Feature Selection for Highly Skewed Sentiment Analysis Tasks

Can Liu
Indiana University
Bloomington, IN, USA
liucan@indiana.edu

Sandra Kübler
Indiana University
Bloomington, IN, USA
skuebler@indiana.edu

Ning Yu
University of Kentucky
Lexington, KY, USA
ning.yu@uky.edu

Abstract

Sentiment analysis generally uses large feature sets based on a bag-of-words approach, which results in a situation where individual features are not very informative. In addition, many data sets tend to be heavily skewed. We approach this combination of challenges by investigating feature selection in order to reduce the large number of features to those that are discriminative. We examine the performance of five feature selection methods on two sentiment analysis data sets from different domains, each with different ratios of class imbalance.

Our finding shows that feature selection is capable of improving the classification accuracy only in balanced or slightly skewed situations. However, it is difficult to mitigate high skewing ratios. We also conclude that there does not exist a single method that performs best across data sets and skewing ratios. However we found that $TF * IDF_2$ can help in identifying the minority class even in highly imbalanced cases.

1 Introduction

In recent years, sentiment analysis has become an important area of research (Pang and Lee, 2008; Bollen et al., 2011; Liu, 2012). Sentiment analysis is concerned with extracting opinions or emotions from text, especially user generated web content. Specific tasks include monitoring mood and emotion; differentiating opinions from facts; detecting positive or negative opinion polarity; determining opinion strength; and identifying other opinion properties. At this point, two major approaches exists: lexicon and machine learning based. The lexicon-based approach uses high quality, often manually generated features. The machine learning-based approach uses automatically generated feature sets, which are from various sources of evidence (e.g., part-of-speech, n-grams, emoticons) in order to capture the nuances of sentiment. This means that a large set of features is extracted, out of which only a small subset may be good indicators for the sentiment.

One major problem associated with sentiment analysis of web content is that for many topics, these data sets tend to be highly imbalanced. There is a general trend that users are willing to submit positive reviews, but they are much more hesitant to submit reviews in the medium to low ranges. For example, for the YouTube data set that we will use, we collected comments for YouTube videos from the comedy category, along with their ratings. In this data set, more than 3/4 of all ratings consist of the highest rating of 5. For other types of user generated content, they opposite may be true.

Heavy skewing in data sets is challenging for standard classification algorithms. Therefore, the data sets generally used for research on sentiment analysis are balanced. Researchers either generate balanced data sets during data collection, by sampling a certain number of positive and negative reviews, or by selecting a balanced subset for certain experiments. Examples for a balanced data set are the movie review data set (Pang and Lee, 2004) or the IMDB review data set (Maas et al., 2011). Using a balanced data set allows researchers to focus on finding robust methods and feature sets for the problem. Particularly, the movie review data set has been used as a benchmark data set that allows for comparisons of various sentiment analysis models. For example, Agarwal and Mittal (2012), Agarwal and Mittal (2013), Kummer

Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP), pages 2–11,
Dublin, Ireland, August 24 2014.

and Savoy (2012), O'Keefe and Koprinska (2009), and Paltoglou and Thelwall (2010) all proposed competitive feature selection methods evaluated on the movie review data set. However, the generalizability of such feature selection methods to imbalanced data sets, which better represent real world situations, has not been investigated in much detail. Forman (2003) provides an extensive study of feature selection methods for highly imbalanced data sets, but he uses document classification as task.

This current paper investigates the robustness of three feature selection methods that Forman (2003) has shown to be successful, as well as two variants of $TF * IDF$. The three methods are Odds-Ratio (OR), Information Gain (IG), and Binormal Separation (BNS). BNS has been found to perform significantly better than other methods in more highly skewed tasks (Forman, 2003). The two variants of $TF * IDF$ differ in the data set used for calculating document frequency. We investigate the behavior of these methods on a subtask of sentiment analysis, namely the prediction of user ratings. For this, we will use data sets from two different domains in order to gain insight into whether or not these feature selection methods are robust across domains and across skewing ratios: One set consists of user reviews from Epicurious[1], an online community where recipes can be exchanged and reviewed, the other set consists of user reviews of YouTube comedy videos.

The remainder of this paper is organized as follows: In section 2, we explain the rationale for applying feature selection and introduce the feature selection methods that are examined in this paper. Section 3 introduces the experimental settings, including a description of the two data sets, data preprocessing, feature representation, and definition of the binary classification tasks. In section 4, we present and discuss the results for the feature selection methods, and in section 5, we conclude.

2 Feature Selection and Class Skewing

In a larger picture, feature selection is a method (applicable both in regression and classification problems) to identify a subset of features to achieve various goals: 1) to reduce computational cost, 2) to avoid overfitting, 3) to avoid model failure, and 4) to handle skewed data sets for classification tasks. We concentrate on the last motivation, even though an improvement of efficiency and the reduction of overfitting are welcome side effects. The feature selection methods studied in this paper have been used in text classification as well, which is a more general but similar task using n-gram features. However, since all measures are intended for binary classification problems, we reformulate the rating prediction into a binary classification problem (see section 3.5).

Feature selection methods can be divided into wrapper and filter methods. Wrapper methods use the classification outcome on a held-out data set to score feature subsets. Standard wrapper methods include forward selection, backward selection, and genetic algorithms. Filter methods, in contrast, use an independent measure rather than the error rate on the held-out data. This means that they can be applied to larger feature sets, which may be unfeasible with wrapper methods. Since sentiment analysis often deals with high dimensional feature representation, we will concentrate on filter methods for our feature selection experiments.

Previous research (e.g. (Brank et al., 2002b; Forman, 2003)) has shown that Information Gain and Odds Ratio have been used successfully across different tasks and that Binormal Separation has good recall for the minority class under skewed class distributions. So we will investigate them in this paper. Other filter methods are not investigated in this paper due to two main concerns: We exclude Chi-squared and Z-score, statistical tests because they require a certain sample size. Our concern is that their estimation for rare words may not be accurate. We also exclude Categorical Proportion Difference and Probability Proportion Difference since they do not normalize over the sample of size of positive and negative classes. Thus, our concern is that they may not provide a fair estimate for features from a skewed data sets.

2.1 Notation

Following Zheng et al. (2004), feature selection methods can be divided into two groups: one-sided and two-sided measures. One-sided measures assign a high score to positively-correlated features and a low

[1] www.epicurious.com

score to negative features while two-sided measures prefer highly distinguishing features, independent of whether they are positively or negatively correlated. Zheng et al. (2004) note that the ratio of positive and negative features affects precision and recall of the classification, especially for the minority class. For one-sided methods, we have control over this ratio by selecting a specified number of features one each side; for two-sided methods, however, we do not have this control. In this paper, we will keep a 1:1 ratio for one-sided methods. For example, if we select 1 000 features, we select the 500 highest ranked features for the positive class, and the 500 highest ranked features for the negative class. When using two-sided methods, the 1 000 highest ranked features are selected.

For the discussion of the feature selection methods, we use the following notations:

- S: target or positive class.
- $\overline{S}$: negative class.
- D_S: The number of documents in class S.
- $D_{\overline{S}}$: The number of documents in class $\overline{S}$.
- D_{Sf}: The number of documents in class S where feature f occurs.
- $D_{\overline{S}f}$: The number of documents in class $\overline{S}$ where feature f occurs.
- T_{Sf}: The number of times feature f occurs in class S.

2.2 Feature Selection Methods

In addition to Information Gain, Odds Ratio and Bi-Normal Separation, $TF * IDF$ is included for comparison purposes. We define these measures for binary classification as shown below.

Information Gain (IG): IG is a two-sided measure that estimates how much is known about an unobserved random variable given an observed variable. It is defined as the entropy of one random variable minus the conditional entropy of the observed variable. Thus, IG is the reduced uncertainty of class S given a feature f:

$$IG = H(S) - H(S|f) = \sum_{f \in \{0,1\}} \sum_{S \in \{0,1\}} P(f,S) log \frac{P(f,S)}{P(f)P(S)}$$

Brank et al. (2002b) analyzed feature vector sparsity and concluded that IG prefers common features over extremely rare ones. IG can be regarded as the weighted average of Mutual Information, and rare features are penalized in the weighting. Thus they are unlikely to be chosen (Li et al., 2009). Forman (2003) observed that IG performs better when only few features (100-500) are used. Both authors agreed that IG has a high precision with respect to the minority class.

Odds Ratio (OR): OR (Mosteller, 1968) is a one-sided measure that is defined as the ratio of the odds of feature f occurring in class S to the odds of it occurring in class $\overline{S}$. A value larger than 1 indicates that a feature is positively correlated with class S, a value smaller than 1 indicates it is negatively correlated:

$$OR = log \frac{P(f,S)(1 - P(f,S))}{P(f,\overline{S})(1 - P(f,\overline{S}))}$$

Brank et al. (2002b) showed that OR requires a high number of features to achieve a given feature vector sparsity because it prefers rare terms. Features that occur in very few documents of class S and do not occur in $\overline{S}$ have a small denominator, and thus a rather large OR value.

Bi-Normal Separation (BNS): BNS (Forman, 2003) is a two-sided measure that regards the probability of feature f occurring in class S as the area under the normal distribution bell curve. The whole area under the bell curve corresponds to 1, and the area for a particular feature has a corresponding threshold along the x-axis (ranging from negative infinite to positive infinite). For a feature f, one can find the threshold that corresponds to the probability of occurring in the positive class, and the threshold corresponding to the probability of occurring in $\overline{S}$. BNS measures the separation in these two thresholds:

$$BNS = |F^{-1}(\frac{D_{Sf}}{D_S}) - F^{-1}(\frac{D_{\overline{S}f}}{D_{\overline{S}}})|$$

where F^{-1} is the inverse function of the standard normal cumulative probability distribution. As we can see, the F^{-1} function exaggerates an input more dramatically when the input is close to 0 or 1 which means that BNS perfers rare words.

Term Frequency * Inverse Document Frequency (TF*IDF): $TF * IDF$ was originally proposed for information retrieval tasks, where it measures how representative a term is for the document in which it occurs. When $TF * IDF$ is adopted for binary classification, we calculate the $TF * IDF$ of a feature w.r.t. the positive class (normalized) and the $TF * IDF$ w.r.t. the negative class (normalized). We obtain the absolute value of the difference of these two measures. If a feature is equally important in both classes and thus would not contribute to classification, it receives a small value. The larger the value, the more discriminative the feature. We apply two variants of $TF * IDF$, depending on how IDF is calculated:

$$TF * IDF_1 = (0.5 + \frac{0.5 \times T_{Sf}}{max_i(T_{Sf_i})}) \times log(\frac{D_S + D_{\overline{S}}}{D_{Sf}})$$

$$TF * IDF_2 = (0.5 + \frac{0.5 \times T_{Sf}}{max_i(T_{Sf_i})}) \times log(\frac{D_S}{D_{Sf}})$$

In the first variant, $TF * IDF_1$, document frequency is based on the whole set of examples while in the second variant, $TF * IDF_2$, document frequency is based only on the class under consideration, S.

3 Experimental Setup

3.1 Data Sets

Epicurious Data Set: We developed a web crawler to scrape user reviews for 10 146 recipes, published on the Epicurious website before and on April 02, 2013. On the website, each recipe is assigned a rating of 1 to 4 forks, including the intermediate values of 1.5, 2.5, and 3.5. This is an accumulated rating over all user reviews. (Reviews with ratings of 0 were excluded, they usually indicate that recipes have not received any ratings.) We rounded down all the half ratings, e.g., 1.5 forks counts as 1 fork, based on the observation that users are generous when rating recipes. Our experiments classify each recipe by aggregating over all its reviews. While a little more than half of the recipes received 1 to 10 reviews, there are recipes with more than 100 reviews. To avoid an advantage for highly reviewed recipes, we randomly selected 10 reviews if a recipe has more than 10 reviews. Recipes with less than 3 reviews were eliminated since they do not provide enough information. After these clean-up steps, the data set has the distribution of ratings shown in table 1.

YouTube Data Set: Using the Google YouTube Data API, we collected average user ratings and user comments for a set of YouTube videos in the category *Comedy*. Each video is rated from 1 to 5. The distribution of ratings among all YouTube videos is very skewed, as illustrated in figure 1. Most videos are rated highly; very few are rated poorly. The 1% quantile is 1.0; the 6.5% quantile is 3.0; the 40% quantile is 4.75; the 50% quantile is 4.85; and the 77% quantile is 5.0. We selected a set of 3 000 videos. Videos with less than 5 comments or with non-English comments are discarded.

rating	no.
1 fork	44 recipes
2 forks	304 recipes
3 forks	1416 recipes
4 forks	1368 recipes

Table 1: The distribution of ratings in the Epicurious data set.

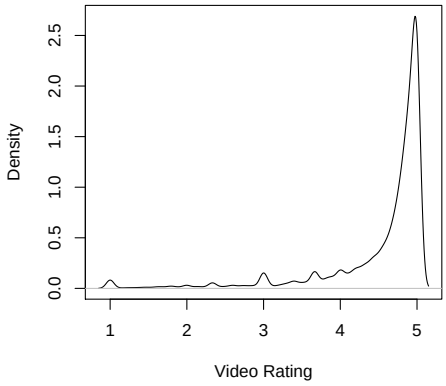

Figure 1: Skewing in the YouTube data set.

3.2 Data Preprocessing

Before feature extraction, basic preprocessing is conducted for both data sets individually. For the Epicurious data set, we perform stemming using the Porter Stemmer (Porter, 1980) to normalize words, and rare words (≤ 4 occurrences) are removed. On the YouTube data set, we perform spelling correction and normalization because the writing style is rather informal. Our normalizer collapses repetitions into their original forms plus a suffix "RPT", thus retaining this potentially helpful clue for reviewer's strong emotion without increasing the features due to creative spelling. For example, "looooooove" is changed to "loveRPT" and"lolololol" to "lolRPT". The normalizer also replaces all emoticons by either"EMOP" for positive emoticons or "EMON" for negative emoticons. Besides a standard English dictionary, we also use the Urban Dictionary[2] since it has a better coverage of online abbreviations.

We do not filter stop words for two reasons: 1) Stop words are domain dependent, and some English stop words may be informative for sentiment analysis, and 2) uninformative words that are equally common in both classes will be excluded by feature selection if the method is successful.

3.3 Feature representation

Since our focus is on settings with high numbers of features, we use a bag-of-words approach, in which every word represents one feature, and its term frequency serves as its value. Different feature weighting methods, including binary weighting, term frequency, and $TF*IDF$ have been adopted in past sentiment analysis studies (e.g., (Pang et al., 2002; Paltoglou and Thelwall, 2010)). (Pang et al., 2002) found that simply using binary feature weighting performed better than using more complicated weightings in a task of classifying positive and negative movie reviews. However, movie reviews are relatively short, so there may not be a large difference between binary features and others. Topic classification usually uses term frequency as feature weighting. $TF * IDF$ and variants were shown to perform better than binary weighting and term frequency for sentiment analysis (Paltoglou and Thelwall, 2010).

Since our user rating prediction tasks aggregate all user comments into large documents and predict ratings per recipe/YouTube video, term frequency tends to capture richer information than binary fea-

[2] `http://www.urbandictionary.com/`

Epicurious			YouTube		
ratio	no. NEG	no. POS	ratio	no. NEG	no. POS
1:8	348	2 784	1:10	56	559
1:1.57	348	547	1:1.57	356	559
1:1	348	348	1:1	559	559

Table 2: Skewing ratios and sizes of positive and negative classes for both data sets.

tures. Thus, we use term frequency weighting for simplicity, not to deviate from the focus of feature selection methods. Since there is a considerable variance in term frequency in the features, we normalize the feature values to [0,1] to avoid large feature values from dominating the vector operations in classifier optimization.

For the Epicurious data, the whole feature set consists of 10 677 unigram features. For YouTube, the full feature set of features consists of 23 232 unigram features. We evaluate the performance of feature selection methods starting at 500 features, at a step-size of 500. For the Epicurious data, we include up to 10 500 features. For the YouTube data, we stop at 15 000 features due to prohibitively long classification times.

3.4 Classifier

The classifier we use in this paper is Support Vector Machines (SVMs) in the implementation of SVM^{light} (Joachims, 1999). Because algorithm optimization is not the focus of this study, we use the default linear kernel and other default parameter values. Classification results are evaluated by accuracy as well as precision and recall for individual classes.

3.5 Binary Classification

Since all feature selection methods we use in our experiments are defined under a binary classification scenario, we need to redefine the rating prediction task. For both data sets, this means, we group the recipes and videos into a *positive* and a *negative* class. A baseline classifier predicts every instance as the majority class. For both data sets, the majority class is *positive*.

For the Epicurious data set, 1-fork and 2-fork recipes are grouped into the negative class (NEG), and 3-fork and 4-fork recipes are grouped into the positive class (POS), yielding a data set of 348 NEG and 2 784 POS recipes (skewing ratio: 1:8). The different skewing ratios we use are shown in table 2. $2/3$ of the data is used for training, and $1/3$ for testing, with the split maintaining the class ratio. Note that for the less skewed settings, all NEG instances were kept while POS instances were sampled randomly.

For the YouTube data set, we sample from all videos with rating 5 for the positive class and from all videos with ratings between and including 1 and 3 for the negative class. This yields 559 POS and 559 NEG videos. The different skewing ratios we use are shown in table 2. $7/8$ of the data is used for training, and $1/8$ for testing, with the split maintaining the class ratio.

4 Results

4.1 Results for the Epicurious Data Set

The results for the Epicurious data set with different skewing ratios are shown in figure 2. The accuracy of the baseline is 50% for the 1:1 ratio, 61% for 1:1.57, and 88.9% for 1:8.

The results show that once we use a high number of features, all the feature selection methods perform the same. This point where they conflate is reached at around 4 000 features for the ratios of 1:1 and 1:1.57. There, accuracy reaches around 71%. For the experiment with the highest skewing, this point is reached much later, at around 8 000 features. For this setting, we also reach a higher accuracy of around 89%, which is to be expected since we have a stronger majority class. Note that once the conflation point is reached, the accuracy also corresponds to the accuracy when using the full feature set. This accuracy is always higher than that of the baseline.

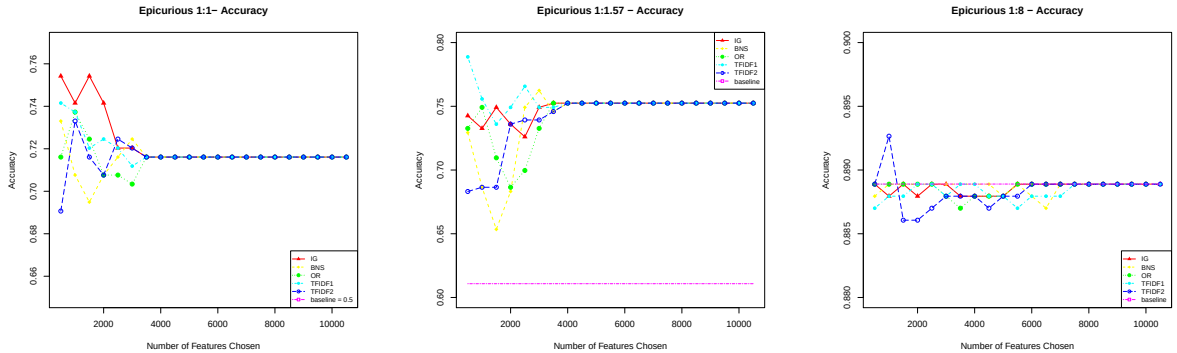

Figure 2: The results for the Epicurious data set.

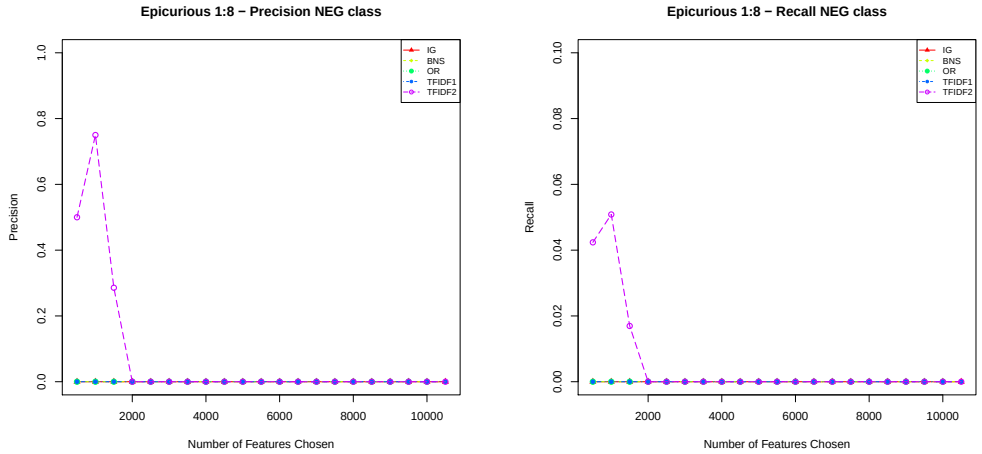

Figure 3: Precision and recall for the negative cases in the Epicurious set, given a 1:8 skewing.

The results also show that the most pronounced differences between feature selection methods occur in the balanced data set. In the set with the highest skewing, the differences are minor, and only $TF*IDF_2$ improves over the baseline when using 1 000 features.

Another surprising result is that $TF*IDF_2$, OR, and BNS have a tendency to fluctuate between higher and lower results than the setting using all features. This means that it is difficult to find a good cut-off point for these methods. $TF*IDF_1$ and IG show clear performance gains for the balanced setting, but they also show more fluctuation in settings with higher skewing.

From these results, we can conclude that for sentiment analysis tasks, feature selection is useful only in a balanced or slightly skewed cases if we are interested in accuracy. However, a look at the precision and recall given the highest skewing (see figure 3) shows that $TF*IDF_2$ in combination with a small number of features is the only method that finds at least a few cases of the minority class. Thus if a good performance on the minority class examples is more important than overall accuracy, $TF*IDF_2$ is a good choice. One explanation is that $TF*IDF_2$ concentrates on one class and can thus ignore the otherwise overwhelming positive class completely. $TF*IDF_1$ and OR have the lowest precision, and BNS fluctuates. Where recall is concerned, $TF*IDF_1$ and IG reach the highest recall given a small feature set.

4.2 Results for the YouTube Data Set

The results for the YouTube data set with different skewing ratios are shown in figure 4. The accuracy of the baseline is 50% for the 1:1 ratio, 61.08% for 1:1.57, and 90.9% for 1:10.

The results show that even though the YouTube data set is considerably smaller than the Epicurious one, is does profit from larger numbers of selected features: For the balanced and the low skewing, there

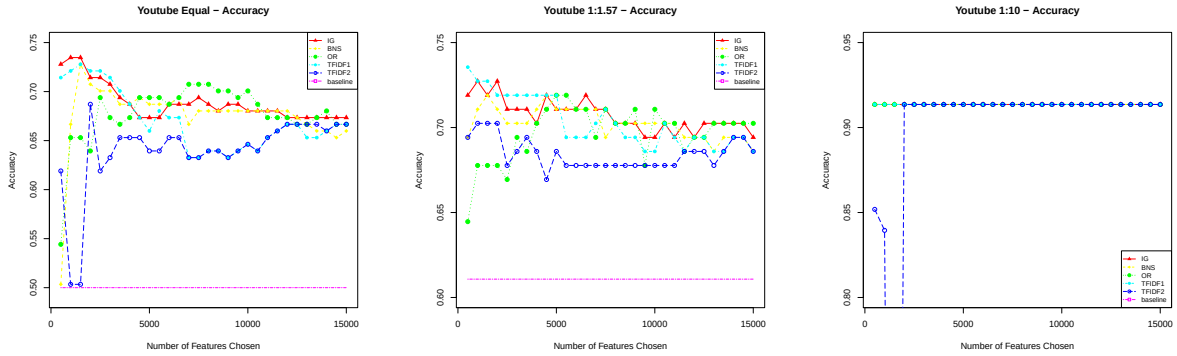

Figure 4: The results for the YouTube set.

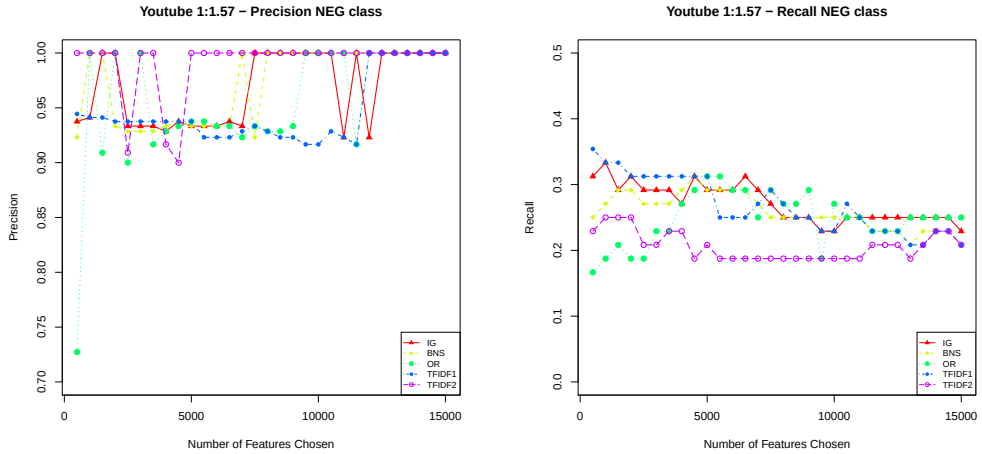

Figure 5: Precision and recall for the negative cases in the YouTube set, given a 1:1.57 skewing.

is no point at which the methods conflate. The results for the highly skewed case show that no feature selection method is capable of finding cases of the minority class: all methods consistently identify only one instance of the negative class. However, this may be a consequence of the small data set. In terms of accuracy, we see that a combination of a small number of features with either IG or $TF * IDF_1$ provides the best results. For the YouTube data set, BNS also performs well but requires a larger set of features for reaching its highest accuracy. We assume that this is the case because BNS has a tendency to prefer rare words. Note that we did not test for number of features greater than 15 000 because of the computation cost, but we can see that the performance curve for different feature selection methods tends to conflate to the point that represents the full feature set.

If we look at the performance of different feature selection methods on identification of minority class instances, we find that $TF * IDF_2$ again manages to increase recall in the highly skewed case, but this time at the expense of precision. For the 1:1.57 ratio, all methods reach a perfect precision when a high number of features is selected, see figure 5. $TF * IDF_2$ is the only method that reaches this precision with small numbers of features, too. However, this is not completely consistent.

4.3 Discussion

If we compare the performance curves across data sets and skewing ratios and aim for high accuracy, we see that there is no single feature selection method that is optimal in all situations. Thus, we have to conclude that the choice of feature selection method is dependent on the task. In fact, the performance of a feature selection method could depend on many factors, such as the difficulty of the classification task, the preprocessing step, the feature generation decision, the data representation scheme (Brank et al., 2002a), or the classification model (e.g., SVM, Maximum Entropy, Naive Bayes).

We have also shown on two different data sets, each with three skewing ratios, that it is difficult for feature selection methods to mitigate the effect of highly skewed class distributions while we can still improve performance by using a reduced feature set for slightly skewed cases. Thus, the higher the skewing of the data set is, the more difficult it is to find a feature selection method that has a positive effect, and parameters, such as feature set size, have to be optimized very carefully. Thus, feature selection is much less effective in highly skewed user rating tasks than in document classification.

However, if the task requires recall of the minority class, our experiments have shown that $TF * IDF_2$ is able to increase this measure with a small feature set, even for highly imbalanced cases.

5 Conclusion and Future Work

In this paper, we investigated whether feature selection methods reported to be successful for document classification perform robustly in sentiment classification problems with a highly skewed class distribution. Our findings show that feature selection methods are most effective when the data sets are balanced or moderately skewed, while for highly imbalanced cases, we only saw an improvement in recall for the minority class.

In the future, we will extend feature selection methods – originally defined for binary classification scenarios – to handle multi-class classification problems. A simple way of implementing this is be to break multi-class classification into several 1-vs.-all or 1-vs.-1 tasks, perform feature selection on these binary tasks and then aggregate them. Another direction that we want to take is integrating more complex features, such as parsing or semantic features, into the classification task to investigate how they influence feature selection. In addition, we will compare other approaches against feature selection for handling highly skewed data sets, including classification by rank, ensemble learning methods, and memory-based learning with a instance-specific weighting. Finally, a more challenging task is to select features for sentiment analysis problems where no annotation (feature selection under unsupervised learning) or few annotations (feature selection under semi-supervised learning) are available.

References

Basant Agarwal and Namita Mittal. 2012. Categorical probability proportion difference (CPPD): A feature selection method for sentiment classification. In *Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology (SAAIP)*, pages 17–26.

Basant Agarwal and Namita Mittal. 2013. Sentiment classification using rough set based hybrid feature selection. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA)*, pages 115–119, Atlanta, GA.

Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2:1–8.

Janez Brank, Marko Grobelnik, Nataša Milic-Frayling, and Dunja Mladenic. 2002a. An extensive empirical study of feature selection metrics for text classification. In *Proceedings of the Third International Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields*, Bologna, Italy.

Janez Brank, Marko Grobelnik, Nataša Milic-Frayling, and Dunja Mladenic. 2002b. Feature selection using Linear Support Vector Machines. Technical Report MSR-TR-2002-63, Microsoft Research.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT-Press.

Olena Kummer and Jaques Savoy. 2012. Feature selection in sentiment analysis. In *Proceeding of the Conférence en Recherche d'Infomations et Applications (CORIA)*, pages 273–284, Bordeaux, France.

Shoushan Li, Rui Xia, Chengqing Zong, and Chu-Ren Huang. 2009. A framework of feature selection methods for text categorization. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 692–700, Suntec, Singapore.

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

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, OR.

Frederick Mosteller. 1968. Association and estimation in contingency tables. *Journal of the American Statistical Association*, 63(321):1–28.

Tim O'Keefe and Irena Koprinska. 2009. Feature selection and weighting methods in sentiment analysis. In *Proceedings of the 14th Australasian Document Computing Symposium (ADCS)*, pages 67–74, Sydney, Australia.

Georgios Paltoglou and Mike Thelwall. 2010. A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1386–1395, Uppsala, Sweden.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Barcelona, Spain.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, Philadelphia, PA.

Martin Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130—137.

Zhaohui Zheng, Xiayun Wu, and Rohini Srihari. 2004. Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter*, 6(1):80–89.