Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

Positive, Negative, or Neutral: Learning an Expanded Opinion Lexicon from Emoticon-Annotated Tweets

Felipe Bravo-Marquez, Eibe Frank and Bernhard Pfahringer
Department of Computer Science, University of Waikato
fjb11@students.waikato.ac.nz, {*eibe,bernhard*}*@waikato.ac.nz*

Abstract

We present a supervised framework for expanding an opinion lexicon for tweets. The lexicon contains part-of-speech (POS) disambiguated entries with a three-dimensional probability distribution for positive, negative, and neutral polarities. To obtain this distribution using machine learning, we propose word-level attributes based on POS tags and information calculated from streams of emoticon-annotated tweets. Our experimental results show that our method outperforms the three-dimensional word-level polarity classification performance obtained by semantic orientation, a state-of-the-art measure for establishing world-level sentiment.

1 Introduction

The language used in **Twitter**[1] provides substantial challenges for sentiment analysis. The words used in this platform include many abbreviations, acronyms, and misspelled words that are not observed in traditional media or covered by popular lexicons. The diversity and sparseness of these informal words make the manual creation of a Twitter-oriented opinion lexicon a time-consuming task.

In this article we propose a supervised framework for opinion lexicon expansion for the language used in Twitter. Taking SentiWordnet as inspiration, each word in our expanded lexicon has a probability distribution, indicating how positive, negative, and neutral it is. The estimated probabilities can be used to represent intensities for a specific sentiment category e.g., the word *awesome* is more positive than the word *adequate*. Furthermore, the neutral dimension may be useful for discarding non-opinion words in text-level polarity classification tasks. In contrast, unsupervised lexicon expansion techniques such as semantic orientation [Turney, 2002] provide a single numerical score for each word, and it is unclear how to impose thresholds on this score for neutrality detection.

All the entries in our lexicon are associated with a corresponding part-of-speech tag. By relying on POS-tagged words, homographs[2] with different POS-tags will be disambiguated [Wilks and Stevenson, 1998]. For instance, the word *apple* will receive different sentiment scores when used to refer to a common noun (a fruit) or a proper noun (a company).

The proposed methodology takes a stream of tweets noisily labelled according to their polarity using *emoticon-based annotation*. In this approach, only tweets with positive or negative emoticons are considered and labelled according to the polarity indicated by the emoticon. This idea has been widely used before to train message-level sentiment classifiers [Bifet and Frank, 2010], [Go *et al.*, 2009].

We calculate two types of word-level attributes from the stream of annotated tweets to train a word classifier: Semantic Orientation (SO) [Turney, 2002], which is based on the mutual information between word and sentiment class, and Stochastic Gradient Descent (SGD) score, which learns a linear relationship between word and sentiment class. Additionally, we consider syntactic information of the word in its context by including the POS tag of the word as a nominal attribute.

To train a word-level sentiment classifier using supervised learning, we also need sentiment labels for the words. These labels are provided by a seed lexicon taken from the union of four different hand-made lexicons after discarding all polarity clashes from the intersection.

To the best of our knowledge, this is the first article in which the lexicon expansion of Twitter opinion words using POS disambiguation and supervised learning is studied and evaluated. Additionally, this is the first study in which scores for positive, negative, and neutral sentiment are provided for Twitter-specific expressions.

We test our approach on two collections of automatically labelled tweets. The results indicate that our supervised framework outperforms semantic orientation when the detection of neutral words is considered. We also evaluate the usefulness of the expanded lexicon for message-level polarity classification of tweets, showing significant improvements in performance.

This article is organised as follows. In Section 2 we provide a review of existing work in opinion lexicon expansion. In Section 3 we describe the seed lexicon used to label the words for the training set. In Section 4 we explain the mechanisms studied to automatically create collections of labelled tweets. The creation of our word-level time-series is described in Section 5, together with the features used for training the classifier. In Section 6 we present the experiments

[1]http://www.twitter.com

[2]Words that share the same spelling but have different meanings.

we conducted to evaluate the proposed approach and discuss results. The main findings and conclusions are discussed in Section 7.

2 Related Work

There are two types of resources that can be exploited for lexicon expansion: thesauri and document collections. The simplest approach using a thesaurus such as WordNet[3] is to expand a seed lexicon of labelled opinion words using synonyms and antonyms from the lexical relations provided by the thesaurus [Hu and Liu, 2004], [Kim and Hovy, 2004]. The hypothesis behind this approach is that synonyms have the same polarity and antonyms have the opposite. This process is normally iterated several times. In [Esuli and Sebastiani, 2005], a supervised classifier was trained using a seed of labelled words that was obtained through synonyms and antonyms expansion. For each word, a vector space model is created from the definition or *gloss* provided by the WordNet dictionary. This representation is used to train a word-level classifier that is used for lexicon expansion. An equivalent approach was applied later to create SentiWordnet[4] in [Esuli and Sebastiani, 2006], [Baccianella *et al.*, 2010]. In SentiWordNet, each WordNet *synset* or group of synonyms is assigned into classes positive, negative and neutral in the range $[0, 1]$.

A limitation of thesaurus-based approaches is their inability to capture domain-dependent words. Corpus-based approaches exploit syntactic or co-occurrence patterns to expand the lexicon to the words found within a collection of documents. In [Turney, 2002], the expansion is done through a measure referred to as *semantic orientation*, which is based on the the point-wise mutual information (PMI) between two random variables:

$$\mathrm{PMI}(term_1, term_2) = \log_2 \left(\frac{Pr(term_1 \wedge term_2)}{Pr(term_1)Pr(term_2)} \right) \tag{1}$$

The semantic orientation of a word is the difference between the PMI of the word with a positive emotion and a negative emotion. Different ways have been proposed to represent the joint probabilities of words and emotions. In Turney's work [Turney, 2002], they are estimated using the number of hits returned by a search engine in response to a query composed of the target word together with the word "excellent" and another query using the word "poor".

The same idea was used for Twitter lexicon expansion in [Becker *et al.*, 2013], [Mohammad *et al.*, 2013], [Zhou *et al.*, 2014], which all model the joint probabilities from collections of tweets labelled in automatic ways. In [Becker *et al.*, 2013], the tweets are labelled with a trained classifier using thresholds for the different classes to ensure high precision. In [Zhou *et al.*, 2014], they are labelled with emoticons to create domain-specific lexicons. In [Mohammad *et al.*, 2013], they are labelled with emoticons and hashtags associated with emotions to create two different lexicons. These lexicons were tested for tweet-level polarity classification.

[3]http://wordnet.princeton.edu/

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

3 Ground-Truth Word Polarities

In this section, we describe the seed lexicon used to label the training dataset for our word sentiment classifier. We create it by taking the union of the following manually created lexical resources:

MPQA Subjectivity Lexicon: This lexicon was created by Wilson et al. [Wilson *et al.*, 2005] and is part of OpinionFinder[5], a system that automatically detects subjective sentences in document corpora. The lexicon has positive, negative, and neutral words.

Bing Liu: This lexicon is maintained and distributed by Bing Liu[6] and was used in several of his papers [Liu, 2012]. It has positive and negative entries.

Afinn: This strength-oriented lexicon [Nielsen, 2011] has positive words scored from 1 to 5 and negative words scored from -1 to -5. It includes slang, obscene words, acronyms and Web jargon. We tagged words with negative and positive scores to negative and positive classes respectively.

NRC emotion Lexicon: This emotion-oriented lexicon [Mohammad and Turney, 2013] was created by conducting a tagging process on the crowdsourcing Amazon Mechanical Turk platform. In this lexicon, the words are annotated according to eight emotions: joy, trust, sadness, anger, surprise, fear, anticipation, and disgust, and two polarity classes: positive and negative. There are many words that are not associated with any emotional state and are tagged as neutral. In this work, we consider positive, negative, and neutral tags from this lexicon.

As we need to reduce the noise in our training data, we discard all words where a polarity clash is observed. A polarity clash is a word that receives two or more different tags in the union of lexicons. The number of words for the different polarity classes in the different lexicons is displayed in Table 1.

	Positive	Negative	Neutral
AFINN	564	964	0
Bing Liu	2003	4782	0
MPQA	2295	4148	424
NRC-Emo	2312	3324	7714
Seed Lexicon	3730	6368	7088

Table 1: Lexicon Statistics

From the table, we can observe that the number of words per class is significantly reduced after removing the clashes from the union. The total number of clashes is 1074.

This high number of clashes found among different handmade lexicons indicates two things: 1) Different human annotators can disagree when tagging a word to polarity classes. 2) There are several words that can belong to more than one sentiment class. Due to this, we can say that word-level polarity classification is a hard and subjective problem.

[5]http://mpqa.cs.pitt.edu/opinionfinder/opinionfinder_2/

[6]http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

4 Obtaining Labelled Tweets

In order to calculate the attributes for our word-level classifier, we require a collection of time-stamped tweets with their corresponding polarity labels.

We rely on the emoticon-based annotation approach in which tweets exhibiting positive :) and negative :(emoticons are labelled according to the emoticon's polarity [Go *et al.*, 2009]. Afterwards, the emoticon used to label the passage is removed from the content.

We consider two collections of tweets covering multiple topics: The Edinburgh corpus (ED) [Petrović *et al.*, 2010], and the Stanford Sentiment corpus (STS) [Go *et al.*, 2009].

The ED corpus has 97 million tweets which were collected using the Twitter streaming API in a period spanning November 11th 2009 until February 1st 2010. This collection includes tweets in multiple languages. As was done in [Bifet and Frank, 2010], non-English tweets were filtered out, and tweets without emoticons were discarded.

The STS corpus was created by periodically sending queries :) and :(to the Twitter search API between April 6th 2009 to June 25th 2009. All the tweets in this collection are written in English.

	ED	STS
Positive	$1,813,705$	$800,000$
Negative	$324,917$	$800,000$
Total	$2,138,622$	$1,600,000$

Table 2: Collection statistics

The number of tweets for each polarity class in the two corpora is given in Table 2. We can observe that when using the streaming API (ED), positive emoticons occur much more frequently than negative ones.

5 Word-level Features

All the tweets from the annotated collection are lowercased, tokenised and POS-tagged. We use the TweetNLP library [Gimpel *et al.*, 2011], which provides a tokeniser and a tagger specifically for the language used in Twitter. We prepend a POS-tag prefix to each word in order to differentiate homographs exhibiting different POS-tags.

To calculate the proposed features, we treat the time-sorted collection of tweets as a data stream and create two time-series for each POS-tagged word observed in the vocabulary: the SGD series, and the SO series. These time-series intend to capture the evolution of the relationship between a word and the sentiment that it expresses.

The first time-series is calculated by incrementally training a linear support vector machine using stochastic gradient descent (SGD) [Zhang, 2004]. The weights of this linear model correspond to POS-tagged words and are updated in an incremental fashion. We optimise the hinge loss function with an L_2 penalty and a learning rate equal to 0.1:

$$\frac{\lambda}{2}||w||^2 + \sum[1 - y(\mathbf{x}\mathbf{w} + b)]_+. \qquad (2)$$

The variables $\mathbf{w}$, b, and λ correspond to the weight vector, the bias, and the regularisation parameter, respectively. The class labels y are assumed to be in $\{+1, -1\}$, corresponding to positively and negatively labelled tweets, respectively. The regularisation parameter was set to 0.0001. The model's weights determine how strongly the presence of a word influences the prediction of negative and positive classes [Bifet and Frank, 2010]. The SGD time-series is created by applying this learning process to a collection of labelled tweets and storing the word's coefficients in different time windows. We use time windows of $1,000$ examples.

The second time-series corresponds to the accumulated semantic orientation (SO) introduced in Section 2. Let *count* be a function that counts the number of times that a word or a sentiment label has been observed during a certain period of time. We calculate the SO score for each POS-tagged word in an accumulated way according to the following expression:

$$\text{SO}(word) = log_2\left(\frac{\text{count}(\text{word} \wedge \text{pos}) \times \text{count}(\text{neg})}{\text{count}(\text{word} \wedge \text{neg}) \times \text{count}(\text{pos})}\right) \qquad (3)$$

We use time windows of $1,000$ examples and the Laplace correction to avoid the zero-frequency problem.

Feature	Description
mean	The mean of the time-series.
trunc.mean	The truncated mean of the time-series.
median	The median of the time-series.
last.element	The last observation of the time-series.
sd	The standard deviation of the time-series .
iqr	The inter-quartile range.
sg	The fraction of times the time-series changes its sign.
sg.diff	The sg value for the differenced time-series.

Table 3: Time-series features

We rely on our time-series to extract features that are used to train our world-level polarity classifier. These features summarise location and dispersion properties of the time-series, and are listed in Table 3. Location-oriented features *mean*, *trimm.mean* and *median* measure the central tendency of the time-series. The feature *last.element* corresponds to the last value observed in the time-series. This attribute would be equivalent to the traditional semantic orientation measure for the SO time-series. The features *sd*, *iqr*, *sg*, and *sg.diff* measure the level of dispersion of the time-series.

In addition to these time-series features, we include the POS-tag of the word as a nominal attribute. We include this attribute based on the hypothesis that non-neutral words are more likely to exhibit certain POS tags than neutral ones [Zhou *et al.*, 2014].

6 Experiments

In this section, we present our experimental results for Twitter lexicon expansion. In the first part, we study the word-level polarity classification problem. In the second part, we expand the seed lexicon using the trained classifier and use it for message-level polarity classification of tweets.

6.1 Word-level polarity classification

We calculated the time-series described in Section 5 for the most frequent $10,000$ POS-tagged words found in each of our two datasets. The time-series were calculated using MOA[7], a data stream mining framework.

	ED	STS
Positive	1027	1023
Negative	806	985
Neutral	1814	1912
Total	3647	3920

Table 4: Word-level polarity classification datasets.

To create training and test data for machine learning, all the POS-tagged words matching the seed lexicon are labelled according to the lexicon's polarities. It is interesting to consider how frequently positive, negative, and neutral words occur in a collection of tweets. The number of words labelled as positive, negative, and neutral for both the ED and STS datasets is given in Table 4. As shown in the table, neutral words are the most frequent words in both datasets. Moreover, positive words are more frequent than negative ones.

Next, we focus on the word-level classification problem. With the aim of gaining a better understanding of the problem, we study three word-level classification problems:

1. *Neutrality*: Classify words as neutral (objective) or non-neutral (subjective). We label positive and negative words as non-neutral for this task.
2. *PosNeg*: Classify words to positive or negative classes. We remove all neutral words for this task.
3. *Polarity*: Classify words to classes positive, negative or neutral. This is the primary classification problem we aim to solve.

We study the information provided by each feature with respect to the three classification tasks described above. This is done by calculating the information gain of each feature. This score is normally used for decision tree learning and measures the reduction of entropy within each class after performing the best split induced by the feature. The information gain obtained for the different attributes in relation to the three classification tasks is shown in Table 5. The attributes achieving the highest information gain per task are marked in bold.

We can observe that variables measuring the location of the SO and SGD time-series tend to be more informative than the ones measuring dispersion. Moreover, the information gain of these variables is much higher for PosNeg than for neutrality. SGD and SO are competitive measures for neutrality, but SO is better for PosNeg. An interesting insight is that features that measure the central tendency of the time-series tend to be more informative than the last values, especially for SGD. These measures smooth the fluctuations of the SGD time-series. We can see that the feature *sgd.mean* is the best attribute for neutrality classification in both datasets. We can also see that POS tags are useful for neutrality detection, but useless for PosNeg. Therefore, we can conclude that positive and negative words have a similar distribution of POS tags.

We trained supervised classifiers for the three different classification problems in both datasets STS and ED. The classification experiments were performed using WEKA[8], a machine learning environment. We studied the following learning algorithms in preliminary experiments: RBF SVM, Logistic regression, C4.5, and Random Forest. As the RBF SVM produced the best performance among the different methods, we used this method in our classification experiments with a nested grid search procedure for parameter tuning where internal cross-validation is used to find C and γ.

The evaluation was done using 10 times 10-folds-cross-validation and different subsets of attributes were compared. All the methods are compared with the baseline of using the last value of the semantic orientation, based on a corrected resampled paired *t*-student test with an α level of 0.05 [Nadeau and Bengio, 2003]. We used the following subsets of attributes: 1) *SO*: Includes only the feature *SO.last*. This is the baseline and is equivalent to the standard semantic orientation measure with the decision boundaries provided by the SVM. 2) *ALL*: Includes all the features. 3)*SGD.TS+POS*: Includes all the features from the SGD time-series and the POS tag. 4)*SO.TS+POS*: Includes all the features from the SO time-series and the POS tag. 5)*SO+POS*: Includes the features *so.last* and the POS tag.

We evaluate the weighted area under the ROC curves (AUCs) (to deal with class imbalances) for the four different subsets of attributes in the two datasets. The classification results are presented in Table 6. The symbols ∘ and • correspond to statistically significant improvements and degradations with respect to the baseline, respectively.

We can observe a much lower performance in Neutrality detection than in PosNeg. This indicates that the detection of neutral Twitter words is much harder than distinguishing between positive and negative words. The performance on both datasets tends to be similar. However, the results for STS are better than for ED. This suggests that a collection of balanced positively and negatively labelled tweets may be more suitable for lexicon expansion. Another result is that the combination of all features leads to a significant improvement over the baseline for neutrality and polarity classification. In the PosNeg classification task, we can see that the baseline is very strong. This suggests that SO is very good for discriminating between positive and negative words, but not strong enough when neutral words are included. Regarding SO and SGD time-series, we can conclude that they are competitive for neutrality detection. However, SO-based features are better for PosNeg and Polarity tasks.

6.2 Lexicon expansion

The ultimate goal of the polarity-classification of words is to produce a Twitter-oriented opinion lexicon emulating the properties of SentiWordet, i.e., a lexicon of POS-tagged disambiguated entries with their corresponding distribution for positive, negative, and neutral classes. To do this, we fit a lo-

[7]http://moa.cs.waikato.ac.nz/

[8]http://www.cs.waikato.ac.nz/ml/weka/

Dataset	ED			STS		
Task	Neutrality	PosNeg	Polarity	Neutrality	PosNeg	Polarity
pos-tag	0.062	0.017	0.071	0.068	0.016	0.076
sgd.mean	**0.082**	0.233	0.200	**0.104**	0.276	**0.246**
sgd.trunc.mean	0.079	0.237	0.201	**0.104**	0.276	0.242
sgd.median	0.075	0.233	0.193	0.097	0.275	0.239
sgd.last	0.057	0.177	0.155	0.086	0.258	0.221
sgd.sd	0.020	0.038	0.034	0.030	0.030	0.052
sgd.sg	0.029	0.000	0.030	0.049	0.017	0.062
sgd.sg.diff	0.000	0.000	0.008	0.005	0.000	0.000
sgd.iqr	0.018	0.012	0.019	0.015	0.014	0.017
so.mean	0.079	0.283	**0.219**	0.081	**0.301**	0.232
so.trunc.mean	0.077	**0.284**	0.215	0.079	0.300	0.229
so.median	0.077	0.281	0.215	0.076	0.300	0.228
so.last	0.069	0.279	0.211	0.084	0.300	0.240
so.sd	0.000	0.015	0.008	0.000	0.012	0.007
so.sg	0.013	0.216	0.126	0.019	0.239	0.142
so.sg.diff	0.000	0.012	0.009	0.000	0.000	0.000
so.iqr	0.000	0.000	0.000	0.000	0.008	0.000

Table 5: Information gain values.

Dataset	SO	ALL	SGD.TS+POS	SO.TS+POS	SO+POS
ED-Neutrality	0.62 ± 0.02	$\mathbf{0.65} \pm 0.02$ ○	$\mathbf{0.65} \pm 0.02$ ○	$\mathbf{0.65} \pm 0.02$ ○	0.64 ± 0.02 ○
ED-PosNeg	0.74 ± 0.03	$\mathbf{0.75} \pm 0.03$	0.71 ± 0.03 ●	0.74 ± 0.03	0.73 ± 0.03
ED-Polarity	0.62 ± 0.02	$\mathbf{0.65} \pm 0.02$ ○	0.64 ± 0.02	$\mathbf{0.65} \pm 0.02$ ○	0.64 ± 0.02 ○
STS-Neutrality	0.63 ± 0.02	$\mathbf{0.67} \pm 0.02$ ○	0.66 ± 0.02 ○	0.66 ± 0.02 ○	0.66 ± 0.02 ○
STS-PosNeg	$\mathbf{0.77} \pm 0.03$	$\mathbf{0.77} \pm 0.03$	0.75 ± 0.03 ●	$\mathbf{0.77} \pm 0.03$	$\mathbf{0.77} \pm 0.03$
STS-Polarity	0.64 ± 0.02	$\mathbf{0.66} \pm 0.01$ ○	0.65 ± 0.02 ○	$\mathbf{0.66} \pm 0.02$ ○	$\mathbf{0.66} \pm 0.02$ ○

Table 6: World-level classification performance.

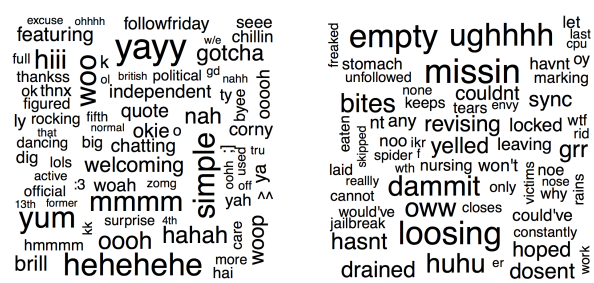
gistic regression model to the output of the support vector machine trained for the *polarity* problem using all the attributes. The resulting model is then used to classify the remaining unlabelled words. This process is performed for both STS and ED datasets.

A sample from the expanded word list is given in Table 7. We can see that each entry has the following attributes: the word, the POS-tag, the sentiment label that corresponds to the class with maximum probability, and the distribution. We inspected the expanded lexicon and observed that the estimated probabilities are intuitively plausible. However, there are some words for which the estimated distribution is questionable, such as the word *same* in Table 7.

word	POS	label	negative	neutral	positive
alrighty	interjection	positive	0.021	0.087	0.892
boooooo	interjection	negative	0.984	0.013	0.003
lmaoo	interjection	positive	0.19	0.338	0.472
french	adjective	neutral	0.357	0.358	0.285
handsome	adjective	positive	0.007	0.026	0.968
saddest	adjective	negative	0.998	0.002	0
same	adjective	negative	0.604	0.195	0.201
anniversary	common.noun	neutral	0.074	0.586	0.339
tear	common.noun	negative	0.833	0.124	0.044
relaxing	verb	positive	0.064	0.244	0.692
wikipedia	proper.noun	neutral	0.102	0.644	0.254

Table 7: Expanded words example.

The provided probabilities can also be used to explore the sentiment intensities of words. In Figure 1, we visualise the expanded lexicon intensities of words classified as positive and negative through word clouds. The sizes of the words are proportional to the log odds ratios $\log_2(\frac{P(pos)}{P(neg)})$ and $\log_2(\frac{P(neg)}{P(pos)})$ for positive and negative words, respectively.



(a) (b)

Figure 1: Word clouds of positive and negative words using log odds proportions.

Next, we study the usefulness of our expanded lexicons based on ED and STS for Twitter polarity classification. This involves categorising entire tweets into positive or negative sentiment classes.

The experiments were performed on three collections of tweets that were manually assigned to positive and negative classes. The first collection is *6HumanCoded*[9], which was used to evaluate SentiStrength [Thelwall *et al.*, 2012]. In this dataset, tweets are scored according to positive and negative numerical scores. We use the difference of these scores to create polarity classes and discard messages where it is equal to zero. The other datasets are *Sanders*[10], and *SemEval*[11]. The number of positive and negative tweets per corpus is given in Table 8.

	Positive	Negative	Total
6Coded	1340	949	2289
Sanders	570	654	1224
SemEval	5232	2067	7299

Table 8: Message-level polarity classification datasets.

We train a logistic regression on the labelled collections of tweets based on simple count-based features extracted using the lexicons. We compute features in the following manner. We count the number of positive and negative words from the seed lexicon matching the content of the tweet. From the expanded lexicons we create a positive and a negative score. The positive score is calculated by adding the positive probabilities of POS-tagged words labelled as positive within the tweet's content. The negative score is calculated in an analogous way from the negative probabilities. Words are discarded as non-opinion words whenever the expanded lexicons labelled them as neutral.

We study three different setups based on these attributes. 1) *Baseline*: It includes the attributes calculated from the seed lexicon. 2) *ED*: It includes the baseline and the attributes from the ED expanded lexicon. 3) *STS*: This one is analogus to ED, but using the STS lexicon.

In the same way as in the word-level classification task, we rely on the weighted AUC as evaluation measure, and we compare the different setups with the baseline using the corrected paired *t*-tests. The classification results obtained for the different setups are shown in Table 9.

Dataset	Baseline	ED	STS
6-coded	0.77 ± 0.03	$\mathbf{0.82} \pm 0.03 \circ$	$\mathbf{0.82} \pm 0.02 \circ$
Sanders	0.77 ± 0.04	$0.83 \pm 0.04 \circ$	$\mathbf{0.84} \pm 0.04 \circ$
SemEval	0.77 ± 0.02	$0.81 \pm 0.02 \circ$	$\mathbf{0.83} \pm 0.02 \circ$

Table 9: Message-level polarity classification performance.

The results indicate that the expanded lexicons produce meaningful improvements in performance over the baseline on the different datasets. The performance of STS is slightly better than that of ED. This pattern was also observed in the word-level classification performance shown in Table 6. This suggests that the two different ways of evaluating the lexicon expansion, one at the word level and the other at the message level, are consistent with each other.

7 Conclusions

In this article, we presented a supervised method for opinion lexicon expansion in the context of tweets. The method creates a lexicon with disambiguated POS entries and a probability distribution for positive, negative, and neutral classes.

The experimental results show that the supervised fusion of POS tags, SGD weights, and semantic orientation, produces a significant improvement for three-dimensional word-level polarity classification compared to using semantic orientation alone. We can also conclude that, as attributes describing the central location of SGD and SO time-series smooth the temporal fluctuations in the sentiment pattern of a word, they tend to be more informative than the last values of the series for word-level polarity classification.

To the best of our knowledge, our method is the first approach for creating Twitter opinion lexicons with these characteristics. Considering that these characteristics are very similar to those of SentiWordNet, a well-known publicly available lexical resource, we believe that several sentiment analysis methods that are based on SentiWordnet can be easily adapted to Twitter by relying on our lexicon[12].

Our supervised framework for lexicon expansion opens several directions for further research. The method could be used to create domain-specific lexicons by relying on tweets collected from the target domain. However, there are many domains such as politics, in which emoticons are not frequently used to express positive and negative opinions. New ways for automatically labelling collections of tweets should be explored. We could rely on other domain-specific expressions such as hashtags, or use message-level classifiers trained from domains in which emoticons are frequently used.

Other types of word-level features based on the context of the word can be explored. We could rely on well-known opinion properties such as negations, opinion shifters, and intensifiers, to create these features.

As our word-level features are based on time-series, they could be easily calculated in an on-line fashion from a stream of time-evolving tweets. Based on this, we could study the dynamics of opinion words. New opinion words could be discovered because the change of the distribution in certain words could be tracked. This approach could be used for online lexicon expansion in specific domains, and potentially be useful for high impact events on Twitter, such as elections and sports competitions.

References

[Baccianella *et al.*, 2010] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced

[9] http://sentistrength.wlv.ac.uk/documentation/6humanCodedDataSets.zip

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

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

[12] The expanded lexicons and the source code used to generate them are available for download at http://www.cs.waikato.ac.nz/ml/sa/lex.html#ijcai15.

lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, LREC'10, pages 2200–2204, Valletta, Malta, 2010.

[Becker *et al.*, 2013] Lee Becker, George Erhart, David Skiba, and Valentine Matula. Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises*, SemEval'13, pages 333–340, 2013.

[Bifet and Frank, 2010] Albert Bifet and Eibe Frank. Sentiment knowledge discovery in twitter streaming data. In *Proceedings of the 13th international conference on Discovery science*, DS'10, pages 1–15, Berlin, Heidelberg, 2010. Springer-Verlag.

[Esuli and Sebastiani, 2005] Andrea Esuli and Fabrizio Sebastiani. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 617–624, New York, NY, USA, 2005. ACM.

[Esuli and Sebastiani, 2006] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation*, LREC'06, pages 417–422, 2006.

[Gimpel *et al.*, 2011] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics, 2011.

[Go *et al.*, 2009] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 2009.

[Hu and Liu, 2004] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.

[Kim and Hovy, 2004] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, pages 1367–1373, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[Liu, 2012] Bing Liu. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[Mohammad and Turney, 2013] Saif Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[Mohammad *et al.*, 2013] Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises*, SemEval'13, pages 321–327, 2013.

[Nadeau and Bengio, 2003] Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52(3):239–281, 2003.

[Nielsen, 2011] Finn Å. Nielsen. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, #MSM2011, pages 93–98, 2011.

[Petrović *et al.*, 2010] Saša Petrović, Miles Osborne, and Victor Lavrenko. The edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, WSA '10, pages 25–26, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Thelwall *et al.*, 2012] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *JASIST*, 63(1):163–173, 2012.

[Turney, 2002] Peter D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[Wilks and Stevenson, 1998] Yorick Wilks and Mark Stevenson. The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Natural Language Engineering*, 4(02):135–143, 1998.

[Wilson *et al.*, 2005] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[Zhang, 2004] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 919–926, New York, NY, USA, 2004. ACM.

[Zhou *et al.*, 2014] Zhixin Zhou, Xiuzhen Zhang, and Mark Sanderson. Sentiment analysis on twitter through topic-based lexicon expansion. In Hua Wang and MohamedA. Sharaf, editors, *Databases Theory and Applications*, volume 8506 of *Lecture Notes in Computer Science*, pages 98–109. Springer International Publishing, 2014.