Combining Strengths, Emotions and Polarities for Boosting Twitter Sentiment Analysis

Felipe Bravo-Marquez
PRISMA Research Group
Department of Computer Science, University of Chile, Chile
Yahoo! Labs Santiago, Chile
fbravo@dcc.uchile.cl

Marcelo Mendoza
Universidad Técnica Federico Santa María, Chile
Yahoo! Labs Santiago, Chile
marcelo.mendoza@usm.cl

Barbara Poblete
PRISMA Research Group
Department of Computer Science, University of Chile, Chile
Yahoo! Labs Santiago, Chile
bpoblete@dcc.uchile.cl

ABSTRACT

Twitter sentiment analysis or the task of automatically retrieving opinions from tweets has received an increasing interest from the web mining community. This is due to its importance in a wide range of fields such as business and politics. People express sentiments about specific topics or entities with different strengths and intensities, where these sentiments are strongly related to their personal feelings and emotions. A number of methods and lexical resources have been proposed to analyze sentiment from natural language texts, addressing different opinion dimensions. In this article, we propose an approach for boosting Twitter sentiment classification using different sentiment dimensions as meta-level features. We combine aspects such as opinion strength, emotion and polarity indicators, generated by existing sentiment analysis methods and resources. Our research shows that the combination of sentiment dimensions provides significant improvement in Twitter sentiment classification tasks such as polarity and subjectivity.

Categories and Subject Descriptors

I.2.7.7 [**Artificial Intelligence**]: Natural Language Processing—*Text Analysis*

General Terms

Experimentation, Measurement

Keywords

Sentiment Classification, Twitter, Meta-level features

1. INTRODUCTION

Inherent in human nature is the need to express particular points of view and feelings about specific topics or entities. Opinions reveal beliefs about specific matters commonly considered to be subjective.

Social media has opened new possibilities for people to interact. Microblogging platforms allow real-time sharing of comments and opinions. Twitter, which has become the most popular microblogging platform, has millions of users that spread millions of personal posts on a daily basis. The rich and enormous volume of data propagated through social media offers enormous opportunities for the study of social human subjectivity.

Manual classification of millions of posts for opinion mining tasks is an unfeasible effort at human scale. Several methods have been proposed to automatically infer human opinions from natural language texts. Due to the inherent subjectivity of the data, this problem is still an open problem in the field.

Opinions are multidimensional semantic artifacts. When people are exposed to information regarding a topic or entity, they normally respond to this external stimuli by developing a personal point of view or orientation. This orientation reveals how the opinion holder is polarized by the entity. Additionally, people manifest emotions through opinions, which are the driving forces behind motivations and personal dispositions. That means that emotions and polarities are mutually influenced by each other, conditioning opinion intensities and emotional strengths.

Computational sentiment analysis methods attempt to measure different opinion dimensions. A number of methods for polarity estimation have been proposed in [3, 6, 7, 16] discussed in depth in Section 2. By transforming polarity estimation into a classification problem with three polarity classes -positive, negative and neutral-supervised and unsupervised approaches have been explored to fulfill this task. In the case of the unsupervised approaches, a number of lexicon resources with with positive and negative scores for words have been released. Another related task is the detection of *subjectivity*, which is the specific task of separating factual from opinionated text. This problem has also been addressed by using supervised approaches [25]. Opinion intensities (strengths) have also been measured. From a strength scored method, SentiStrength [23] can estimate positive and negative strength scores at sentence level. Finally, emotion estimation has also been addressed by developing lexicons. The Plutchik's wheel of emotions was proposed in [21]. The wheel is composed by four pairs of opposite emotion states: **joy-trust**, **sadness-anger**, **surprise-fear**, and **anticipation-disgust**. Mohammad et.al [14] labeled a number of words according to Plutchik emotional categories, developing the NRC word-emotion association lexicon.

According to the previous paragraphs, we can see that sentiment analysis tools focus on different scopes within opinions. Although these scopes are very difficult to categorize explicitly, we propose the following categories:

1. **Polarity**: These methods and resources aim towards extracting polarity information from a passage. Polarity-oriented

WISDOM'13, August 11 2013, Chicago, USA.

methods normally return a categorical variable whose possible values are positive, negative and neutral. On the other hand, polarity-oriented lexical resources are composed by lists of positive and negative words.

2. **Emotion**: Methods and resources focused on extracting emotion or mood states from a text passage. An emotion-oriented method should classify the message to an emotional category such as sadness, joy, surprise, among others. An emotion-oriented lexical resources should provide a list of words or expressions marked according to different emotion states.

3. **Strength**: These methods and resources provide intensity levels according to a certain sentiment dimension which can have a polarity or an emotional scope. Strength-oriented methods return different numerical scores indicating the intensity or the strength of an opinion dimension expressed in a text passage. For instance, numerical scores indicating the level of positivity, negativity or another emotional dimension. Strength-oriented lexical resources provide lists of opinion words together with intensity scores regarding an opinion dimension.

In this article we propose to efficiently combine existing sentiment analysis methods and resources focused the main scopes discussed above. Our goal is to improve two major sentiment analysis tasks: 1) Subjectivity classification, and 2) Polarity classification. We combine all of these aspects as input features in a sentiment classifier using supervised learning algorithms. To validate our approach we evaluate our classifiers on two existing datasets. Our results show that the composition of these features achieves significant improvements over single approaches. This, indicates that strength, emotion and polarity-based resources are complementary, addressing different dimensions of the same problem. Therefore, a tandem approach should be more appropriate.

To the best of our knowledge, this is the first study to combine polarity, emotion, and strength oriented sentiment analysis lexical resources with existing opinion mining methods as meta-level features for boosting sentiment classification performance.

This article is organized as follows. In Section 2 we provide a review of existing lexical resources and discuss related work on Twitter sentiment analysis. In Section 4.4 we describe our approach for Twitter sentiment classification as well as the features that are used in our classification scheme. The experimental results are presented in Section 4. Finally, we conclude in Section 5 with a brief discussion.

2. RELATED WORK

2.1 Lexical Resources for Sentiment Analysis

The development of lexical resources for sentiment analysis has gathered attention from the computational linguistic community. Wilson et al. [25] labeled a list of English words in positive and negative categories, releasing the Opinion Finder lexicon. Bradley and Lang [3] released ANEW, a lexicon with affective norms for English words. The application of ANEW to Twitter was explored by Nielsen [16], leveraging the AFINN lexicon. Esuli and Sebastiani [6] and later Baccianella et al. [1] extended the well known Wordnet lexical database [13] by introducing sentiment ratings to a number of synsets, creating SentiWordnet. The development of lexicon resources for strength estimation was addressed by Thelwall et al. [23], leveraging SentiStrength. Finally, NRC, a lexicon resource for emotion estimation was recently released by Mohammad and Turney [14], where a number of English words were tagged with emotion ratings, according to the emotional wheel taxonomy introduced by Plutchik [21].

Besides the syntactic-level resources for sentiment analysis presented above, other type of resources have been elaborated for a semantic-level analysis refered as concept-based. Concept-based approaches conduct a semantic analysis of the text using semantic knowledge bases such as web ontologies [17] and semantic networks [18]. In this manner, concept-based methods allow the detection of subjective information which can be expressed implicitly in a text passage. A publicly available concept-based resource to extract sentiment information from common sense concepts is *SenticNet 2*[1]. This resource was built using both graph-mining and dimensionality-reduction techniques [4].

2.2 Twitter Sentiment Analysis

Twitter users tend to post opinions about products or services [19]. **Tweets** (user posts on Twitter) are short and usually straight to the point messages. Therefore, tweets are considered as an interesting resource for sentiment analysis. Common tasks of opinion mining that can be applied to Twitter data are sentiment classification and opinion identification. As Twitter messages are at most, 140-characters long, a sentence-level classification approach can be adopted, assuming that tweets express opinions about one single entity. Furthermore, retrieving messages from Twitter is a straightforward task, through the use of the Twitter API.

As the creation of a large corpus of manually-labeled data for sentiment classification tasks involves significant human effort, a number of studies has explored the use of emoticons as labels [7, 5, 22]. The use of emoticons assumes that they could be associated with positive and negative polarities regarding the subject mentioned in the tweet. Although there are cases where this basic assumption holds, there are some cases where the relation between the emoticon and the tweet subject is not clear. Hence, the use of emoticons as tweet's labels can introduce noise. However, this drawback is counterweighted by the large amount of data that can easily be labeled. In this direction, Go et al. [7] reported the creation of a large Twitter dataset with more than $1,600,000$ tweets. By using standard machine learning algorithms, accuracies greater than 80% were reported for label prediction. Recently, Liu et al. [11] explored the combination of emoticon labels and human labeled tweets in language models, outperforming previous approaches.

Sentiment Lexical resources were used as features in a supervised classification scheme in [10, 9, 26] among other works. In [10] a supervised approach for Twitter sentiment classification based on linguistic features was proposed. In addition of using n-grams and part-of-speech tags as features, the authors used sentiment lexical resources and aspects particular from microblogging platforms such as the presence of emoticons, abbreviations and intensifiers. A comparison of the different types of features was carried out, showing that although features created from the opinion lexicon are relevant, microblogging-oriented features are the most useful.

Recently, The Semantic Evaluation (SemEval) workshop has organized a Sentiment Analysis in Twitter task (SemEval-2013)[2]. This task provides training and testing datasets for Twitter sentiment classification at both expression and message levels [24].

For further details about sentiment analysis methods and applications we refer the reader to the survey of Pang and Lee [20] and to the book of Liu [12].

[1] http://sentic.net/

[2] http://www.cs.york.ac.uk/semeval-2013/task2/

3. CLASSIFICATION APPROACH

In this section we describe the proposed Twitter sentiment classification approach. We consider two classification tasks: subjectivity and polarity classification. In the former, tweets are classified as subjective (non neutral) or objective (neutral), and in the latter as positive or negative. Moreover, positive and negative tweets are considered as subjective.

We propose a supervised approach for which we model each tweet as a vector of sentiment features. Additionally, a dataset of manually annotated tweets is required for training and evaluation purposes. Once the feature vectors of all the tweets from the dataset have been extracted, they are used together with the annotated sentiment labels as input for supervised learning algorithms. Several learning algorithms can be used to fullfill this task, eg. naive Bayes, SVM, decision trees. Finally, the resulting learned function can be used to infer automatically the sentiment label regarding an unseen tweet. All the resources and methods considered in this work are publicly available, facilitating repeatability of our experiments.

In contrast to the common text classification approach, in which the words contained within the passage are used as features (e.g., unigrams, n-grams), our meta-level features are based on existing lexical resources and sentiment analysis methods. These resources and methods, summarize the main efforts discussed in Section 2, and cover three different dimensions of the problem: polarity, strength, and emotions.

From each lexical resource we calculate a number of features according to the number of matches between the words from the tweet and the words from the lexicon. If the lexical resource provides strength values associated to the words, then features are calculated through a weighted sum. Finally, for each sentiment analysis method, its outcome is included as a dimension in the feature vector. The features are summarized in Table 1, and are described together with their respective methods and resources in the following parapraphs.

OpinionFinder Lexicon.

The **OpinionFinder Lexicon** (**OPF**) is a polarity oriented lexical resource created by Wilson et al. [25]. It is an extension of the Multi-Perspective Question-Answering dataset (MPQA), that includes phrases and subjective sentences. A group of human annotators tagged each sentence according to the polarity classes: positive, negative, neutral[3]. Then, a pruning phase was conducted over the dataset to eliminate tags with low agreement. Thus, a list of sentences and single words was consolidated, with their polarity tags. In this study we consider single words (unigrams) tagged as positive or negative, that correspond to a list of $6,884$ English words. We extract from each tweet two features related to the OpinionFinder lexicon, **OpinionFinder Positive Words** (OPW) and **OpinionFinder Negative Words** (ONW), that are the number of positive and negative words of the tweet that matches the OpinionFinder lexicon, respectively.

AFINN Lexicon.

This lexicon is based on the **Affective Norms for English Words** lexicon (ANEW) proposed by Bradley and Lang [3]. ANEW provides emotional ratings for a large number of English words. These ratings are calculated according to the psychological reaction of a person to a specific word, being "valence" the most useful value for sentiment analysis. "Valence" ranges in the scale pleasant-unpleasant. ANEW was released before the rise of microblogging and hence, many slang words commonly used in social media were not included. Considering that there is empirical evidence about significant differences between microblogging words and the language used in other domains [2] a new version of ANEW was required. Inspired in ANEW, Nielsen [16] created the **AFINN** lexicon, which is more focused on the language used in microblogging platforms. The word list includes slang and obscene words as also acronyms and web jargon. Positive words are scored from 1 to 5 and negative words from -1 to -5, reason why this lexicon is useful for strength estimation. The lexicon includes $2,477$ English words. We extract from each tweet two features related to the AFINN lexicon, **AFINN Positivity** (APO) and **AFINN Negativity** (ANE), that are the sum of the ratings of positive and negative words of the tweet that matches the AFINN lexicon, respectively.

SentiWordNet Lexicon.

SentiWordNet 3.0 (**SWN3**) is a lexical resource for sentiment classification introduced by Baccianella et al. [1], that it is an improvement of the original SentiWordNet proposed by Esuli and Sebastiani [6]. SentiWordNet is an extension of **WordNet**, the well-known English lexical database where words are clustered into groups of synonyms known as **synsets** [13]. In SentiWordNet each synset is automatically annotated in the range $[0, 1]$ according to positivity, negativity and neutrality. These scores are calculated using semi-supervised algorithms. The resource is available for download[4]. In order to extract strength scores from SentiWordNet, we use the word's scores to compute a real value from -1 (extremely negative) to 1 (extremely positive), where neutral words receive a zero score. We extract from each tweet two features related to the SentiWordnet lexicon, **SentiWordnet Positiveness** (SWP) and **SentiWordnet Negativeness** (SWN), that are the sum of the scores of positive and negative words of the tweet that matches the SentiWordnet lexicon, respectively.

SentiStrength Method.

SentiStrength is a lexicon-based sentiment evaluator that is specially focused on short social web texts written in English [23]. SentiStrength considers linguistic aspects of the passage such as a negating word list and an emoticon list with polarities. The implementation of the method can be freely used for academic purposes and is available for download[5]. For each passage to be evaluated, the method returns a positive score, from 1 (not positive) to 5 (extremely positive), a negative score from -1 (not negative) to -5 (extremely negative), and a neutral label taking the values: -1 (negative), 0 (neutral), and 1 (positive). We extract from each tweet three features related to the SentiStrength method, **SentiStrength Negativity** (SSN) and **SentiStrength Positivity** (SSP), that correspond to the strength scores for the negative and positive classes, respectively, and **SentiStrength Polarity** (SSPOL), that is a polarity-oriented feature corresponding to the neutral label.

Sentiment140 Method.

Sentiment140[6] is a Web application that classifies tweets according to their polarity. The evaluation is performed using the distant supervision approach proposed by Go et al. [7] that was previously discussed in the related work section. The approach relies on supervised learning algorithms and due to the difficulty of obtaining a large-scale training dataset for this purpose, the problem is tackled using positive and negative emoticons and noisy labels. The

[3] The lexicon also includes 17 words having mixed positive and negative tags tagged as "both", which were omitted in this work.

[4] `http://sentiwordnet.isti.cnr.it/`

[5] `http://sentistrength.wlv.ac.uk/`

[6] `http://www.sentiment140.com/`

Scope	Feature	Source	Description	Range
Polarity	SSPOL	SentiStrength	method label (negative, neutral, positive)	$\{-1, 0, +1\}$
	S140	Sentiment140	method label (negative, neutral, positive)	$\{-1, 0, +1\}$
	OPW	OpinionFinder	number of positive words that matches OpinionFinder	$\{0, 1, ..., n\}$
	ONW		number of negative words that matches OpinionFinder	$\{0, 1, ..., n\}$
Strength	SSP	SentiStrength	method score for the positive category	$\{1, \ldots, 5\}$
	SSN		method score for the negative category	$\{-5, \ldots, -1\}$
	SWP	SentiWordNet	sum of the scores for the positive words that matches the lexicon	$\{0, \ldots, n\}$
	SWN		sum of the scores for the negative words that matches the lexicon	$\{0, \ldots, n\}$
	APO	AFINN	sum of the scores for the positive words that matches the lexicon	$\{0, ..., n\}$
	ANE		sum of the scores for the negative words that matches the lexicon	$\{-n, ..., 0\}$
Emotion	NJO	NRC	number of words that matches the joy word list	$\{0, 1, ..., n\}$
	NTR		... matches the trust word list	$\{0, 1, ..., n\}$
	NSA		... matches the sadness word list	$\{0, 1, ..., n\}$
	NANG		... matches the anger word list	$\{0, 1, ..., n\}$
	NSU		... matches the surprise word list	$\{0, 1, ..., n\}$
	NFE		... matches the fear word list	$\{0, 1, ..., n\}$
	NANT		... matches the anticipation word list	$\{0, 1, ..., n\}$
	NDIS		... matches the disgust word list	$\{0, 1, ..., n\}$

Table 1: Features can be grouped into three classes having as scope Polarity, Strength, and Emotion, respectively.

method provides an API[7] that allows to classify tweets to polarity classes positive, negative and neutral. We extract from each tweet one feature related to the Sentiment140 output, **Sentiment140** class (S140), that corresponds to the output returned by the method.

NRC Lexicon.

NRC is a lexicon that includes a large set of human-provided words with their emotional tags. By conducting a tagging process in the crowdsourcing Amazon Mechanical Turk platform, Mohammad and Turney [14] created a word lexicon that contains more than 14,000 distinct English words annotated according to the Plutchik's wheel of emotions. These words can be tagged to multiple categories. Eight emotions were considered during the creation of the lexicon, joy-trust, sadness-anger, surprise-fear, and anticipation-disgust, which compounds four opposing pairs. Additionally, NRC words are tagged according to polarity classes positive and negative, which are not considered in this work. The word list is available under request[8]. We extract from each tweet eight features related to the NRC lexicon, **NRC Joy** (NJO), **NRC Trust** (NTR), **NRC Sadness** (NSA), **NRC Anger** (NANG), **NRC Surprise** (NSU), **NRC Fear** (NFE), **NRC Anticipation** (NANT), and **NRC Disgust** (NDIS), that are the number of words of the tweet that matches each category.

4. EXPERIMENTS

4.1 Lexical Resource Interaction

In this section we study the interaction of words between the different lexical resources: SWN3, NRC, OpinionFinder, and AFINN. The number of words that overlap between each pair of resources is shown in Table 2. From the table we can see that SWN3 is much larger than the other resources. Nevertheless, the resource includes many neutral words provided by WordNet that lack of useful information for sentiment analysis purposes.

	SWN3	NRC	AFINN	OPFIND
SWN3	$147,306$	×	×	×
NRC	$13,634$	$14,182$	×	×
AFINN	$1,783$	$1,207$	$2,476$	×
OPFIND	$6,199$	$3,596$	$1,245$	$6,884$
Distinct Words	$149,114$			

Table 2: Intersection of words between different Lexical Resources

	SWN3	NRC	AFINN	OPFIND
SWN3	$33,313$	×	×	×
NRC	$2,932$	$3,071$	×	×
AFINN	$1,203$	721	$1,871$	×
OPFIND	$3,703$	$1,658$	900	$4,311$
Distinct Words	$34,649$			

Table 3: Intersection of non-neutral words

Table 3 shows the overlap of words after discarding the neutral words from SentiWordNet, the neutral and mixed words from OpinionFinder and the words without emotion tags from NRC. We can see that although the size of SWN3 was strongly reduced it stills has much more words than the others. The interaction of all the non-neutral words, is better represented in the Venn diagram shown in Figure 1. From the diagram we can see that SWN3 covers the majority of the words within the lexical resources. However, if we discard SWN3 we keep with three different sets of words: NRC having words related to emotions, OpinionFinder whose words are related to polarity, and AFINN whose words are also related to polarity with additional strength information. These resources, in addition to having different sentiment scopes, cover many different words from each other. It is also revealed from the figure that the AFINN lexicon, despite being smaller, contains some words that are not included in SWN3 nor in the others. We inspected these words included only in AFINN and we found many Internet acronyms and slang words such as “lmao”, “lol”, “rofl”, “wtf” among other expressions.

We compare the sentiment values assigned by each lexical resource to a sample of words that appear in the intersection of all lexicons in Table 4. We can observe a tendency of the different resources to support each other, eg. words that received negative strength values from SWN3 and AFINN normally receive a nega-

[7] http://help.sentiment140.com/api

[8] mailto: saif.mohammad@nrc-cnrc.gc.ca

word	SWN3	AFINN	OPFIND	NRC
abuse	-0.51	-3	negative	ang,disg,fear,sadn
adore	0.38	3	positive	ant, joy, trust
cheer	0.13	2	positive	ant, joy, surp, trust
shame	-0.52	-2	negative	digs,fear,sadn
stunned	-0.31	-2	positive	fear, surpr
sympathy	-0.13	2	negative	sadn
trust	0.23	1	positive	trust
ugly	-0.63	-3	negative	disg
wonderful	0.75	4	positive	joy, surp, trust

Table 4: Sentiment Values of Words included in all the Resources

tive tag from OpinionFinder and are associated as well with negative NRC emotions states. A similar pattern is observed for positive words. However, we can also see controversial examples such as words “stunned” and “sympathy” which receive contrary sentiment values from polarity and strength oriented resources. These words may be used to express either positive and negative opinions, depending on the context. Considering that it is very hard to associate them to a single polarity class, we think that emotion tags explain in a better manner the diversity of sentiment states triggered by these kind of words.

These insights indicate that the resources considered in this work complement each other, providing different sentiment information.

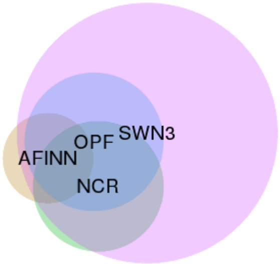

Figure 1: Non-neutral words interaction Venn diagram

4.2 Training and Testing Datasets

We consider two collections of tweets for our experiments: *Stanford Twitter Sentiment* (STS) [9] which was used by Go et al. [7] in their experiments, and *Sanders*[10]. Each tweet includes a **positive**, **negative** or **neutral** tag. Table 5 summarizes both datasets.

Negative and positive tweets were considered as subjective. Neutral tweets were considered as objective. Subjective/objective tags favor the evaluation of subjectivity detection. For polarity detection tasks, positive and negative tweets were considered, discarding neutral tweets.

Both datasets were balanced. Class imbalance was tackled by sampling 139 subjective tweets in STS from the 359 positive and negative tagged tweets, achieving a balance with the 139 neutral tweets. In the case of Sanders, the neutral collection was sampled recovering 1,196 tweets from the 2,429 neutral tweets achieving a balance with the 1,196 positive and negative tagged tweets. A similar process was conducted for class imbalance in the case of polarity recovering 354 and 1,120 tweets from STS and Sanders respectively. Table 6 summarizes the balanced datasets.

	STS	Sanders
#negative	177	636
#neutral	139	2,429
#positive	182	560
#total	498	3,625

Table 5: Datasets Statistics

Subjectivity	STS	Sanders
#neutral	139	1,196
#subjective	139	1,196
#total	278	2,392
Polarity	STS	Sanders
#negative	177	560
#positive	177	560
#total	354	1,120

Table 6: Balanced Datasets

4.3 Feature Analysis

For each tweet of the two datasets we calculated the features summarized in Table 1. In a first analysis we explored how well each feature splits each dataset regarding polarity and subjectivity detection tasks. We do this by calculating the information gain criterion of each feature in each category. The information gain criterion measures the reduction of the entropy within each class after performing the best split induced by the feature. Table 7 shows the information gain values obtained.

Scope	Feature	Subjectivity		Polarity	
		STS	Sanders	STS	Sanders
Polarity	SSPOL	**0.179**	0.089	**0.283**	**0.192**
	S140	**0.103**	0.063	**0.283**	**0.198**
	OPW	0.088	0.024	0.079	0.026
	ONW	0.097	0.024	**0.135**	0.075
Strength	SSP	0.071	0.037	**0.200**	**0.125**
	SSN	0.090	0.044	**0.204**	**0.118**
	SWN	0.090	0.023	**0.147**	0.089
	SWP	**0.104**	0.030	0.083	0.015
	APO	0.088	0.024	0.079	0.026
	ANE	**0.134**	0.048	**0.200**	**0.143**
Emotion	NJO	0.000	0.000	0.055	0.065
	NTR	0.000	0.000	0.000	0.000
	NSA	0.000	0.017	0.000	0.056
	NANG	0.000	0.016	0.046	0.055
	NSU	0.000	0.000	0.000	0.017
	NFE	0.000	0.008	0.039	0.024
	NANT	0.000	0.000	0.000	0.000
	NDIS	0.000	0.014	0.056	0.030

Table 7: Feature information gain for each sentiment analysis task. Bold fonts indicate the best splits.

As Table 7 shows, the best polarity splits are achieved by using the outcomes of the methods(see SSPOL, S140, SSP, and SSN). SentiWordNet, OpinionFinder and AFINN-based features are useful for negative polarity detection. These features are also useful for subjectivity detection. In addition, we can observe that the best splits are achieved in the STS. The Sanders dataset is hard to split.

[9] `http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip`

[10] `http://www.sananalytics.com/lab/twitter-sentiment/`

By analyzing the scope, we can observe that polarity-based features are the most informative. This fact is intuitive because the target variables belong to the same scope. Finally, although emotion features provide almost no information for subjectivity, some of them like joy, sadness and disgust are able to provide some information for the polarity classification task.

We also explored feature-subsets extracted by the correlation feature selection algorithm (CFS) [8]. This algorithm is a best-first feature selection method that considers different types of correlation as selection criteria. Selected features for each classification task on the two datasets are displayed in Table 8.

	Neu.STS	Neu.San	Pol.STS	Pol.San
ANE	✓	✓	✓	✓
APO	✓		✓	✓
ONW	✓		✓	✓
OPW	✓			
NJO				✓
S140	✓	✓	✓	✓
SSN			✓	✓
SSP			✓	
SSPOL	✓	✓	✓	✓
SWN	✓		✓	✓
SWP	✓	✓		

Table 8: Selected Features by CFS algorithm

From the table we can see that the two features that come from polarity-oriented methods (S140 and SSPOL), are selected in all the cases. We can also observe that the algorithm tends to include more features for polarity than for subjectivity classification in the Sanders dataset. Regarding the emotion-oriented features, the only feature that is selected by the CFS algorithm is the NJO feature. Moreover, the feature is only selected for the polarity task on the Sanders dataset. These results agree with the information gain values discussed above, and support the evidence that most of the features are more informative for polarity than for subjectivity classification.

4.4 Classification Results

We evaluate a number of learning algorithms on the STS and Sanders datasets, for both subjectivity and polarity detection. We conducted a 10-fold cross-validation evaluation. As learning algorithms we considered CART, J48, Naive Bayes, Logistic regression, and RBF SVMs. The experiments were performed using R 2.15.2 packages using the following packages: **rpart**[11] for CART, **rWeka**[12] for J48 and Logistic regression, and **e1071**[13] for Naive Bayes and SVMs.

The performance of many machine learning techniques are highly dependent on the calibration of parameters. Different parameters such as the min-split criterion for trees, γ and C for radial SVMs, among others were tuned using a grid-search procedure with nested 10-fold cross validation.

An example of the tuning process for the radial SVM for polarity classification on the Sanders dataset is shown in Figure 2. The x-axis and y-axis of the chart represent the **gamma** and **cost** parameter respectively. The color of the region corresponds to the classification error obtained using the corresponding parameter values. From the figure we can see that the classification performance varies considerably for different parameters values. Therefore, it is important to remark that the tuning process of machine learning parameters is crucial to obtain accurate classifiers.

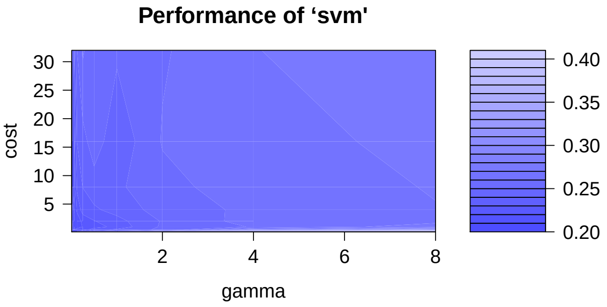

Figure 2: RBF SVM parameters performance for Polarity classification on Sanders dataset

A relevant issue regarding our feature-set is its heterogeneity. Most of the features are numerical but two of them are categorical (S140 and SSPOL). A number of supervised learning algorithms are not capable to handle mixed-type features and hence some transformations must be applied before the learning task. For CART and J48 the numerical features are “discretized” as part of the learning process. Naive Bayes handles numerical features by assuming Gaussian distributions. For the SVM and Logistic regression algorithms, we transformed categorical features into dummy variables by mapping the c possible categories to binary values using 1-of-c encoding. Afterwards, these binary variables are handled as numerical features by these learning algorithms.

The performance of our classifiers in both classification tasks is compared with baselines created from isolated methods or resources. In the subjectivity task we considered the features Sent140 and SSPOL as the **Baseline.1** and **Baseline.2**, respectively. For both methods the positive and negative outputs are interpreted as subjective. We chose these features because they are the only ones which explicitly distinguish between subjective and neutral tweets.

Nevertheless, these methods could not be used as baselines for the polarity task, because it is not clear how to handle their neutral outcomes in this context. Therefore, we created two categorical variables whose outcomes are restricted to **positive** and **negative** values. The **Baseline.1** is calculated from strength features SSP and SSN as follows: if the sum of **SSP** and **SSN** is positive the baseline takes a **positive** value, otherwise takes a **negative** value. Then, the second baseline (**Baseline.2**), is calculated in the same manner from the features **APO** and **ANE**. Considering that for SentiStrength and AFINN, positivity and negativity are assessed independently, basically what we are doing in our baselines is to combine these dimensions into categorical variables that are constrained to distinguish between positive and negative tweets.

In addition of the feature-subset obtained by the best first CFS algorithm, we also explored feature-subsets constrained to the scope. Thus, we evaluate five groups of features -all, best first, polarity, strength, and emotion- and for each group five learning algorithms -CART, J48, naive Bayes, logistic regression, and SVMs-.

We consider as performance measures **accuracy**, **precision**, **recall** and F_1. We believe that the costs of misclassifying each type of observation for each classification task are equally important. Thus, considering that our datasets are balanced, we will pay more attention to the measures accuracy and F_1 than to precision and

[11] `http://cran.r-project.org/web/packages/rpart/`

[12] `http://cran.r-project.org/web/packages/RWeka`

[13] `http://cran.r-project.org/web/packages/e1071/`

Dataset		STS				Sanders			
Features	Methods	accuracy	precision	recall	F_1	accuracy	precision	recall	F_1
Baseline.1	Sent140	0.655	0.812	0.403	0.538	0.615	0.686	0.424	0.524
Baseline.2	SSPOL	0.734	0.712	0.784	0.747	0.659	0.632	0.760	0.690
All	CART	0.694	0.696	0.691	0.693	0.686	0.688	0.683	0.685
	J48	0.716	0.742	0.662	0.700	0.694	0.703	0.673	0.688
	Naive Bayes	0.737	0.784	0.655	0.714	0.649	0.718	0.491	0.583
	Logistic	0.755	0.775	0.719	0.746	0.678	0.679	0.675	0.677
	SVM	0.763	0.766	0.755	0.761	**0.701**	0.696	0.713	0.705
Best.First	CART	0.730	0.735	0.719	0.727	0.677	0.639	**0.816**	**0.717**
	J48	0.701	0.730	0.640	0.682	0.673	0.639	0.796	0.709
	Naive Bayes	0.759	0.821	0.662	0.733	0.651	**0.727**	0.483	0.581
	Logistic	0.748	0.756	0.734	0.745	0.683	0.676	0.704	0.690
	SVM	0.773	0.757	**0.806**	**0.780**	0.680	0.663	0.732	0.696
Polarity	CART	0.734	0.712	0.784	0.747	0.677	0.639	**0.816**	**0.717**
	J48	0.676	0.684	0.655	0.669	0.673	0.639	0.797	0.709
	Naive Bayes	0.748	0.772	0.705	0.737	0.671	0.688	0.625	0.655
	Logistic	0.748	0.767	0.712	0.739	0.676	0.656	0.742	0.696
	SVM	0.759	0.765	0.748	0.756	0.674	0.637	0.810	0.713
Strength	CART	0.719	0.729	0.698	0.713	0.661	0.653	0.686	0.669
	J48	0.701	0.697	0.712	0.705	0.646	0.628	0.716	0.669
	Naive Bayes	0.766	**0.830**	0.669	0.741	0.636	0.711	0.460	0.558
	Logistic	0.763	0.797	0.705	0.748	0.662	0.688	0.593	0.637
	SVM	**0.777**	0.824	0.705	0.760	0.694	0.683	0.725	0.703
Emotion	CART	0.579	0.634	0.374	0.471	0.586	0.638	0.398	0.490
	J48	0.590	0.647	0.396	0.491	0.575	0.628	0.370	0.465
	Naive Bayes	0.579	0.628	0.388	0.480	0.573	0.647	0.320	0.428
	Logistic	0.583	0.624	0.417	0.500	0.585	0.635	0.402	0.492
	SVM	0.597	0.622	0.496	0.552	0.594	0.627	0.462	0.532

Table 9: 10-fold Cross-Validation Subjectivity Classification Performances

recall measures, as was also done in [11]. This is because accuracy and F_1 measures are affected by both false positive and false negative results.

Table 9 shows the results for the subjectivity classification task. We can observe that **Baseline.2** outperforms **Baseline.1** in both datasets. This is because Sentiment140 is not focused on subjectivity classification.

There are significant performance differences between both datasets. We hypothesize that STS's tweets have good properties for classification because they show clear differences between neutral and non neutral tweets. On the other hand, in the Sanders dataset, we found tweets marked as neutral that contain mixed positive and negative opinions. Two examples of this kind of tweets are presented below.

1. *Hey @Apple, pretty much all your products are amazing. You blow minds every time you launch a new gizmo. That said, your hold music is crap.*
2. *#windows sucks... I want #imac so bad!!! why is it so damn expensive :(@apple please give me free imac and I will love you :D*

Both tweets are about the company **Apple**. The first tweet shows a positive opinion about Apple's products and at the same time shows a negative opinion about Apple's hold music. This example contains contrary opinions about two different aspects of the entity Apple. The second example is even more complicated because it expresses opinions on two different entities: **Windows** and **Apple**. The tweet compares two products and shows a clear preference for Apple's product **iMac**. Additionally, the message indicates that the product **iMac** is too expensive, something that could be interpreted as a negative opinion about the product. By inspection, that kind of tweets are not included in STS. Due to this fact, we believe that in addition of being larger, Sanders captures in a better way than the STS corpus the sentiment diversity of tweets. Nevertheless, considering that tweets with mixed positive and negative indicators are subjective, we believe that labeling them as **neutral** may increase the level of noise in the data.

Regarding learning algorithms, SVM tends to outperform other methods in accuracy and F_1, and most of the best results are achieved using the best feature selection algorithm. As was expected, the emotion feature subset achieves poor classification results for this task.

Polarity performance results are showed in Table 10. In this case, both baselines are strongly competitive, being the SentiStrength-based baseline better than the other one. This result agrees with the results reported by Nielsen [16] where it was shown that the AFINN lexicon was not able to outperform SentiStrength. We can observe also that the detection of polarity is a more difficult task in Sanders than in STS, as was also observed for the subjectivity detection task.

The best tree obtained for polarity classification by the CART algorithm using all the features on the Sanders dataset is shown in Figure 3. From the figure with can see that top level nodes of the tree correspond to features related to SentiStrength, Sentiment140 and AFINN. This results correspond with the information gain values obtained and explains in some manner why these methods are competitive as baselines. The tree also indicates that negative words from the different lexical resources are more useful than the positive ones.

In a similar way as in the subjectivity task, SVM achieves the best results in accuracy and F_1. This fact suggests that there are non-linearities between the features that are successfully tackled by using the RBF kernel. The performance tends also in both datasets to be better for the polarity task than for the subjectivity problem. This is because most of the lexical resources and methods are more

Dataset		STS				Sanders			
Features	Methods	accuracy	precision	recall	F_1	accuracy	precision	recall	F_1
Baseline.1	SentiStrength	0.777	0.766	0.797	0.781	0.733	0.735	0.729	0.732
Baseline.2	AFINN	0.771	0.804	0.718	0.758	0.713	0.747	0.643	0.691
All	CART	0.788	0.790	0.785	0.788	0.780	0.759	0.821	0.789
	J48	0.788	0.768	0.825	0.796	0.775	0.769	0.786	0.777
	Naive Bayes	0.794	0.757	0.864	0.807	0.774	0.729	0.873	0.794
	Logistic	0.805	0.784	0.842	0.812	**0.801**	0.782	0.834	0.807
	SVM	0.808	**0.808**	0.808	0.808	**0.801**	0.775	0.848	**0.810**
Best.First	CART	0.791	0.775	0.819	0.797	0.789	**0.790**	0.788	0.789
	J48	0.802	0.789	0.825	0.807	0.781	0.778	0.788	0.783
	Naive Bayes	0.811	0.775	**0.876**	0.822	0.788	0.750	0.863	0.802
	Logistic	0.814	0.803	0.831	0.817	0.778	0.765	0.802	0.783
	SVM	**0.816**	0.795	0.853	**0.823**	0.792	0.760	0.854	0.804
Polarity	CART	0.802	0.796	0.814	0.804	0.779	0.736	0.870	0.797
	J48	0.791	0.764	0.842	0.801	0.775	0.728	0.877	0.796
	Naive Bayes	0.805	0.787	0.836	0.811	0.756	0.736	0.800	0.766
	Logistic	0.799	0.779	0.836	0.807	0.786	0.771	0.813	0.791
	SVM	0.799	0.770	0.853	0.810	0.776	0.728	0.882	0.797
Strength	CART	0.780	0.783	0.774	0.778	0.705	0.686	0.757	0.720
	J48	0.777	0.772	0.785	0.779	0.746	0.732	0.775	0.753
	Naive Bayes	0.780	0.746	0.847	0.794	0.762	0.711	0.880	0.787
	Logistic	0.797	0.800	0.791	0.795	0.752	0.747	0.761	0.754
	SVM	0.799	0.805	0.791	0.798	0.779	0.747	0.845	0.793
Emotion	CART	0.684	0.637	0.853	0.729	0.658	0.630	0.766	0.691
	J48	0.681	0.629	0.881	0.734	0.650	0.620	0.777	0.689
	Naive Bayes	0.641	0.599	0.853	0.704	0.654	0.604	**0.891**	0.720
	Logistic	0.661	0.623	0.814	0.706	0.671	0.637	0.795	0.707
	SVM	0.624	0.598	0.757	0.668	0.656	0.624	0.784	0.695

Table 10: 10-fold Cross-Validation Polarity Classification Performances

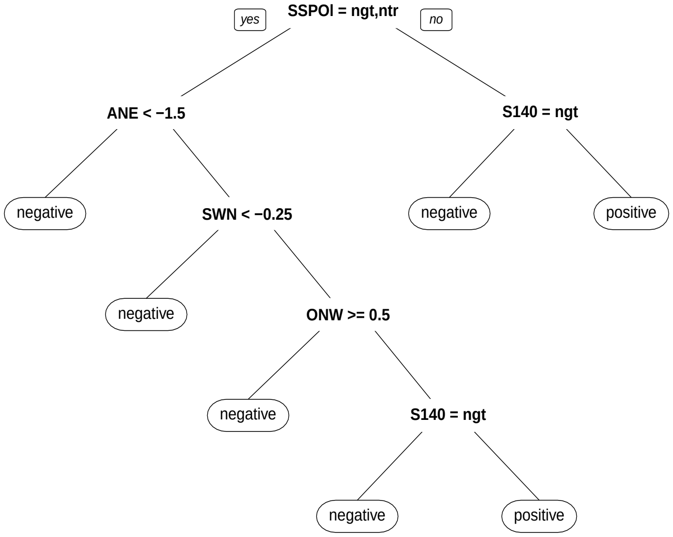

Figure 3: Best Tree trained with CART for polarity classification on the Sanders dataset

focused on the detection of polarity rather than detecting subjectivity.

As was discussed before, emotion-oriented features tend to have low information gain values and also present a poor classification performance. Therefore, it would make sense to think that emotion-oriented features are not useful for sentiment classification. However, if we consider the accuracies obtained by RBF SVMs on the Sanders dataset for both classification tasks, we can see that in addition of outperforming the others learning algorithms, they achieved the best accuracies when all type of features were included. That means, that emotion-oriented features are useful for sentiment classification when they are combined with polarity and strength oriented features in a non-linear fashion.

The best learned functions obtained for each classification task outperformed the results achieved by the baselines created from isolated methods. Thus, our results validate the hypothesis that the combinations of different sentiment analysis methods and resources enhances the overall sentiment classification.

5. CONCLUSIONS AND FUTURE WORK

We present a novel approach for sentiment classification on microblogging messages or short texts based on the combination of several existing lexical resources and sentiment analysis methods. Our experimental validation shows that our classifiers achieve very significant improvements over any singe method, outperforming state-of-the-art methods by more than 5% accuracy and F_1 points.

Considering that the proposed feature representation does not depend directly on the vocabulary size of the collection, it provides a considerable dimensionality reduction in comparison to word-based representations such as unigrams or n-grams. Likewise, our approach also avoids the sparsity problem presented by word-based feature representations for Twitter sentiment classification discussed in [15]. Due to this, our low-dimensional feature representation allows us to efficiently use several learning algorithms.

The classification results varied significantly from one dataset to another. The manual sentiment classification of tweets is a subjective task that can be biased by the evaluator's perceptions. This fact should serve as a warning call against bold conclusions from inadequate evidence in sentiment classification. It is very important to check beforehand whether the labels in the training dataset correspond to the desired values, and if the training examples are able to capture the sentiment diversity of the target domain.

Finally, it is important to recall that opinions are multidimensional objects. In this way, when we classify tweets into polarity classes, we are essentially projecting these multiple dimensions to one single categorical dimension. Furthermore, it is not clear how to project tweets having mixed positive and negative expressions to a single polarity class. Therefore, we have to be aware that the sentiment classification of tweets may lead to the loss of valuable sentiment information.

As future work we expect to expand this study by including other sentiment resources and methods which were not considered at this moment. For instance we expect to create semantic-level features from concept-based resources such as *SenticNet*. Additionally, we plan to evaluate our approach on the *SemEval* task datasets in order to compare our results with other works that participated in the task.

6. ACKNOWLEDGMENT

This work has been partially supported by FONDEF-CONICYT project D09I1185. Felipe Bravo-Marquez was supported by CONICYT's Master Scholarship. Marcelo Mendoza was supported by project FONDECYT 11121435. Barbara Poblete was partially supported by FONDECYT grant 11121511 and Program U-INICIA VID 2012, grant U-INICIA 3/0612; University of Chile.

7. REFERENCES

[1] Baccianella, S., Esuli, A., and Sebastiani, F. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 2010. Valletta, Malta.

[2] Baeza-Yates, R., and Rello, L. How Bad Do You Spell?: The Lexical Quality of Social Media. The Future of the Social Web, Papers from the 2011 ICWSM Workshop, Barcelona, Catalonia, Spain. AAAI Workshops, 2011.

[3] Bradley, M. M., and Lang, P. J. Affective Norms for English Words (ANEW) Instruction Manual and Affective Ratings. *Technical Report C-1, The Center for Research in Psychophysiology* University of Florida, 2009.

[4] Cambria, E., Speer R., Havasi C., and Hussain A. SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. In *FLAIRS Conference*, pages 202–207, 2012.

[5] Carvalho, P., Sarmento, L., Silva, M. J., and de Oliveira, E. Clues for detecting irony in user-generated contents: oh...!! it's "so easy" ;-). In *Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion* Hong Kong, China, 2009.

[6] Esuli, A., and Sebastiani, F. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation, 2006*.

[7] Go, A., Bhayani, R., and Huang, L. Twitter sentiment classification using distant supervision. *Technical report Stanford University*, 2010.

[8] Hall, M. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, University of Waikato, 1999.

[9] Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 151–160, 2011.

[10] Kouloumpis, E., Wilson, T., and Moore, J. Twitter Sentiment Analysis: The Good the Bad and the OMG! In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[11] Liu, K., Li, W., and Guo, M. Emoticon smoothed language models for Twitter sentiment analysis. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*. Toronto, Ontario, Canada, 2012.

[12] Liu, B. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies series, Morgan & Claypool Publishers, 2012.

[13] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244, 1990.

[14] Mohammad, S. M., and Turney, P. D. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 2012.

[15] Saif, H., He, Y., and Alani, H. Alleviating data sparsity for twitter sentiment analysis. In *Workshop of Making Sense of Microposts co-located with WWW 2012*, 2012.

[16] Nielsen, F. Å. A new anew: Evaluation of a word list for sentiment analysis in microblogs. In *ESWC2011 Workshop on Making Sense of Microposts*, May 2011.

[17] Grassi, M., Cambria, E., Hussain, A., and Piazza, F. Sentic web: A new paradigm for managing social media affective information. *Cognitive Computation*, 3(3):480–489, 2011.

[18] Olsher, D J. Full spectrum opinion mining: Integrating domain, syntactic and lexical knowledge. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pages 693–700. IEEE, 2012.

[19] Pak, A., and Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta, 2010.

[20] Pang, B., and Lee, L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135. 2008.

[21] Plutchik, R. 2002. Nature of emotions, American Scientist, 89, 349.

[22] Read, J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. Michigan, USA, 2005.

[23] Thelwall, M., Buckley, K., and Paltoglou, G. Sentiment strength detection for the social web. *JASIST* 63(1):163–173. 2012.

[24] Wilson, T., Kozareva, Z., Nakov, P., Ritter A., Rosenthal, S., and Stoyonov V. Semeval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation*. Association for Computation Linguistics, 2013.

[25] Wilson, T., Wiebe, J., and Hoffmann, P. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*. Vancouver, British Columbia, Canada, 2005.

[26] Zirn, C., Niepert M., Stuckenschmidt H., and Strube M. Fine-grained sentiment analysis with structural features. In *IJCNLP*, pages 336–344, 2011.