Feature Selection for Twitter Sentiment Analysis: An Experimental Study

Riham Mansour[1], Mohamed Farouk Abdel Hady[2], Eman Hosam[1], Hani Amr[1], and Ahmed Ashour[1]

[1] Microsoft Research, Advanced Technology Lab, Cairo, Egypt
{rihamma,t-emadel,v-haniam,aash}@microsoft.com
[2] Microsoft, Redmond, WA, USA
Mohamed.Abdel-Hady@microsoft.com

Abstract. Feature selection is an important problem for any pattern classification task. In this paper, we developed an ensemble of two Maximum Entropy classifiers for Twitter sentiment analysis: one for subjectivity and the other for polarity classification. Our ensemble employs surface-form, semantic and sentiment features. The classification complexity of this ensemble of linear models is linear with respect to the number of features. Our goal is to select a compact feature subset from the exhaustive list of extracted features in order to reduce the computational complexity without scarifying the classification accuracy. We evaluate the performance on two benchmark datasets, CrowdScale and SemEval. Our selected 20K features have shown very similar results in subjectivity classification to the NRC state-of-the-art system with 4 million features that has ranked first in 2013 SemEval competition. Also, our selected features have shown a relative performance gain in the ensemble classification over the baseline of uni-gram and bi-gram features of 9.9% on CrowdScale and 11.9% on SemEval.

1 Introduction

Twitter is a popular micro-blogging service where users post status messages (called tweets). The users use tweets to share their personal feelings and some- times express opinions in the form of user-defined hashtags, emoticons or normal words about different topics such as events, movies, products or celebrities. Sentiment Analysis (SA) can be formulated as a text classification task where the categories are polarities such as positive and negative. There has been a large amount of NLP research on this user-generated content in the area of sentiment classification. Traditionally most of the research work has focused on large pieces of text, such as product and movie reviews that represent summarized thoughts of authors. Although tweets became publicly available for the research community, they are different from reviews primarily because they are limited to 140 characters and have a more colloquial linguistic style. The frequency of misspellings and slang in tweets is much higher than in reviews.

The sentiment classification task can be handled either by a lexicon-based approach that requires an extensive set of manually supplied sentiment-bearing words or a supervised machine learning approach that requires a large amount of hand-labeled training tweets. The sentiment analysis research has shown that a two-stage approach is more effective [4]: the first stage is subjectivity classification in which subjective instances are

A. Gelbukh (Ed.): CICLing 2015, Part II, LNCS 9042, pp. 92–103, 2015.
DOI: 10.1007/978-3-319-18117-2_7

distinguished from objective ones, then whether the subjective instances has ”positive” or ”negative” polarity is detected.

In this paper, we aim to investigate all types of features introduced in the literature for sentiment analysis. Then evaluate their discrimination ability on a number of benchmark datasets. By the end of this study, we find out the compact feature subset of the exhaustive list of that features that can maintain the classification accuracy while reduce the computational complexity as it is linear with respect of feature set size.

The remainder of the paper is organized as follows. In Section 2, related work on machine learning based on sentiment classification and feature selection for sentiment analysis are reviewed. Section 3 defines the feature types investigated in this paper. In Section 4, presents the performance evaluation of our approach for feature selection. Finally, some conclusive remarks are given in Section 5.

2 Related Work

Machine learning is used extensively to automatically extract sentiment from text. While traditional work [17] focused on movie reviews, more recent research has explored social networks for sentiment analysis. The methods involved differ somewhat since texts like tweets have a different purpose and a more colloquial linguistic style [9]. Go et al. [8] have trained a sentiment classifier to label tweets' sentiment polarities as “positive” or “negative”. Pak et al. [16] trained classifiers to also detect “neutral” tweets that do not contain sentiment. Sentiment classifier training requires a large amount of labeled training data, but the manual annotation of tweets is expensive and time-consuming. To collect the training data, Go, Pak and others used a heuristic method introduced by Read [19] to assign sentiment labels to tweets based on emoticons.

The n-grams representation and specifically Bag-of-Words are commonly used for sentiment classification, resulting in high-dimensional feature space. Agarwal and Mittal [2] has extracted uni-grams and bi-grams from product and movie review text then they have used Information Gain (IG) and Minimum Redundancy maximum Relevancy [18] feature selection criteria to select prominent features.

Barbosa and Feng [4] followed the two-stage approach but instead of n-gram features they have used polarity lexicons, part-of-speech tags, lexical and special micro-blogging features such as emoticons, hashtags, punctuation and character repetitions and words in capital letters to build SVM classifiers. Because the language used on Twitter is often informal and differs from traditional text types [9], most approaches include a preprocessing step. Usually emoticons are detected, URLs removed, abbreviations expanded and twitter markup is removed or replaced by markers. Zhang et al. [23] combined a lexicon-based classifier with an SVM to increase the recall of the classification.

3 Feature Selection

Our goal is to select a compact feature subset from the exhaustive list of extracted features in order to reduce the computational complexity without scarifying the classification accuracy. An ensemble of two binary classifiers is composed of two members: a subjectivity classifier indicating whether the tweet carries sentiment or not and the other

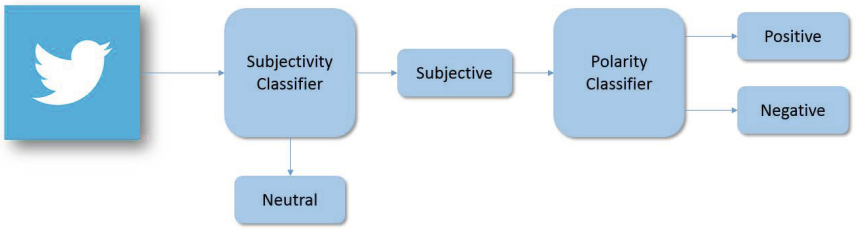

Fig. 1. Ensemble of Subjectivity and Polarity Classifiers

is a polarity classifier indicating whether the subjective tweet is positive or negative. Figure 1 depicts the ensemble of classifiers. We trained Maximum Entropy classifiers on the Azure Machine Learning (AzureML) platform for both subjectivity and polarity on two training data sets.

Because the language used on Twitter is often informal and differs from traditional text types [9]. We pre-process the tweets by removing stop words, using the Natural Language Toolkit (NLTK) library[1] and non-alphabetic characters. We detect emoticons using a regular expression adopted from Christopher Potts' tokenizing script [2]. We normalized URLs to http://someurl and userids to @someuser. We tokenized and part-of-speech tagged the tweets with the Carnegie Mellon University (CMU) tool [7]. All negations (e.g. not, no, never, n't, cannot) are replaced by "'NOT"'. Repeated character sequences in words like "cooooool" are replaced with three characters, so it becomes "coool". In this paper, we adopt multiple features from the literature[13,1,21] in an attempt to find the best combinations for sentiment subjectivity and polarity classification. We use surface-form, semantic and sentiment features. The following subsections describe the three feature sets along with the baseline features.

3.1 Baseline Features

The baseline features of both our subjectivity and polarity classifiers are the unigrams and bigrams of the tweets. We apply feature reduction using Log Likelihood Ratio (LLR) to select the top 20K features that highly co-relate with the training data. All feature types are combined into a single feature vector. Pang et al. [17] have shown that feature presence (binary value) is more useful than feature frequency. Therefore, we use binary feature presence instead of feature frequency.

3.2 Senti-Features

These features refer to the subset of features we adopted from the 100 senti-features presented by Agarwal et al. [1]. Additional pre-processing steps have been performed

[1] `http://nltk.org/`
[2] `http://sentiment.christopherpotts.net/tokenizing.html`

on the tweets for the extraction of the senti-features as follows: All emoticons are replaced with their sentiment polarity specified in the sentiment lexicon presented in [1]. Social acronyms are expanded using an online resource [3] that has 5,184 expansions.

A number of the senti-features are based on prior polarity of words and emoticons. We adopted the same emoticon dictionary in [1] that is composed of 170 manually-annotated emoticons listed on Wikipedia [4] with their emotional state. We adopted the subjectivity lexicon [5] from Opinion Finder [22], an English subjectivity analysis system which classifies sentences as subjective or objective. The lexicon was compiled from manually developed resources augmented with entries learned from corpora. It contains 6,856 entries including 99 multi-word expressions. The entries in the lexicon have been labeled (1) for polarity either positive, negative, or neutral, (2) for part of speech, and (3) for reliability those that appear most often in subjective contexts are strong cludes of subjectivity, while those appearing less are often labeled weak. In order to increase the coverage of the lexicon, we adopt an approach similar to [3]. We used synonym relations from English WordNet[6]to expand the initial seed MPQA English polarity lexicon. The assumption is that synonyms carry same sentiment/polarity as compared to the root words. We make a hypothesis of traversing WordNet like a graph where words are connected to each other based on synonym or antonym relations. Consider each word in this list as a node of the graph. Each node has many in-links and many out-links. This is an undirected graph which is not fully connected i.e. not all the nodes are connected to every other node. For every word in the seed lexicon, we identify its synonym and append with appropriate polarity label in the seed lexicon. Unlike [3], we performed one iteration of traversal, we did not identify the synonyms of the seed lexicon words synonyms. As a post-processing step, we exclude any term that appears more than once with different polarity.

Unlike Agarwal et al. in [1], the features are calculated for the whole tweet only instead of calculating them for the last one-third as well. The senti-features are either counts of other features, lexicon-based features, and Boolean features including bag of words, presence of exclamation marks and capitalized text. We only re-implemented the counts and lexicon-based features as our unigram/bigram baseline covers the rest of the senti-features. Features whose prior polarity are calculated from the emoticon dictionary or the lexicon are classified as polar while all other features are non-polar. The following list depicts the polar and non-polar features we adopted.

Polar features:

- # of (+/-) POS (JJ, RB, VB, NN)
- # of negation words, positive words, negative words
- # of positive and negative emoticons
- # of (+/-) hashtags and capitalized words
- For POS JJ, RB, VB, NN, sum of the prior polarity scores of words of that POS
- Sum of prior polarity scores of all words

[3] http://www.noslang.com/
[4] http://www.wikipedia.org/wiki/List_of_emoticons
[5] mpqa.cs.pitt.edu/lexicons/subj_lexicon/
[6] http://wordnet.princeton.edu/

Non-polar features:

- # of JJ, RB, VB, NN
- # of slangs, latin alphabets, dictionary words, words
- # of hashtags, URLs, targets
- Percentage of capitalized text
- Exclamation, capitalized text

3.3 Sentiment-Specific Word Embedding Features (SSWE)

Sentiment-specific word embedding features (SSWE) are introduced in [21] where a model for learning SSWE is introduced. Unlike continuous word representations that typically models the syntactic context of words only, SSWE encodes sentiment information in the continuous word representation. SSWE addresses the problem of mapping words with similar syntactic context but opposite sentiment polarity such as "'good"' and "'bad"' to neighboring word vectors. The SSWEs are obtained in [21] from large-scale training corpora of distant-supervised tweets collected by positive and negative emoticons. The training data are passed to three neural networks whose loss functions incorporate the supervision from sentiment polarity of text.

The work in [21] extends the existing word embedding learning algorithm in [6] where the C&W model is introduced to learn word embeddings based on the syntactic contexts of words. Given an ngram, C&W replaces the center word with a random word to derive a corrupted ngram. The training objective is that the original ngram is expected to obtain a higher language model score than the corrupted ngram by a margin of 1. Following the C&W model, SSWE incorporate the sentiment information into the neural network through predicting the sentiment distribution of text based on input ngram. A sliding window on the ngram input is used to predict the sentiment polarity based on each ngram with a shared neural network. The higher layers of the neural network are interpreted as features describing the input. Instead of hand-crafted features for Twitter sentiment classification under a supervised framework like in [17], SSWE framework incorporates the continuous representations of words and phrases as the features of a tweet. The sentiment classifier is built from the tweets with manually annotated sentiment polarity. We have obtained the SSWEs from the authors of [21]. They trained their model on 10M tweets labeled automatically using emoticons in [11].

3.4 NRC Features

Mohammad et al. [13] presented the top-performed system in SemEval 2013 Twitter sentiment classification track. The work incorporates diverse sentiment lexicons and many hand-crafted features. We re-implemented this system as the codes are not available publicly. [13] presented two lexicons namely the hashtag sentiment lexicon and the sentiment140 lexicon. The former contains 308,808 entries of terms and their associated sentiment score calculated from tweets selected with seed positive and negative hashtags. The sentiment140 lexicon has been generated in the same way as the hashtag sentiment lexicon, but from the sentiment140 corpus [8] that is labeled automatically

annotated based on emoticons. The sentiment140 lexicon has entries for 62,468 unigrams, 677,698 bigrams, and 480,010 non-contiguous pairs. The authors also use three other lexicons in their features from that were previously presented in [10,14] along with MPQA. Mohammad et al.[13] proposed the following features:

- Word ngrams for the presence or absence of contiguous sequences of 1, 2, 3, and 4 tokens. We have not used this feature as it is very similar to our unigram and bigram features. 3 and 4 grams are costly to calculate especially when training on large text corpus.
- Character ngrams for the presence or absence of contiguous sequences of 3, 4 and 5 characters. We have not used these features as we used unigram and bigram features as a replacement.
- all-caps for the number of words with all characters in upper case.
- POS capturing the number of occurences of each part-of-speech tag.
- hashtags for the number of hashtags.
- lexicon features are based on the 5 lexicons mentioned above. The lexicon features are created for all tokens in the tweet, each POS, hashtags, and all-caps tokens. For each token w, the polarity score p in the lexicon is used to determine
  - total count of tokens in the tweet with score(w,p)¿0.
  - total score adding all the score(w,p) for all tokens.
  - the maximal score among all tokens.
  - the score of the last token in the tweet with score(w,p)¿0.
- Punctuation for the number of contiguous sequences of exclamation marks, question marks, and both exclamation and question marks. Another feature checks whether the last token contains an exclamation or question mark.
- Emoticons where the polarity of an emoticon is determined with a regular expression adopted from Christopher Potts' tokenizing script as described earlier in this section. We did not implement this feature as it is already part of our pre-processing steps.
- Elongated words refers to the number of words with repeated character sequences. This feature is a little different from our pre-processing steps as it counts the number of such elongated words.
- Clusters refer the presence or absence of tokens from 1000 clusters provided by the CMU pos-tagger and produced with the Brown clustering algorithm on 56 million English tweets.
- Negation counts the number of negated contexts. A negated context is defined as a segment of a tweet that starts with a negation word and ends with one of the punctuation marks"' ',', '.', ':', ';', '!', '?'. Throughout the rest of the paper, we refer to the whole set of NRC features as the full feature set of NRC while the set that we implemented as NRC features. The subset of features that we employed from NRC are all the features except for the word ngram and character ngram features. The reason we excluded these features from NRC is mainly their size. For example, the full NRC are around 4 million features for the subjectivity classifier when trained on the SemEval dataset. Similar results in this example could be retained with 20,200 features with some of our feature combinations as shown in Section 4.

4 Experiments

In this section, we describe the set of experiments conducted in order to select a compact feature subset from the different feature types described in section 3. The aim is to reduce the dimensionality of the feature space used for sentiment classification without scarifying the classification accuracy.

4.1 Datasets

We have employed the following two benchmark datasets for our experiments as summarized in Table 1:

SemEval. This dataset was constructed for the Twitter sentiment analysis task (Task 2) [15] in the Semantic Evaluation of Systems challenge (SemEval-2013) [7]. All the tweets were manually annotated by 5 Amazon Mechanical Turk workers with negative, positive and neutral labels. The turkers were also asked to annotate expressions within the tweets as subjective or objective. The statistics of each sentiment label is shown in Table 1. Participants in the SemEval-2013 Task 2 used this dataset to evaluate their systems for expression-level subjectivity detection [5,13] as well as tweet-level subjectivity detection [12,20].

CrowdScale Dataset. The CrowdScale dataset [8] is the sentiment analysis judgment dataset in CrowdScale 2013. The tweets in the dataset is from the weather domain. Each tweet was evaluated by at least 5 raters. The possible answers are: "Negative", "Neutral"; the author is just sharing information, "Positive", "Tweet not related to weather condition" and "I can't tell". The total number of tweets in addition to the number of tweets of each sentiment for training and testing is shown in Table 1.

Table 1. Description of datasets used in our experiments

Dataset	Training Set				Development Set				Testset			
	Positive	Negative	Neutral	Total	Positive	Negative	Neutral	Total	Positive	Negative	Neutral	Total
SemEval	3,168	1,380	4,111	8,659	500	340	1160	2000	1,570	601	1,638	3,809
CrowdScale	14,253	15,513	20,234	50,000	3,237	3,496	4,510	11,243	3,237	3,496	4,510	11,243

4.2 Experimental Setup

We have conducted extensive experiments to explore the various combinations of features among the four most significant feature sets in the literature. Our experiments target exploring the best features for the subjectivity, polarity and ensemble classifier. All the experiments have been performed on the development sets to select the features while the test sets have been used to compare our system to the baseline system

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

[8] http://www.crowdscale.org/shared-task/sentiment-analysis-judgment-data

using the selected features. The four feature sets discussed in Sections 3.1, 3.2, 3.3, and 3.4 resulted in 16 combinations on each of the three data sets discussed in Section refsec:allDatasets. The baseline is the model trained with the unigram and bigram featuers only. We then add one or more feature set(s) to the baseline to measure the improvement both in the subjectivity, polarity and ensemble classifiers independently. We used macro-F1 as an evaluation metric for all our experiment results. Macro-F1 is the average F1 across the positive, negative and neutral classes. For each feature set combination, we trained 2 types of models: subjectivity model trained using the neutral tweets as class 0 and the subjective tweets (positive or negative) as class 1 and polarity model trained using the positive tweets as class 1 and the negative tweets as class 0.

4.3 Results

This section depicts the results of the subjectivity and polarity classifiers independently followed by the results of the ensemble. This is motivated by our research goals in finding the best combination of features for each classifier alone as well as the ensemble.

Subjectivity Classification. The macro-F1 results of the subjectivity classifier for each feature set combination on the three datasets are shown in Figure 2 For the SemEval dataset, the maximum macro-F1 0.74 is from the NRC and SSWE combination while the maximum macro-F1 for CrowdScale is 0.83 from the full NRC features. For deciding the best feature set combination for the subjectivity classifier, we took the average macro-F1 across the two datasets and found that the best feature set combination for subjectivity classifier is the full NRC features for an average macro-F1 of 0.796. However, the feature vector spans multiple millions. For instance, the SemEval dataset, the feature vector had about 4 million features. The combination of features that retained the second best average macro-F1 of 0.79 is the NRC, the unigram-bigragm and the senti-features with a difference in the macro-F1 of 0.006 less than the full NRC features. We chose the second best combination of (NRC + unigram-bigram + senti-features) combination to serve as the feature set for the subjectivity classifier when integrated in the ensemble.

Polarity Classification. Figure 3 shows the macro-F1 scores of the polarity classifier for each dataset and for each feature set combination. The NRC and SSWE combination is the best combination for SemEval while unigram-bigram and SSWE is the best for CrowdScale. After averaging the macro-F1 of all the datasets, we found that the best feature set combination is NRC, unigram-bigram and SSWE with macro-F1 of 0.89.

Ensemble Results. The ensemble results shown in Table 2 are retained from the best combination of features for the subjectivity classifier (NRC + unigram-bigram + senti-features) and the polarity classifier (NRC + unigram-bigram + SSWE). The results are obtained from applying the best combinations on the test sets. Our baseline system is the ensemble with unigram-bigram features for both the subjectivity and polarity classifiers. The baseline results are shown for each dataset in Table 2. We can see from the results

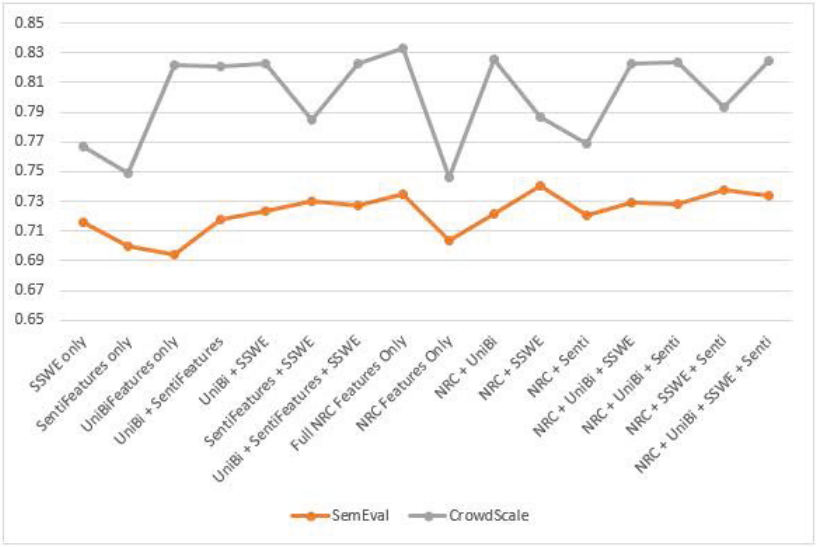

Fig. 2. Macro-F1 of subjectivity classifier for each feature set on each dataset

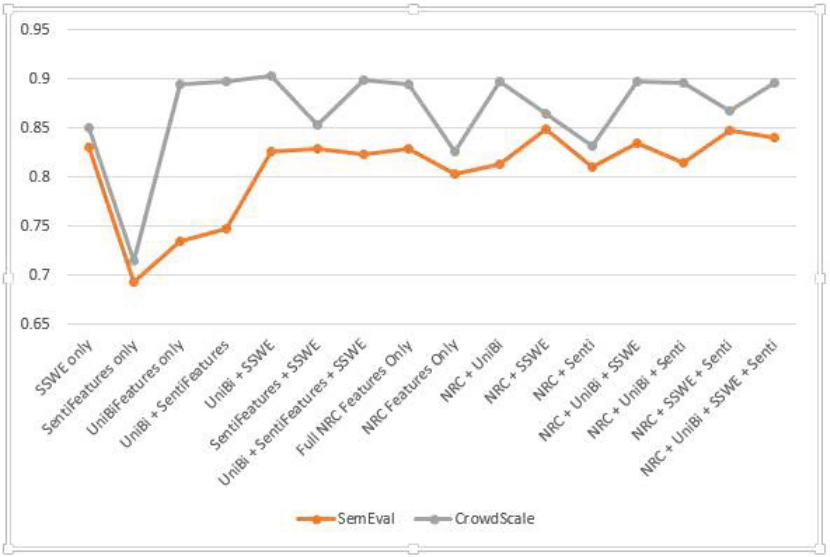

Fig. 3. Macro-F1 of polarity classifier for each feature set on each dataset

that our proposed combination of features have achieved 9.9% and 11.0% relative gain in the macro-F1 of both CrowdScale and SemEval over the unigram-bigram baseline. The size of the feature vector in each data set is few hundreds above the 20,000 unigram-bigram selected features, which maps to much reduced classification complexity.

Table 2. Macro-F1 Results for the Sentiment Ensemble

Dataset	Features	Positive	Negative	Neutral	Macro-F1	Relative Gain
CrowdScale	Baseline	*0.71*	*0.63*	*0.79*	*0.71*	
	Best	*0.77*	*0.76*	*0.82*	*0.78*	**9.9%**
SemEval	Baseline	*0.66*	*0.43*	*0.67*	*0.59*	
	Best	*0.70*	*0.57*	*0.70*	*0.66*	**11.9%**

4.4 Discussion

Figure 4 shows the results of McNemar significance test for subjectivity classifier trained on different feature subsets. The significance test has shown that there is not significant difference in terms of classification accuracy between the full NRC feature set and the selected feature subset. That means that we have reduced the computational complexity of the subjectivity classifier without sacrificing its prediction ability. Our explanation to this result is that the unigram and bigrams features with LLR feature selection capture the highly correlated ngram features for subjectivity classification. The selected features in the significance test we performed are the NRC (the statistical features that do not include the character and word ngrams), unigram-bigram and senti-features. The combination of the unigram-bigram and senti-features are the main difference between the two models we compared in the significance test. This demonstrates that the selected ngram features are equally effective compared to the large size of character and word ngrams in the full NRC feature set. Our ongoing work involves more automatic feature selection techniques to further reduce the computational complexity of both the subjectivity and polarity classifiers.

	[Full NRC] vs. [UniBi + NRC + Senti]		[Full NRC] vs. [Unigram]	
	McNemar Test	p-value	McNemar Test	p-value
SemEval	2.538	Not Significant Difference p = 0.1	13.26	Significant Difference p is less than 0.001
CrowdScale	4.232	Not Significant Difference p = 0.05	6.064	Significant Difference p = 0.01

Fig. 4. McNemar significance test for subjectivity classifier for different pairs of feature set combinations on each dataset

5 Conclusions

In this paper, we explore multiple sets of features used in the literature for the task of sentiment classification including surface-form, semantic and sentiment features. An ensemble of a subjectivity and polarity classifiers is used for sentiment classification. The classification complexity of this ensemble of linear models is linear with respect to the number of features. Our approach aims to select a compact feature subset from the exhaustive list of extracted features in order to reduce the computational complexity without scarifying the classification accuracy. We evaluate the performance on two benchmark datasets. Our selected 20K features have shown very similar results in subjectivity classification to the NRC state-of-the-art system with 4 million features that

has ranked first in 2013 SemEval competition. Also, our selected features have shown a relative performance gain in the ensemble classification over the ngram baseline of 9.9% on CrowdScale and 11.9% on SemEval.

References

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proceedings of the Workshop on Languages in Social Media, pp. 30–38 (2011)
2. Agarwal, B., Mittal, N.: Optimal feature selection for sentiment analysis. In: Gelbukh, A. (ed.) CICLing 2013, Part II. LNCS, vol. 7817, pp. 13–24. Springer, Heidelberg (2013)
3. Bakliwal, A., Arora, P., Varma, V.: Hindi subjective lexicon: A lexical resource for hindi adjective polarity classification. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012) (2012)
4. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), pp. 36–44 (2010)
5. Chalothorn, T., Ellman, J.: Tjp: Using twitter to analyze the polarity of contexts, Atlanta, Georgia, USA, p. 375 (2013)
6. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. The Journal of Machine Learning Research 12, 2493–2537 (2011)
7. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: 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, vol. 2, pp. 42–47. Association for Computational Linguistics (2011)
8. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision, vol. 150, pp. 1–6. Ainsworth Press Ltd (2009)
9. Han, B., Baldwin, T.: Lexical normalisation of short text messages: Makn sens a# twitter. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (2011)
10. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177. ACM (2004)
11. Hu, X., Tang, J., Gao, H., Liu, H.: Unsupervised sentiment analysis with emotional signals. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 607–618. International World Wide Web Conferences Steering Committee (2013)
12. Martınez-Cámara, E., Montejo-Ráez, A., Martın-Valdivia, M., Urena-López, L.: Sinai: Machine learning and emotion of the crowd for sentiment analysis in microblogs, Atlanta, Georgia, USA, p. 402 (2013)
13. Mohammad, S.M., Kiritchenko, S., Zhu, X.: Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. arXiv preprint arXiv:1308.6242 (2013)
14. Mohammad, S.M., Turney, P.D.: Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 26–34. Association for Computational Linguistics (2010)
15. Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., Wilson, T.: Semeval-2013 task 2: Sentiment analysis in twitter (2013)

16. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010) (2010)
17. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques, pp. 79–86 (2002)
18. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), 1226–1238 (2005)
19. Read, J.: Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: Proceedings of the ACL Student Research Workshop, ACLstudent 2005, pp. 43–48 (2005)
20. Remus, R.: Asvuniofleipzig: Sentiment analysis in twitter using data-driven machine learning techniques. Atlanta, Georgia, USA 3(1,278), 450 (2013)
21. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B.: Learning sentiment-specific word embedding for twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 1555–1565 (2014)
22. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of HLT/EMNLP 2005 (2005)
23. Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., Liu, B.: Combining lexicon-based and learning-based methods for twitter sentiment analysis, vol. 89 (2011)