Necessity of Feature Selection when Augmenting Tweet Sentiment Feature Spaces with Emoticons

Joseph D. Prusa, Taghi M. Khoshgoftaar, Amri Napolitano

jprusa@fau.edu, khoshgof@fau.edu, amrifau@gmail.com

Abstract

Tweet sentiment classification seeks to identify the emotional polarity of a tweet. One potential way to enhance classification performance is to include emoticons as features. Emoticons are representations of faces expressing various emotions in text. They are created through combinations of letters, punctuation marks and symbols, and are frequently found within tweets. While emoticons have been used as features for sentiment classification, the importance of their inclusion has not been directly measured. In this work, we seek to determine if the addition of emoticon features improves classifier performance. We also investigate how high dimensionality impacts the addition of emoticon features. We conducted experiments testing the impact of using emoticon features, both with and without feature selection. Classifiers are trained using four different learners and either emoticons, unigrams, or both as features. Feature selection was conducted using five filter based feature rankers with four feature subset sizes. Our results showed that the choice of feature set (emoticon, unigram or both) had no significant impact in our initial tests when using no feature selection; however, with any of the tested feature selection techniques, augmenting unigram features with emoticon features resulted in significantly better performance than unigrams alone. Additionally, we investigate how the addition of emoticons changes the top features selected by the rankers.

Introduction

When performing tweet sentiment classification, features describing the text of the tweet have to be extracted. In text mining, a common approach for feature extraction is to use words found in the tweet as features. This approach is attractive as its implementation is relatively straightforward and language independent. Despite its simplicity, this approach is effective, since the words used in a tweet should convey the emotional polarity of the tweet. Word features, often referred to as n-grams, are popular and have been used in many studies on sentiment classification; however, there are challenges associated with their use. One issue is that extracting n-grams as features leads to the creation of thousands of features. Additionally, each tweet contains only a

small number of features as they are limited to being 140 characters or less. The resulting feature space is high dimensional and sparse, both of which can negatively impact classifier performance (Yang and Pedersen 1997). Because of this, it is common to use additional methods of feature extraction as alternatives or additions to n-gram features.

One additional source of emotional information in a text passage is an emoticon. An emoticon, also called a smiley, is a text representation of a facial expression and often expresses a feeling that the text alone may not convey.For example, ":-)" may accompany a sentence with unclear sentiment and could help a reader understand it to be positive as the emoticon is a smiley face. Emoticons logically should be beneficial to sentiment classification performance, since they describe what we are interested in classifying, i.e., emotional polarity of a text passage. Unfortunately, emoticons alone are insufficient for classifying tweets as many instances may not include emoticons and the benefit of adding emoticons to other types of features, such as n-grams, may compound the problem of high dimensionality due to increasing the size of the feature space. To address this issue the combination of multiple feature types could be accompanied by the use of feature selection techniques. Feature selection techniques select a subset of features in an effort to reduce the effects of high dimensionality.

In this study, we seek to determine if adding emoticon features improves classification performance. We trained classifiers using four learners, C4.5 decision tree, Multilayer Perceptron, 5-Nearest Neighbors and Logistic Regression. We conducted two sets of experiments. First, no feature selection was employed. Each learner was trained on three datasets, created using the same instances but with different methods of feature extraction (unigrams, emoticons or both unigrams and emoticons). The second set of experiments were conducted using either unigram features or both unigram and emoticon features and classifiers trained using feature selection. We combined the four learners with five feature rankers and four different feature subset sizes to again test the impact of using emoticon features, while addressing concerns about high dimensionality impacting results.

Our experimental results showed that there is no significant difference between using unigram, emoticon or both unigram and emoticon features when not conducting feature selection; however, when performing feature selection using

any of the five rankers and four subset sizes adding emoticon features improves classification performance. This improvement is observed for all learners, subset sizes and rankers. Furthermore, statistical tests performed using Analysis of Variance (ANOVA) and Tukey's Honestly Significant Difference test (HSD) (Berenson, Goldstein, and Levine 1983) confirm this improvement is significant. Additionally, we examine changes in the top features selected by each ranker when emoticons are added to the feature space. The results of our study indicate that while the inclusion of emoticons significantly improves classifier performance, their addition has little impact if techniques such as feature selection are not employed to reduce the feature space.

The paper is divided into the following sections. Related Works discusses previous research on how emoticons and feature selection have been used in the context of tweet sentiment. The Dataset and Feature Extraction section details how we constructed our training sets and how we extracted both our unigram and emoticon features. Experimental Methodology provides details on the learners we selected, how they were configured and the process used to train and evaluate our models. The performance of each model, accompanying statistical tests and a discussion on emoticons is provided in Results. Finally, the Conclusion section contains our closing remarks and suggestions for future work.

Related Works

Emoticons are common in informal text, such as tweets, and have been used by multiple research groups for a variety of purposes, both for preparing and labeling datasets and training classifiers. While emoticons have been considered as part of a larger feature extraction framework, their individual impact on classifier performance has not been evaluated despite the relative ease with which they can be extracted compared to other feature types which may require complicated natural language processing tools.

Go et al. (Go, Bhayani, and Huang 2009) demonstrated that the emoticons could successfully be used to identify the sentiment of tweets by using the Twitter search API to find and collect tweets containing emoticons. Using the sentiments associated with the emoticons used in the search, they constructed a corpus of 800,000 positive and 800,000 negative tweets. While this method of labeling training instances likely contains errors, the work of Go et al. and related experiments show that emoticons are a strong indication of the sentiment of a tweet.

A similar study was conducted by Davidov et al. (Davidov, Tsur, and Rappoport 2010), who generated a dataset that was comprised of 15 sets of 1000 tweets using 15 emoticons. They compared classification results for models trained using this dataset against one created using hashtags (tags created by twitter users). When performing binary classification they found their models performed better when trained on the data labeled using emoticons, again demonstrating the relationship between emoticons and the sentiment of a text passage.

Researchers have also extracted and used emoticons as features. Flekova et al. (Flekova, Ferschke, and Gurevych

2014) evaluated models trained on the SemEval 2014 dataset using a combination of lexical, negation, n-gram, and "other" features (which included emoticons); however direct measurement of the impact of emoticon features was not included as the study focused primarily on the addition of syntactic and lexical features. Kouloumpis et al. (Kouloumpis, Wilson, and Moore 2011) conducted a study of n-gram, lexicon, part of speech and microblog features in various combinations. Their microblog features included emoticon features in addition to features checking for intensifiers such as character repetitions, and common abbreviations. Using n-grams as a baseline, they found adding microblogging features to n-grams and lexicon features improved performance, but they did not measure using emoticons or other elements of the microblog features alone.

Feature selection techniques have been demonstrated to be effective for a wide range of text classification problems, including sentiment analysis. Feature selection reduces data dimensionality and can improve classification performance due to reduction of over-fitting and the removal of redundant or misleading features (Guyon and Elisseeff 2003). Saif et al. (Saif, He, and Alani 2012) performed feature selection on unigram features using Information Gain (IG) and determined that there was no benefit to using over 500 features when training models using Naïve Bayes on a tweet dataset with 34,855 initial features. In our previous work, we investigated tweet sentiment and feature selection using 10 feature rankers with 10 subset sizes and found 5 rankers significantly improved classifier performance (Prusa, Khoshgoftaar, and Dittman May 2015).

Unlike other studies using emoticon features, the goal of this paper is to measure the effect adding emoticons has on classifier performance. Additionally, as adding emoticon features to unigrams will inherit the high dimensionality and feature sparsity resulting from using unigram features, we measure the impact of adding emoticons when conducting feature selection. Based on our previous work, we have selected five feature rankers and four subset sizes to test, with unigram and unigram plus emoticon features. Statistical tests are performed on our results, to determine if the findings are significant. To the best of our knowledge this is the first study to investigate how combining multiple methods of feature extraction is influenced by high dimensionality and the impact feature selection has in this scenario in the tweet sentiment domain.

Dataset and Feature Extraction

Datasets

We constructed three datasets using instances from the SemEval 2015 task 10, subtask B dataset (Rosenthal et al. 2015). The SemEval dataset contains manually labeled tweets that are either positive, negative, neutral or objective (factual or uninfluenced by feelings) in sentiment. We used all tweets labeled as positive or negative. The first dataset contains only emoticon features, the second contains unigram features, and the third contains both unigram and emoticon features. All three of our training sets contain the same 3420 positive sentiment and 1399 negative sentiment

tweets. In our experimental results, the datasets are referred to by the features they contain, either "unigrams", 'emoticons", or "both".

Feature Extraction

In order to train our classifiers, we need features representing the dataset so that we can learn patterns about the training instances. While the text of each tweet is not directly usable by our learners, we can extract usable features from the tweet. The process of creating features to describe our dataset is known as feature extraction. For our case study we performed three methods of feature extraction: unigrams, emoticons and a combination of emoticons and unigrams.

As a baseline method of feature extraction, we extracted the 2500 most frequently occurring unigrams in the dataset. When extracting unigrams as features, we first removed symbols, punctuation marks and web URLs from tweet, and made all letters lowercase. Additionally, any character repeated more than two times is reduced to only repeating twice. The resulting feature set was used to create our first dataset, denoted as "unigrams" in our Results section.

Our second feature type, emoticons, are combinations of letters, punctuation marks and symbols that seek to create a face conveying an emotion. For example ":-)" is a smiley face and ":-(" is a frown. When extracting features for our dataset we used a dictionary of 133 emoticons, compiled from multiple websites listing emoticons and associated emotions. Emoticons are used as binary features, either present or absent. We found 45 of the 133 emoticons contained in our dictionary in the SemEval dataset and extracted these emoticons as features. The use of the training set generated with this method is denoted as "emoticons".

We also created a third set of features by combining unigrams and emoticons into a single feature set containing all 2500 unigrams and all 45 emoticons extracted using the above methods. Results for classifiers trained on the dataset containing this third feature set is denoted with "both".

Feature Selection

Feature selection techniques seek to reduce problems associated with high dimensionality by selecting a subset of features. High dimensionality can reduce classifier performance due to over-fitting and the presence of redundant or useless features. There are many types of feature selection techniques; however, we elected to use five filter based feature rankers found to significantly improve classifier performance in our prior work (Prusa, Khoshgoftaar, and Dittman May 2015). These rankers are: Chi-Squared (CS) (Forman 2003), Mutual Information (MI) (Peng, Long, and Ding 2005), area under the Receiver Operating Characteristic curve (ROC) (Dittman et al. 2010), area under the Precision-Recall Curve (PRC) (Dittman et al. 2010) and Kolmogorov-Smirnov statistic (KS) (Dittman et al. 2010). Due to space limitations these techniques will not be discussed in detail. We use these rankers to select feature subsets containing the top ranked 75, 100, 150 or 200 features. These values were selected as our preliminary research found these subset sizes significantly improved performance compared to using no feature selection, and larger

feature subsets did not significantly improve classifier performance compared to these subset sizes. Tests are not conducted using feature selection with emoticon features alone, since only 45 emoticons were found in the dataset.

Experimental Methodology

Our experiments were conducted using four supervised classification algorithms, C4.5 decision tree (C4.5), Multilayer Perceptron (MLP), K-Nearest Neighbor (K-NN) and Logistic Regression (LR). These learners are commonly used in machine learning and offer a diverse set of algorithms with which to evaluate the use of emoticon features. All learners were implemented in the WEKA toolkit (Witten and Frank 2011). Changes to default parameters, determined by previous experimentation to improve performance, are noted below; however, due to space constraints each learner can not be discussed in detail.

For K-NN, we elected to use five nearest neighbors, chosen by setting "k = 5". Additionally we used inverse distance weighting by setting the "distanceWeighting" parameter to "Weight by 1/distance". For C4.5, we enabled "no pruning" and "Laplace Smoothing" as these settings were demonstrated to improve performance (Weiss and Provost 2003). For the artificial neural network MLP, we set "hiddenLayers" to "3", resulting in the neural network having 1 hidden layer with three nodes. We set it to leave aside 10% of our training data to be used for validation to determine when to stop training the network by setting "validationSetSize" to "10". No changes to default parameters were made for LR.

All models were trained using five-fold cross-validation. This technique partitions the dataset into five equal parts and uses four parts at a time to train a model, while using the fifth for validation. By alternating which partition is left out for validation all portions of the dataset are used to train models without using the same data to test a model as was used to train it. Cross-validation was repeated four times to eliminate bias due partition creation.

Classification performance of our models is evaluated using Area Under the receiver operating characteristic Curve (AUC) (Witten and Frank 2011). Since the receiver operating characteristic curve measures performance across the entire range of class distributions and error costs the classifier may encounter, AUC provides a numeric representation of how well a classifier performs in a variety of situations (Provost, Fawcett, and Kohavi 1998). Additionally, AUC is an appropriate metric for imbalanced data, where there is a majority and minority class. As our dataset is imbalanced, with roughly a 3:1 positive:negative class ratio, AUC is a good metric for evaluating classifier performance.

Results

This section presents the results of our case study comparing the use of emoticon and unigram features and provides statistical tests verifying the significance of our findings. Additionally, a subsection is included that provides a discussion of which emoticons were useful and how they impacted feature selection. Classification performance results for each

Table 1: Classification Results Using all Features

learner	unigram	emoticons	both
C4.5	0.72120	0.58552	**0.73100**
LR	0.59256	**0.59766**	0.59146
MLP	0.46459	**0.59532**	0.44948
5-NN	0.64365	0.60078	**0.64895**

of the four learners using the full feature sets of extracted unigrams, emoticons, and the combination of unigrams and emoticons are presented in Table 1. The highest AUC value for each learner is highlighted in **boldface**.

Using both emoticons and unigrams results in the highest observed performance for C4.5 and 5-NN, but using emoticons alone performs better for LR and MLP. While unigrams never have the highest performance, they have higher performance than using emoticons when using either C4.5 or 5-NN and better than using both types of features for LR and MLP.

As seen in Table 2, when using Feature Selection (FS), the addition of emoticon features is beneficial for all rankers and feature subset sizes. Additionally, a comparison of classifier performance in Table 1 and Table 2 show that using feature selection results in higher classification performance than was observed when using no feature selection. The highest performing classifier was trained using LR with 200 features and either KS or ROC as a ranker; however, CS and MI perform better than KS overall. While the highest performing classifier was trained using 200 features and using 200 features often offers the best performance, using 150 features offers comparable results. The difference between using unigrams and using both unigrams and emoticons is fairly small, generally resulting in a difference of AUC of between 1-2%.

The significance of our results was tested using ANOVA to verify our observations about the impact of using emoticon features. The results when using no feature selection are presented in Table 3 and show that there is no significant difference in classification performance between using any of the three feature sets (emoticons, unigrams or their combination). The results of an ANOVA test for the classifiers trained using feature selection are presented in Table 4 and show that the choice of feature set (unigrams or both unigrams and emoticons), ranker, and feature subset size all significantly impact classification performance. To determine which factors offer the best performance, a Tukey's HSD test was conducted for each factor using MATLAB. The results are presented in Figures 1 through 3.

Looking at the results of the HSD test in Figure 1, we observe that there is a significant difference between using both emoticons and unigrams compared to using unigrams alone. Figure 2 shows that CS is the highest performing ranker, followed by MI. The remaining three rankers are not significantly different from each other. Finally, Figure 3 shows that while using 200 features yields the highest performance, it is not significantly better than using 150, though both are significantly better than using 75 features.
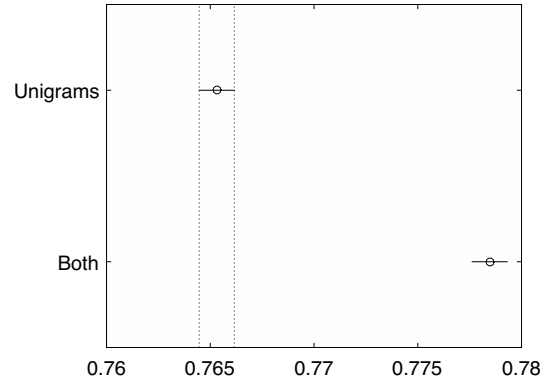

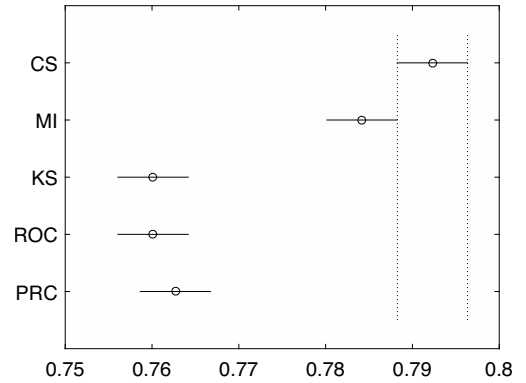Figure 1: HSD Test comparing feature sets when using feature selection

Figure 2: HSD Test for ranker

Emoticon Discussion

As seen in our results, the inclusion of emoticons in combination with unigrams outperforms using unigrams alone when performing feature selection. Without feature selection, the addition of emoticons offers no significant improvement to classifier performance. As previously mentioned, using unigrams as features results in a very large number of features being extracted. Our current experiments show that high dimensionality has the additional effect of hiding or diminishing the impact of including emoticons. Despite emoticons clearly containing valuable information on the sentiment of a tweet, their addition results in no significant change in classifier performance if feature selection is not used. It appears that high dimensionality must be addressed when adding additional types of features to unigrams to ensure the benefit of their inclusion can be observed.

As discussed in the section on feature extraction, 45 emoticons were extracted as features; however, most of these emoticons were not included among the top features when conducting feature selection. Table 5 presents how many different emoticon features were selected for each subset size and ranker combination. It can be observed that the number of emoticons increases with feature subset size and that no more than seven were ever selected as part of the top 200 features by any single ranker. Using PRC results in the

Table 2: Classification Results Using Feature Selection

FS Technique	Learner	Feature Subset Size							
		200		150		100		75	
		unigrams	both	unigrams	both	unigrams	both	unigrams	both
CS	C4.5	0.79051	**0.80275**	0.78915	**0.80133**	0.79002	**0.80588**	0.78515	**0.80022**
	LR	0.82424	**0.83693**	0.82136	**0.83511**	0.81376	**0.82910**	0.80762	**0.82276**
	MLP	0.81734	**0.83494**	0.80754	**0.82627**	0.79547	**0.81427**	0.78945	**0.80453**
	5-NN	0.71732	**0.72843**	0.72389	**0.73621**	0.74123	**0.75856**	0.74290	**0.75957**
ROC	C4.5	0.77558	**0.79030**	0.77083	**0.78532**	0.75395	**0.77154**	0.74308	**0.75696**
	LR	0.82529	**0.84138**	0.81657	**0.83394**	0.80232	**0.81827**	0.79186	**0.80655**
	MLP	0.80413	**0.82328**	0.79516	**0.81223**	0.76714	**0.78478**	0.75915	**0.77554**
	5-NN	0.65437	**0.66439**	0.65847	**0.66482**	0.66254	**0.66902**	0.67081	**0.67441**
PRC	C4.5	0.77584	**0.78561**	0.77070	**0.78116**	0.76801	**0.77757**	0.75741	**0.77278**
	LR	0.81668	**0.82951**	0.81437	**0.82736**	0.80739	**0.82141**	0.79713	**0.81140**
	MLP	0.78973	**0.77833**	0.78399	**0.80433**	0.77685	**0.79356**	0.76600	**0.78207**
	5-NN	0.65702	**0.66439**	0.66773	**0.67102**	0.68166	**0.68604**	0.69082	**0.69833**
MI	C4.5	0.77875	**0.78961**	0.77963	**0.78922**	0.78086	**0.79368**	0.77919	**0.79273**
	LR	0.81829	**0.83202**	0.81739	**0.83150**	0.81356	**0.82628**	0.80457	**0.81815**
	MLP	0.80627	**0.82130**	0.80321	**0.81989**	0.79473	**0.80979**	0.78516	**0.80364**
	5-NN	0.70395	**0.71214**	0.71127	**0.71823**	0.72582	**0.73768**	0.74061	**0.75505**
KS	C4.5	0.77558	**0.79030**	0.77083	**0.78532**	0.75395	**0.77115**	0.74308	**0.75696**
	LR	0.82529	**0.84138**	0.81657	**0.83394**	0.80232	**0.81829**	0.79186	**0.80655**
	MLP	0.80465	**0.82256**	0.79531	**0.81207**	0.76728	**0.78510**	0.75915	**0.77554**
	5-NN	0.65437	**0.66439**	0.65847	**0.66482**	0.66254	**0.66920**	0.67081	**0.67441**

Table 3: ANOVA Test for Results Using all Features

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Feat.Sets	0.00593	2	0.00296	0.27	0.7619
Error	2.58004	237	0.01089		
Total	2.58597	239			

Table 4: ANOVA Test for Classification Results Using Feature Selection

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Feat.Sets	0.13845	1	0.13845	49.15	2.88E-12
FS	0.59492	4	0.14873	52.8	3.91E-43
subset	0.06163	3	0.02054	7.29	7.15E-05
Error	8.98858	3191	0.00282		
Total	9.78359	3199			

Figure 3: HSD Test for subset sizes

least amount of emoticons being selected.

Table 6 presents the emoticon features that were selected along with the frequency of their selection, and the emotion each emoticon is intended to portray. No more than seven of these emoticons were ever selected by a single ranker. Of the eight emoticons, "8D" and "XP" were ranked in the top 75 features by all five rankers and three more emoticons, ":3", "=]" and ":-(", were present in the majority of feature subsets. Interestingly, when examining these five emoticons, four are positive in sentiment while the remaining is negative. Additionally, when considering all eight emoticons, six are clearly positive while only one is clearly negative. This may suggests that positive emoticons were found to be more useful in predicting a tweet sentiment. Alternatively, many negative emoticons may not have sufficient frequency to be of good use, since there are roughly three times as many
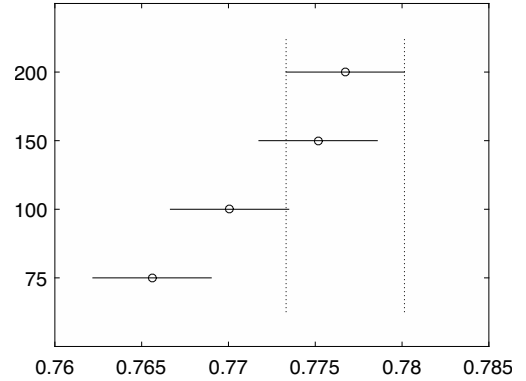
positive instances as negative in the SemEval dataset.

Conclusion

Emoticons contain valuable emotional information and have been used to aid tweet sentiment classification, but their impact on classifier performance, when added as features, had not previously been isolated. Additionally, the effect of high dimensionality when including additional valuable features, such as emoticons, alongside unigrams had yet to be studied. To determine if the inclusion of emoticon features significantly improves the performance of tweet sentiment classifiers, we compared the performance of classifiers trained using only emoticons, only unigrams and the combination of emoticons and unigrams with and without feature selection.

Using no feature selection, we found no significant difference between using the three feature sets. However, ex-

Table 5: Emoticons selected using the Unigrams+Emoticons dataset

FS	subset size			
	75	100	150	200
CS	4	4	6	7
AUC	4	5	7	7
PRC	3	3	5	5
MI	4	4	6	7
KS	4	5	7	7

Table 6: Emoticon Count and Meaning

Emoticon	Count	Emotion
8-0	20	Surprise/Shock
XP	20	Cheeky/Playfull
:3	18	Happy
=]	16	Happy
:-(	12	Sad
XD	7	Laughing/Grin
8D	6	Laughing/Grin
:*	1	Kiss

periments using feature selection found adding emoticon features significantly improved classification performance. Our experiments showed that for all tested combinations of learners, feature selection techniques and subset sizes, using both emoticons and unigrams results in higher performance than using only unigrams. Additionally, the measured difference in AUC, although small, is statistically significant. Further examination of adding emoticons when conducting feature selection showed that only eight of the 45 emoticons in the dataset were responsible for the observed increase in performance, suggesting that the others are either too infrequent in occurrence, or unclear in their use to be helpful when determining the sentiment of a tweet.

We recommend that emoticons be included alongside unigrams in any instance where unigrams are being used as features and recommend using feature selection in order to gain the full benefit of including additional types of features. Future studies should investigate larger sentiment datasets with a wider variety of available emoticons in an effort to create a more comprehensive study of what types of emoticons are important and how the distribution of positive and negative instances influences which emoticons are important.

References

Berenson, M. L.; Goldstein, M.; and Levine, D. 1983. *Intermediate Statistical Methods and Applications: A Computer Package Approach 2nd Edition*. Prentice Hall.

Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 241–249. Association for Computational Linguistics.

Dittman, D. J.; Khoshgoftaar, T. M.; Wald, R.; and Van Hulse, J. 2010. Comparative analysis of dna microarray data through the use of feature selection techniques. In *Proceedings of the Ninth IEEE International Conference on Machine Learning and Applications (ICMLA)*, 147–152. ICMLA.

Flekova, L.; Ferschke, O.; and Gurevych, I. 2014. Ukpdipf: A lexical semantic approach to sentiment polarity prediction in twitter data. *SemEval 2014* 704.

Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* 3:1289–1305.

Go, A.; Bhayani, R.; and Huang, L. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* 1–12.

Guyon, I., and Elisseeff, A. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3:1157–1182.

Kouloumpis, E.; Wilson, T.; and Moore, J. 2011. Twitter sentiment analysis: The good the bad and the omg! *ICWSM* 11:538–541.

Peng, H.; Long, F.; and Ding, C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 27(8):1226–1238.

Provost, F. J.; Fawcett, T.; and Kohavi, R. 1998. The case against accuracy estimation for comparing induction algorithms. In *ICML*, volume 98, 445–453.

Prusa, J. D.; Khoshgoftaar, T. M.; and Dittman, D. J. May 2015. Impact of feature selection techniques for tweet sentiment classification. In *Proceedings of the 28th International FLAIRS conference*, 299–304.

Rosenthal, S.; Nakov, P.; Kiritchenko, S.; Mohammad, S. M.; Ritter, A.; and Stoyanov, V. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval*.

Saif, H.; He, Y.; and Alani, H. 2012. Alleviating data sparsity for twitter sentiment analysis. CEUR Workshop Proceedings (CEUR-WS. org).

Weiss, G. M., and Provost, F. J. 2003. Learning when training data are costly: the effect of class distribution on tree induction. *J. Artif. Intell. Res.(JAIR)* 19:315–354.

Witten, I. H., and Frank, E. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 3rd edition.

Yang, Y., and Pedersen, J. O. 1997. A comparative study on feature selection in text categorization. In *ICML*, volume 97, 412–420.