



Analyzing the Impact of Feature Sets on Sentiment Classification Accuracy

Mr.Y.Praveen Konda Reddy¹, Kasarla Chandana²

1 Assistant Professor, Department of CSE, Malla Reddy College of Engineering for Women.,

Maisammaguda., Medchal., TS, India

2, B.Tech CSE (20RG1A05E8),

Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India

Abstract

Analysing text for emotions like happiness or sadness is called sentiment analysis. In order to narrow down a large feature collection, this work employs a number of feature selection methods, including Mutual information, Chi-square, Information gain, and TF-idf. These procedures are assessed using a dataset of 2000 MOVIE reviews. Support vector machine from the weka9 tool is used to carry out the categorization. We also look at the question of what feature works best for determining reviewers' emotions. Word functions, POS tags, and larger word structures are also part of our feature set.

Introduction

Internet is one of the drastically developing areas. People are often communicating, discussing and sharing information through internet. Due to these reasons' internet is one of the essential parts of human life. The information in it covers a wide range of areas such as academic information, feedback or opinion about products, comments about social issues etc. It helps people to think and make decision in many things. Majority of people always listen to others opinion before taking a final decision. Sentimental analysis is one of the research areas. In this information gathering is performed. And the information which is gathered will be analysed in order to determine the sentiment.

of the information such as negative sentiment or positive sentiment. One of the applications of this area is in product purchasing, before purchasing a product people will often enquire about the opinion of the product by other people. In this paper we are presenting methods to extract the sentiment of text data about movie. The objective of this paper is to predict the sentiment of review about the movie. Also, we are interested in knowing the effective feature that can provide better result as well as the best feature selection method. We also perform a comparative study on how can we reduce unigram feature set to get better accuracy on small feature length. Remaining sections are organized as follows: Section 2 delivers some related works on the domain of sentimental analysis; Section 3 contributes an overview of our proposed methodology; Section 4 gives experiment and result; and the final section is the conclusion and future work.

Related work

Salient al.1 presented methods for normalizing the noisy tweets and classified them according to the polarity. The authors of this paper collected 2 million tweets from September 2009 to June 2010 using Twitter search API. They collected tweets related to the mobile operation. To generate sentimental words, they have employed a mixture model approach, and calculated F-score of each word and the words with F-score greater than 10 % will be selected as raw words. As a future work, they suggested a frame work to gain knowledge of the lexicon that can be extracted from the collected tweets so we can represent the words such as *luv*, *lovwww* and *love* as one entity "love". Linguae al.2 proposed a novel multi-domain active learning framework. This framework jointly selects text data from all domains such as BOOKS, DVD, Electronics and Kitchen from Amazon.com. The data set used by this framework is the Multi-Domain Sentiment Dataset. During data pre-processing, they converted all words to lower case and remove the English stop words from the data. The authors of this paper implemented term frequency for weighting features. To create a better classification model, they used LIBLINEAR. Smithy have planned to joint query instances using a hierarchical structure among domains. In



this paper, authors presented framework in the linearly-separable manner and leave the non-separable case to our future work. In this paper Po-it al.3 proposed a new method which extracts the sentiments of micro blogs.

They found that some tweets are mean positive, but it is negative in case of emotion. To overcome these challenges this paper proposed a method which combines supervised learning that is capable of extracting, learning and classifying tweets with opinion expressions. The system is called Opinion Miner. The feature used for this work was unigram. In order to reduce the features in the set, they used mutual information and chi-square as feature selection methods. They have crawled tweets of three distinct categories (camera, mobile phone, and movie) as their training set from the time period November 1, 2012 to January 31, 2013. Naive Bayes classifier was used to classify the tweets. The accuracy was 91% for chi-square. The best accuracy was 96.6%. In this paper Gang Li et al.4 they proposed a new method called clustering-based approach to overcome the drawbacks of the existing methods. This approach was based on the K-mean clustering algorithm. In this the documents are primarily cluster into positive and negative clusters. Then applied TF-idf weighting method to extract more clustering results. Multiple implementations of clustering process were used to obtain the final result. The dataset used for this experiment was movie reviews; it contains 1000 positive and 1000 negative reviews about movie. Compared to the existing methods (Supervised learning and Symbolic techniques), the proposed method produced the accuracy of 77.17% and it is faster in operation too. Yan Dang et al.5 a group of sentiment words built on sentiment lexicon using a method called lexicon-enhanced. They have used these words as a new feature in this paper. This experiment used three features such as Sentiment.

words along with content specific and content free features. The evaluation was performed using 10-fold cross validation. The dataset used contains reviews about DVD, Books, Digital cameras, Electronics, Kitchen appliances. The highest overall accuracy was 84.15%, it is obtained for the product Kitchen appliances. The experiments show that the combination of F1, F2, F3 was giving more accuracy when compared to the individual feature set. 3. Proposed Methodology

Dataset

The corpus used for this work is MOVIE reviews⁷ contains 1000 POSITIVE and 1000 NEGATIVE sample reviews about movie, each will be in an unprocessed HTML files. We extracted only the review data from the HTML file. This dataset is the widely used benchmark dataset for sentiment analysis and domain adaptation

Pre-processing

The samples in the dataset should be pre-processed before performing any type of operation in it. The pre-processing includes x Upper to lower case conversion: For the easiness of feature selection all the data should be converted into lower cases. x Normalization: All words with apostrophizes should be replace with its original form. E.g., don't-> do not. x non-ASCII removal: All non-ASCII characters are removed from the samples. x Remove new lines: The datasets contain some unwanted new lines that are also removed before the feature selection phase. x Remove unwanted punctuations: All punctuations should be removed before feature selection. x Stop word removal: Stop words in the English language are "a", "an", "the", "is". We have removed all words whose length is less than 3, except no, not, none. To remove stop words, we are using Natural Language Toolkit (NLTK)⁸ provided by python. x Stemming: We observed that some of the words in the dataset have similar roots but they may differ only in affixes for example: computer, computation, computing has same root compute. The main purpose of this step is that reducing the feature set and improves the classification performance. We are using Porter stemmer of NLTK provided by python.

Features

Unigram of words:

We are performing some pre-processing on this feature such as stemming and stop word removal. So that we are considering different categories of unigram features such as unigram with stemming with stop word, unigram with stemming without stop words, unigram without stemming with stop words. We are not considering without stemming without stop words category because without stemming will cause high dimensional features, since it



does not reduce to the root form of words. Also, removal of stop words flips the negative samples to positive samples.

Bigram of words:

Like unigrams, we are considering two categories of Bigram such as Bigram without stemming with stop words and Bigram with stemming with stop words. Due to the high dimensionality of bigram features. We are considering features which are appearing more than three times in our dataset.

Parts of speech tags:

We are considering 36 parts of speech tags such VB, PRP, DT, NN etc. Each text in the reviews will be tagged using POS tagger of NLTK9 . The number of tags of reviews varies, since 36 tags are not used in this experiment. We have 26 tags and 25 tags in positive and negative class respectively .

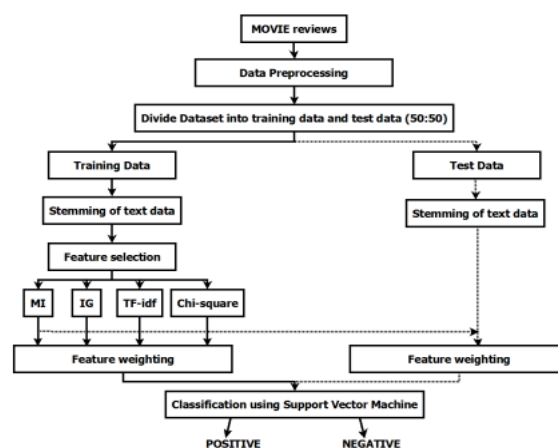


Fig.1. Architecture for Sentiment classification

Function Words:

Function words or grammatical words are the words that have little lexical meaning or have ambiguous meanings. We are considering 375 function words¹⁰. To normalize all these counts we are dividing the count by N (Total number of words).

Word based features:

We are using eight statistical measures¹⁰ exploring that whether the count of words in each sample, helps to extract the sentiments of review. By using these features, we are exploring that whether the count of words in each sample, helps to extract the sentiments of review.

Experimental Results

The experiments are performed on the review data. Before performing the sentiment classification, we apply pre-processing step in the dataset. The features are preferred after pre-processing of the samples in corpus. The pre-processing contains stemming; it is applied to reduce the inflected forms of words to their root form. This is done by the Porter Stemmer module which uses the Porter stemming algorithm⁹. For feature selection, we will use the selection criteria's, which are Mutual information, Information gain, chi-square and TF-idf.

The score is calculated for each feature with respect to the two classes, which are then sorted in decreasing order of their score. SVM classifier is used to perform classification, on features of different length. For features (bag-of-words), the classifier is tested on feature sets of size 100-3000. In the Table 2 and 3 we visualize that accuracy of unigram features in positive class as well as negative class. It is found that unigram with stemming with stop word and unigram with stemming without stop word gives accuracy of 82.9% and 83% with



information gain in positive class. And in negative class also, we are getting good accuracy with information gain. In negative class unigram with stemming and without stop word gives better accuracy of 83.1%. Table 4 and 5 represents accuracy of bigram with different feature selection criteria in both the classes. The performance of unigram is better as compared to the bigram. Bigram without stemming with stop word give better accuracy in both classes. Table 6 depicts the results using Postages. We represent the accuracy (%) observed for both positive reviews and negative reviews such as 53.4% and 53.5% respectively. Function words and Word-based measures produced the results of 64.4% and 67.8% respectively. Comparing the performance of four features such as unigram, bigram, function words and Postages of words, we found that unigram of bag of words is the best feature to extract sentiment from the reviews. In previous work⁶ authors compared results of different dataset and got better results for ensemble feature set. In this experiment we are getting better result for two categories of unigram (A and C).

A : Unigram without stemming with stop word

B : Unigram with stemming with stop word

C : Unigram with stemming without stop word

D: Bigram with stemming with stop word

E: Bigram without stemming with stop word

FL : Feature Length.

Table 2: Accuracy (%) of unigram features in POSITIVE class

FL	MI			IG			Chi-square			TF-idf		
	A	B	C	A	B	C	A	B	C	A	B	C
1300	81.5	80.5	81.2	78.8	78.8	79	79.7	79.4	79.7	78	76.6	74.7
1400	81.8	79.5	81.6	79.6	79	79.8	79.6	79	79.6	79.5	76.9	76.5
1500	81.6	79.6	81.7	80.7	79.7	80.4	80.4	78.6	80.6	79.2	77.1	77
1600	81.1	80	81	81.2	80.3	81.6	81.5	79.2	81.6	79.7	78.6	78.4
1800	80.6	79.6	81	82.4	80.9	83	80.9	78.9	81.3	79.8	79.1	78.6
1900	80.8	79.7	81.4	82.1	80.8	82.5	81.4	79.7	81.3	79.9	80.2	78.4
2000	81.3	79.9	81.5	82.2	80.9	82.6	81	79.7	81.2	79.7	81	79.3
2100	80.3	80.2	80.2	82.9	81.3	82.7	81	79.7	80.8	80.5	81.1	79.9
2300	80.5	80.7	80.3	82.1	81.6	83	81	79.4	80.7	80.9	81.2	80.1
2400	79.9	80.8	80.2	82.2	81.1	82.6	81	79.5	80.8	81.1	80.5	80.2
2500	79.8	80.5	80.4	81.9	81.4	81.9	80	79.8	80.5	80.6	80.8	80

Table 3: Accuracy (%) of unigram features in NEGATIVE class



FL	MI			IG			Chi-square			TF-idf		
	A	B	C	A	B	C	A	B	C	A	B	C
1300	81.6	80.6	81.3	64.2	79.2	79.1	80.1	79.9	80.2	79.2	79.1	76.7
1400	81.9	79.6	81.7	65.8	79	80.7	79.2	79.4	80.7	79.8	79.2	77.4
1500	81.7	79.7	81.8	65.5	79.6	80.7	80.8	79.4	81	79.8	78.9	77.7
1600	81.2	80.1	81.1	65.1	80.6	81.8	80.8	79.4	81.1	80	79.5	78.5
1800	80.7	79.7	81.1	67.1	81	82.7	80.8	80.3	81.1	80.8	79.6	80.6
1900	80.9	79.8	81.5	67.6	80.9	82.3	81	79.2	81.4	81.2	79.9	80.2
2000	81.4	79.7	81.5	66	81.2	82.2	81	80.6	80.8	81.4	80.1	81
2100	80.4	80.3	81.6	67	81.2	82.6	80.4	80.1	80.1	81.5	80	81.6
2300	80.6	80.8	80.4	69.1	81.6	83.1	79.9	80.6	79.8	81.5	80.9	82.2
2400	80	80.9	80.3	69.1	81.4	82.4	81.1	80.4	81	81.9	80.5	82.5
2500	79.9	80.6	80.5	69	81.4	82.2	80.3	80.6	80.4	81.4	80.8	82.1

Table 4: Accuracy (%) of bigram features in POSITIVE class

FL	MI		IG		TF-idf		Chi-square	
	D	E	D	E	D	E	D	E
300	61	60.1	57.5	59.5	57.6	54.7	60.7	54.1
500	61.8	60.8	58.3	58.9	56.6	53.5	63	55.2
600	60.8	60.7	59	58.6	56.2	52.2	63.8	54.6
700	60.4	60.8	59	59	58.7	53.2	63.6	55.8
1900	60.1	59.9	61.7	60	58.7	51.7	61	54.6
2000	60.1	60.7	59.7	61	59.4	52.1	60.9	54.4
2400	60.2	61.6	60	60.3	59.6	53.3	60.9	54.1

Table 5: Accuracy (%) of bigram features in NEGATIVE class

FL	MI		IG		TF-idf		Chi-square	
	D	E	D	E	D	E	D	E
800	57.9	51.4	59.5	51.6	57.8	51.1	57.8	51.5
1500	60.6	49.9	59.4	52.4	58.1	52	58.1	52.8
1900	59.1	50.8	58.6	51.2	59.1	51.5	60.4	53.8
2000	58	52.2	59	51.8	59.1	52.5	59.1	54
2300	58.6	51.7	58.5	52.3	59.4	52.2	59.4	54.2
2500	59.1	52.2	58.4	52.3	58.5	52.1	58.5	52.3
2900	59	51.8	58	52.6	58	52.3	58	53.1

Table 6: Accuracy (%) of POSTag

Number of tags	POSITIVE Class				NEGATIVE Class			
	MI	IG	TF-IDF	Chi-Square	MI	IG	TF-IDF	Chi-Square
5	53.4	49.3	52.1	53.1	53.5	49.4	52.2	53.2
10	50.6	53	52.9	52.6	50.5	53.1	52.8	52.1
15	52.1	52.1	52.2	52.2	52	52.2	52.5	52.7
20	52.1	52.2	52.2	52.2	52	52.6	52.1	52.1
25	52.1	52.2	52.2	52.2	52	52.1	52.1	52.1
26	52.2	52.2	52.2	52.2	-	-	-	-

Table 7: Accuracy (%) of function words and word-based measures



Number features	Name of feature	Accuracy
375	Function words	64.4
8	Word based measures	67.8

Conclusion

The difficulty of categorizing emotions is raised in this research. The best technique for extracting review sentiment was discovered to be unigram. Accuracy in the positive class of 82.9% and 83%, respectively, may be shown when using unigrams with stemming with and without the stop word. Unigrams with stemming and no stop words perform better (83.1%) in the negative class. In terms of gaining knowledge, both groups do well. To further develop this work, we propose experimenting with an ensemble feature selection approach.

References

- [1] . Celikyilmaz, Asli, DilekHakkani-Tur, and Junlan Feng. *Probabilistic model-based sentiment analysis of twitter messages. Spoken Language Technology Workshop (SLT), 2010 IEEE. IEEE, 2010.*
- [2] . Li, Lianghao, et al. *Multi-domain active learning for text classification. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.*
- [3] . Liang, Po-Wei, and Bi-Ru Dai. *Opinion Mining on Social Media Data. Mobile Data Management (MDM), 2013 IEEE 14th International Conference on. Vol. 2. IEEE, 2013.*
- [4] . Li, Gang, and Fei Liu. *A clustering-based approach on sentiment analysis. Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on. IEEE, 2010.*
- [5]. Dang, Yan, Yulei Zhang, and Hsinchun Chen. *A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. Intelligent Systems, IEEE 25.4 (2010): 46-53.*
- [6]. Hsu, Chih-Wei, and Chih-Jen Lin. *A comparison of methods for multiclass support vector machines. Neural Networks, IEEE Transactions on 13.2 (2002): 415-425.*
- [7]. www.cs.cornell.edu/people/pabo/movie-review-data/
- [8]. <http://www.nltk.org/>
- [9]. <http://www.cs.waikato.ac.nz/ml/weka/>
- [10]. Cheng, Na, RajarathnamChandramouli, and K. P. Subbalakshmi. *Author gender identification from text. Digital Investigation 8.1 (2011): 78- 88.*