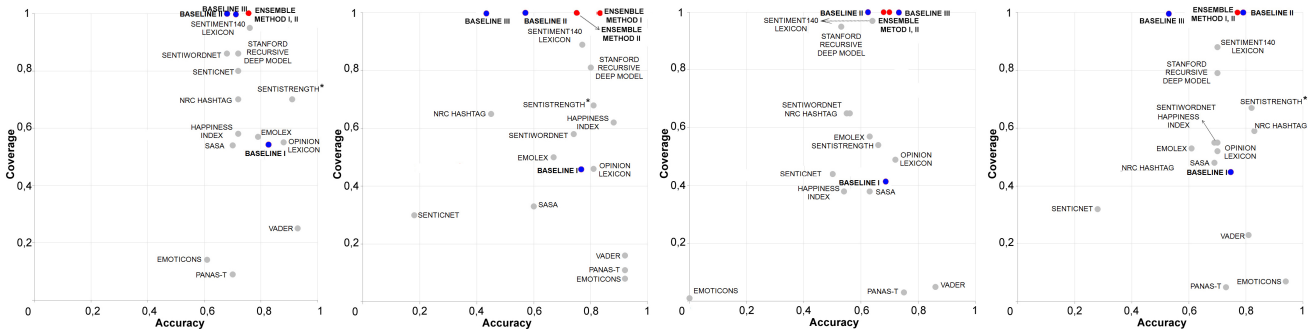
These results may be better summarized by looking at the winning numbers of the methods. This is shown in Figure 5, in which we consider three different performance metrics: accuracy, coverage and Fcov_acc. Fcov_acc, similarly to the well-known F1 measure largely used in classification tasks corresponds to a harmonic mean between accuracy and coverage, producing a single performance measure that covers both aspects, i.e., $Fcov\_acc = \frac{2*coverage*accuracy}{coverage+accuracy}$.

As it can been seen in Figure 5(a), the fact that our ensembles and Baselines II and III always produce full coverage makes them to achieve the largest possible winning number. Sentiment140Lexicon has the best comparative performance among the base methods in terms of coverage. But perhaps even more important, this full coverage capability of our proposals comes with *very low or no loss at all* in accuracy. As discussed before, this a hard task, as both goals may be conflicting. In fact as shown by the winning numbers of our ensembles (Figure 5(b)), they are among the most competitive methods in most datasets under this metric. Although there are a few competitive base methods such as Vader (the best under this metric), SentiStrength and SASA, our ensembles perform very close to them. This means that their performance are close to the best base method in each dataset or even better than them, such as in the Yelp and Reviews I datasets. We can also see that Baselines I and II do not perform well under this metric, with Baseline I being the best baseline, but still losing to several base methods and our ensembles.

Finally, when considering the combined metric $Fcov\_acc$, our ensembles stand out as the most competitive methods. We can also see that methods that are very competitive under only one of the previous metrics do not perform competitively when considering $Fcov\_acc$. We also verify that our baselines do not perform so well under this metric. Baselines II and III presented the 3th and 4th best results, but mostly due to their full coverage. However, all baselines still loose by large margins to our ensembles. Finally, regarding our ensembles, the one which used SVM as the combination technique obtained a very slight advantage in this scenario over the one that exploited Random Forests.

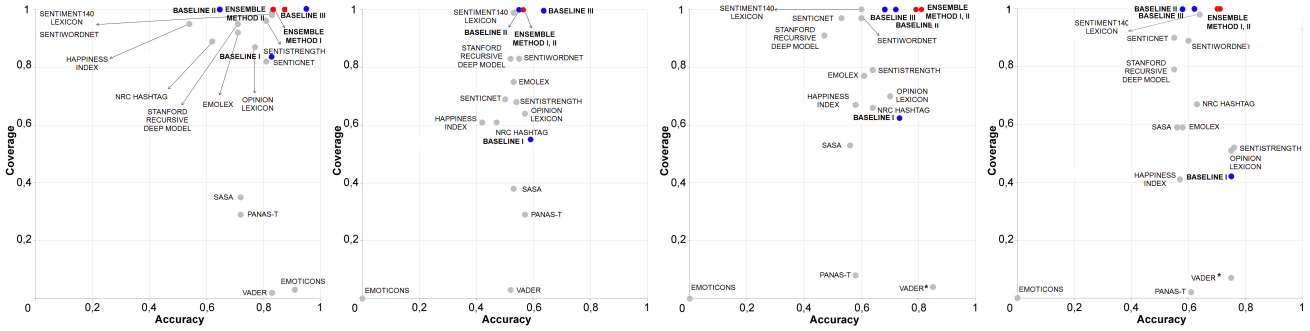
5.1 Impact of the Methods in the Ensemble

In this section, we proceed to analyze the impact of the methods' outputs in the final ensemble result. Table 2 shows the most dis-



(a) Twitter I (b) Myspace (c) Debate (d) Youtube

Figure 2: Coverage vs. Accuracy for base methods, baselines and the proposed ensembles in six OSN datasets.



(a) Yelp (b) Reviews I (c) Reviews II (d) Amazon

Figure 3: Coverage vs. Accuracy for base methods, baselines and the proposed ensembles in review datasets.

criminative feature (i.e., method output) for determining the final polarity, as given by the Variable Importance Measure (VIM) calculated by the RF classifier. Particularly, we use the Mean Decrease Accuracy (MDA) method to calculate this importance. MDA is determined during the out-of-bag error[11] calculation phase of the RF training procedure. The MDA quantifies the importance of an attribute by measuring the change in prediction accuracy, when the values of the attributes are randomly permuted compared to the original observations.

In consonance with previous discussions, no single feature (method) is dominant for all cases, with SentiStrength being the most discriminative one in 5 out of 13 cases, Stanford Recursive Deep Model and Sentiment140 Lexicon coming both in second place (3 cases each), followed by Sentiwordnet (2 cases) and Emolex (or NRC Emotion Lexicon) (1 case). Notice that the discriminative power of the methods within the ensemble does not necessarily correlate directly with their accuracy in the respective dataset[12], as this characteristic has to be considered in combination with the information brought by the other methods.

To further understand the impact of the individual base methods in the ensembles, we conducted an experiment using subsets of the most discriminative attributes, as determined by MDA, that occupy contiguous positions in the feature ranking (i.e., top 13 attributes, top 12 attributes, top 11 attributes and so on) (see Table 3). In other words, we run our ensemble first with all methods, next with the top 13 methods, top 12, and so on, until we consider the accuracy of the sole best method, in this case Sentiment140 Lexicon. We considered in this experiment only one of the datasets in which the ensembles produced the largest accuracy gain over the best base method, in this case, the Yelp dataset. In this experiment we used only the ensemble based on SVM.

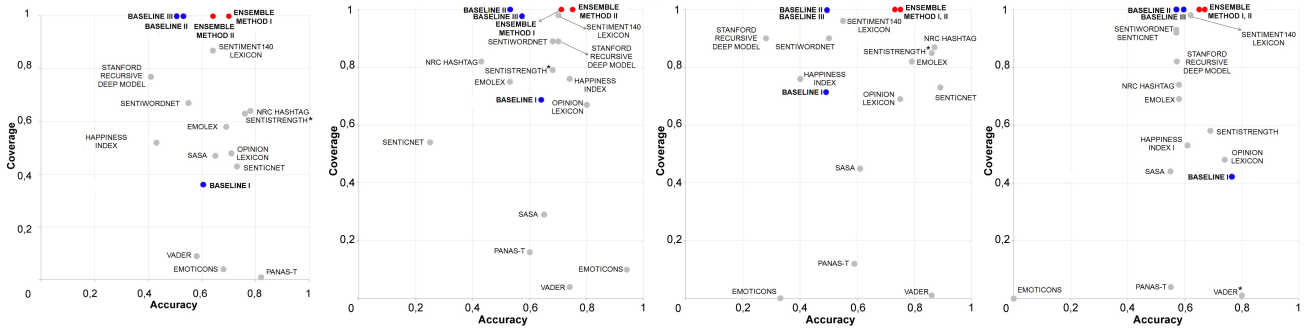
Figure 6 shows the accuracy for the different attribute subsets and for the version that considers all attributes (dashed line). We can notice that the variations are low when we start removing the least discriminative attributes but starts to accelerate when we start removing the most discriminative ones. One interesting observation is that gains are only observed over the best base method when we have at least six methods in the ensemble. But it is also interesting to notice that the inclusion of more base methods still keeps improving the results until certain limit is achieved. This encourages

Table 2: Methods contribution in each dataset

Dataset	Best attribute
Amazon	SentiStrength
BBC	Emolex
Debate	SentiStrength
Digg	Sentiwordnet
Myspace	Sentiment140 Lexicon
NYT	Sentiwordnet
Reviews I	Stanford Recursive Deep Model
Reviews II	Stanford Recursive Deep Model
RW	Sentiment140 Lexicon
Tweets	SentiStrength
Twitter	SentiStrength
Yelp	Sentiment140 Lexicon
Youtube	Stanford Recursive Deep Model
Stanford Twitter Corpus	SentiStrength

[11]The out-of-bag error is an estimate of the performance of the classifier in the portion of the labeled data that was not randomly chosen to train the trees in the forest.

[12]For instance, in the Debate dataset, SentiStrength is the most discriminative but not the most accurate method.



(a) Digg (b) RW (c) BBC (d) NYT

Figure 4: Coverage vs. Accuracy for base methods, baselines and the proposed ensembles in comments, forum opinions and news.

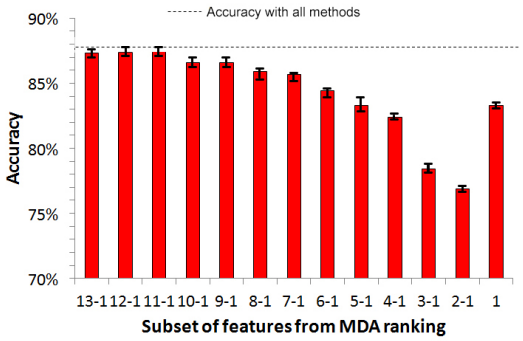


Figure 6: Average accuracy in Yelp when subsets of attributes with less importance are excluded. Results are shown with 95% confidence intervals.

Table 3: Ranking of Attributes - Yelp dataset

MDA ranking	Attribute/Method
14	EmoLex
13	PANAS-t
12	VADER
11	SentiWordNet
10	Emoticons
9	OpinionLexicon
8	NRC Hashtag
7	SentiStrength value 2
6	SASA
5	SentiStrength value 1
4	SenticNet
3	Stanford Recursive Deep Model
2	Hapiness Index
1	Sentiment140 Lexicon

the use of the largest number of base methods as possible since using an "off-the shelf" method comes with almost no cost.

6. RELATED WORK

The idea of combining different strategies to create a new sentiment analysis method has been recently explored by the research community. For instance, [6] take a step on this direction by combining two approaches for sentiment classification: machine learning and semantic-orientation. Unlike the machine learning approach, a semantically-oriented method does not require prior training, since it only needs to consider words expressing positive or negative sentiments. The process of combination consisted of the development of a lexicon-enhanced method to generate a set of positive and negative word measurements to use as new features. The evaluation of the proposed method was performed in a specific domain, a dataset of online product reviews such as books and and electronics. Similarly, [30] and [28] proposed ensemble classifiers for sentiment analysis of data from the Web using lexicon-bases approach and machine learning techniques such as Naïve Bayes and SVM, respectively. In another study, [3] propose the combination of six methods developed for these tasks and incorporate them as input features in a sentiment classifier using supervised learning algorithms. They evaluate their approach in a single specific domain, Twitter. Differently from ours, these studies only combine a few supervised methods, usually trained and tested in the same domain. Our work considers a much large universe of methods and datasets/domains and studies the possibility of reusing and combining knowledge from several different domains.

In [21], authors investigated approaches to detect the polarity of FourSquare tips using supervised (SVM, Maximum Entropy and Naive Bayes) and unsupervised (SentiWordNet lexicon) learning. They also 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.

[38] 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. Differently from our work, this method was proposed for the entity level, while ours was proposed for the sentence-level granularity (where the sentiment detected is associated with the whole sentence and not a single entity). Similarly, [22] 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). Differently from the above efforts, we propose a novel perspective by combining a series of "off-the-shelf" existing methods. Another major difference of our effort is that our evaluation is performed across multiple labeled datasets that cover multiple domains and social media sources.

7. CONCLUSIONS

We present a novel approach for sentence-level sentiment analysis based on the combination of several existing state-of-the-art, "off-the-shelf" sentiment analysis methods. We tested our solutions in a very rich experimentation environment, covering thirteen widely used methods and fourteen labeled datasets from many domains, including messages from social networks, movie and product reviews, opinions and comments in news articles. Our experimental results demonstrate that our approach achieves high significant improvements over the base methods and baselines, both in terms of accuracy and coverage, a hard objective. Moreover, by combining many "off-the-shelf" methods, our ensembles can be very resilient to vocabulary size of the input text as well as to the amount of available training.

Acknowledgements

This research is funded by grants from CNPq, CAPES and FAPEMIG.

8. REFERENCES

[1] M. Araújo, P. Gonçalves, F. Benevenuto, and M. Cha. ifeel: A system that compares and combines sentiment analysis methods. In *WWW*, 2014.

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

[3] F. Bravo-Marquez, M. Mendoza, and B. Poblete. Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In *WISDOM*, 2013.

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

[5] S. Canuto, M. A. Gonçalves, and F. Benevenuto. Exploiting new sentiment-based meta-level features for effective sentiment analysis. In *Proc. of WSDM*, 2016.

[6] Y. Dang, Y. Zhang, and H. Chen. A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 25, 2010.

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

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

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

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

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

[12] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha. Comparing and combining sentiment analysis methods. In *Proc. COSN*, 2013.

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

[14] M. Hu and B. Liu. Mining and summarizing customer reviews. Proc. KDD'04, pages 168–177, 2004.

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

[16] T. Joachims. *Learning to Classify Text using Support Vector Machines*. Kluwer, 2002.

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

[18] S. Mohammad. #emotional tweets. In *SEM*, 2012.

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

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

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

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

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

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

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

[26] R. Plutchik. *A general psychoevolutionary theory of emotion*, pages 3–33. Academic press, New York, 1980.

[27] F. Ribeiro, M. Araújo, P. Gonçalves, F. Benevenuto, and M. A. Gonçalves. A benchmark comparison of state-of-the-practice sentiment analysis methods. *arXiv preprint arXiv:1512.01818*, 2015.

[28] A. Sharma and S. Dey. A boosted svm based ensemble classifier for sentiment analysis of online reviews. *SIGAPP Appl. Comput. Rev.*, 13, 2013.

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

[30] Y. Su, Y. Zhang, D. Ji, Y. Wang, and H. Wu. Ensemble learning for sentiment classification. In *CLSW*, 2012.

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

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

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

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

[35] D. Wolpert. Stacked generalization. *Neural Networks*, 5, 1992.

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

[37] Yelp. Yelp dataset challenge. http://www.yelp.com/dataset_challenge, 2014. Accessed April 23, 2014.

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