공학 학사 학위논문

Ensemble learning for twitter sentiment analysis

(Ensemble 학습을 이용한 트위터 감정 분석)

2016 년 8 월

서울개학교 공과 대학

전기.정보공학부

아미르호세인 그다르지

Ensemble learning for twitter sentiment analysis

지도교수: 윤성로

이 논문을 공학 학사학위 논문으로 제출함.

2016 년 06 월

서울대학교 공과대학

전기·정보공학부

아미르호세인 그다르지

아미르호세인 그다르지 의 학사 학위 논문을 인준함

2016 년 06 월 10 일

지 도 교 수 (인)

Abstract

Social Networking Services (SNS) have become an inseparable part of our daily life and it produces a large amount of data which can be processed for multiple purposes such as market analysis, social studies and financial market prediction. Opinion mining and sentiment analysis are one of the main tools to analyze the feelings and emotions of user in a text. It is important to have an accurate sentiment analysis system to provide correct information for the market analysis and further studies. Through this thesis we propose a very accurate corporate related tweet sentiment analysis system to analyze public idea about a company. This system is based on ensemble learning and a new data filtering technique which enhances the accuracy of the system. We have used several methods to filter misspellings and slangs to increase the total accuracy of system. Then running multiple classifiers and optimization techniques, to find the best ensemble combination and achieve the highest possible accuracy. Tests show that our system has a higher accuracy comparing to the existing systems and can be used to analyze the tweet sentiment of a target company and even find the popularity of that company on daily basis.

Keywords: sentiment analysis, ensemble learning, twitter opinion mining,

Contents

List of Tables

List of figures

1. Introduction

21st century is the era of internet and advanced technologies and this has resulted in the appearance and rapid growth of internet. Social Networking Services (SNS) such as Facebook and Twitter are the results of the 21st century's globalization and it has brought a large amount of data for the data scientists to analyze and process. Sentiment analysis and opinion mining is one of the fields that has attracted many data scientists and investors to find the feeling of users in their comments and articles. In this thesis we are proposing an accurate corporate related Twitter sentiment analysis system to provide a corporate popularity analysis system. In order to analyze the huge amount of data exported from tweeter, there are many techniques that we have discussed some of them in Section 2 and taken advantage of them to create an accurate output. We have used new methods to filter the tweets and achieve a more reliable and accurate system which will be discussed in detailed in the Section 3. After the classification and finding the right combination for the ensemble system we have proposed a system and observed its characteristics in the Section 4. Finally in the Section 6 we have explained the advantages and weaknesses of the current system and proposed the future works related to this thesis.

1.1 Defining sentiment

In this thesis we have defined sentiment of a text to be the general positive or negative feeling which is obtained from a text. Due to complexity of the sentiment analysis classification we haven't considered the neutral or other classes, and it has been mentioned and discussed in the Section 5. However the main problem with the sentiment of a text is the relative measurement of the feeling, which depends on the view of reader towards the text. Some readers might find a tweet positive and some might find it negative. In order to reduce this unwanted mismatch, we have prepared a golden data set, reviewed by four different people and taken the prevalence responses for the testing purposes.

Sentiment analysis of large texts is still an ongoing challenge to data scientists; analyzing the sentiment of the last sentence is one of the ways that researchers have tried to solve the problem [1] but still there is no accurate system for a large article by this time. However tweets have a very special characteristic which makes them exceptionally great for sentiment analysis. Tweets are consisted of maximum 140 characters and average one sentence, so finding the sentiment of that sentence according to emoticons or bag of words would be a more possible task [2]. Therefore we have chosen Twitter data because of its characteristics and limited the search queries to find a higher accuracy in a specific field. In the Section 3 we have further discussed about our approach for extracting and analyzing the tweets.

1.2 Contribution

The main contribution of this paper is using a unique filtering technique and increasing the classification accuracy by ensemble classification method. Through this filtering system we were able to increase the accuracy by detecting the slangs and misspellings in each tweet. We have introduced a different approach of filtering and cleaning, which has resulted in a more accurate output. In our classification approach, we have tested 30 classifiers and optimization techniques to find the best fit for the ensemble classifier.

This research also contributes in finding and labeling the tweets using emoticons by introducing a more accurate emoticon table. Also we have reduced the processing time which means our proposed system requires less processing resources and it benefits users by saving their time and resources. These have been discussed more in the Section 3 and Section 4.

Finally we have created a sentiment analysis system which contributes in company popularity analysis. The range of this application can be extended to products popularity analysis in the further studies.

2. Related Work

Sentiment analysis has a very broad range of algorithms and often it is possible to find alternative algorithms for different problems. Through this thesis, we have used 30 classifiers and optimization techniques, but they mostly fit in one of the main groups that is explained below. Finally in the last part of this section we have introduced and explained about correlation coefficient and the importance of it in machine learning techniques. We will discuss our use case in the Section 3.

2.1 Sentiment Analysis

Twitter has become a great source for data scientists to conduct their researches. Because of the simple and special characteristics of tweeter, there have been many researches in the field of sentiment analysis. Pak and Paroubek [3] has classified the tweets into objective, positive and negative classes. In order to collect and distinguish a corpus of objective posts, they retrieved tweets from newspaper accounts such as "New York Times", "Washington Posts" etc. Their classifier is based on the multinomial Naïve Bayes classifier that uses N-gram and POS-tags as features. Then Parikh and Movassate [4] has conducted a more accurate research, implemented Naive Bayes unigram and bigram models and a Maximum Entropy model to classify the tweets. They found that the accuracy of Naive Bayes classifier especially in bigram was better than the Maximum Entropy model. However we got our inspiration from Go et al. [5] who proposed an emoticon based solution for a distant supervised training. Although this approach was initially introduced by Read [6] but Go et al. [5] build their model using Naive Bayes, Maximum Entropy and Support Vector Machines (SVM). Furthermore they have taken advantage of both unigrams, bigrams and reported that SVM outperformed other models and that unigram were more effective as features.

There have been large amount of research in longer texts as well. One of the important use cases of a longer text would be product and movie reviews. Researchers such as Pang et al. [7] have analyzed the performance of different classifiers on movie and customer products. The main idea for labeling these type of reviews comes from the points or stars that is given by users.

As we have read through many papers and articles, the classification results are different for each field and different classifiers are appearing to be superior by different approaches. In the recent studies, many researchers have implemented ensemble approaches to find a more accurate and generalized result to implement on a larger amount of data. Hassan et al [8] utilizes a bootstrap parametric ensemble framework for Twitter sentiment analysis purpose. They have proposed an accurate ensemble approach but it consumes a large amount of time and resources which has been taken in consideration in our research.

In our research, we show that we can produce a more accurate results on tweets with distant supervision using ensemble learning and data filtering techniques.

2.2 Classifiers

2.2.1 Support Vector Machines (SVM)

Support Vector Machines (SVM) [9] is a supervised machine learning technique that is used as a discriminative classifier. This classification technique is famous for its linear classification superiority. However using different kernels gives us the advantage of working with nonlinear data which makes this classifier an ideal and accurate classification technique. This approach has been depicted in the figure 2-1.

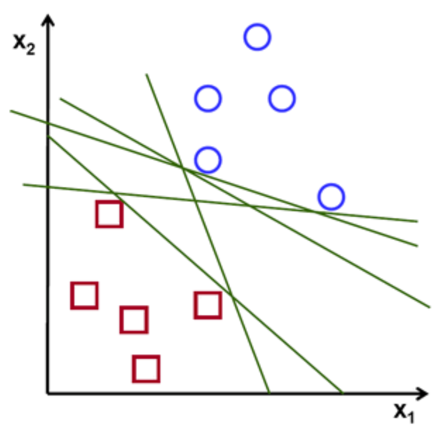

Figure 2-1. Demonstration of SVM classification

In the SVM classification method, given a set of training labeled data, it builds a hyperplane that can be used for classification purposes. It is mainly based on the separation of the data on the hyperplane, therefore a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class. Due to the time restrictions we have only used SVM as a linear model in this experiment but in the future works, other kernels can be adapted to observe a better classification result.

2.2 K-nearest neighbor (KNN)

K-nearest neighbor (KNN) [10, 11] algorithm is one of the other commonly used machine learning techniques. It is one of the simplest machine learning techniques but it shows a good amount of accuracy through the classification.

In KNN classification, k is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point [9]. Therefore by choosing a random point, computer starts calculating the distance from other points and classifies the closest ones into a class and so on. This technique has been demonstrated in figure 2-2. Through this experiment we have used a subset of this algorithm called Nearest Centroid. This method assigns the label of the class of training samples whose mean (centroid) is closest to the observation.

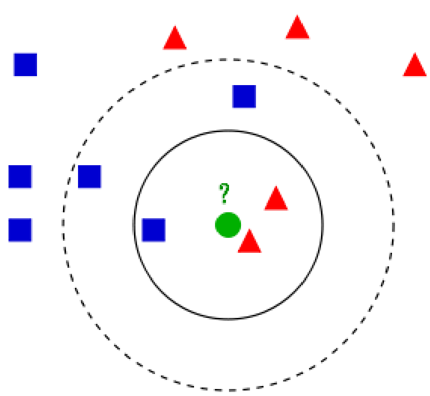

Figure 2-2. Demonstration of KNN classification [11]

2.2.3 Decision Tree

Decision tree [12] is a machine learning technique which maps observations about an item to conclusions about the item's target value. This algorithms uses a tree structured shape in which leaves represent the labels of each target value. If we provide continuous values in these kind of trees, then it is called to be a regression tree. As Decision trees are demonstrated in figure 3-3, they are very powerful tools to find the final result of each entry. However the results are very limited to the existing input and to solve this limit and the misleading problems that can happen by a special tree, we have used the random forest classification in this thesis. Random forest uses several random trees and then makes a voting ensemble according to the output of each tree, therefore it results in a much higher accuracy.

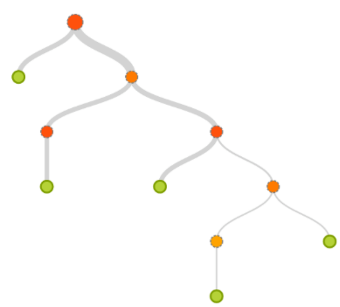

Figure 2-3. Demonstration of decision tree

2.2.4 Stochastic gradient descent (SGD)

Stochastic gradient descent (SGD) [13] is a commonly chosen algorithm for training machine learning models including SVM and Logistic regression. SGD is an approximation of the gradient descent optimization method for minimizing an objective function that is written as a sum of differentiable functions. In another words it is an optimization technique that reduces the errors by making random samples in each classifier and processing it. SGD is great for large amount of data and that is why we have used it in our methods.

2.3 Correlation coefficient

Correlation coefficient [14] is a quantitative measure that illustrates the relationship between two random set of data. As it has been explained in the Section 3, we have taken advantage of this and optimized the ensemble classifier by finding the classifiers with the lowest correlation coefficient comparing to random forest classifier.

Correlation coefficient is one of the main tools in machine learning to find the dependency of data with each other and it has different types; Pearson product, Intraclass correlation and Rank correlation are the most commonly used methods to find the correlation coefficient of two or more sets of data[15].

3. Methods

This experiment is largely divided into three large steps. First preparing, processing and filtering the data, then using 30 classification and optimization methods [15] to find the best match for ensemble classifier. Finally the last step is the ensemble classification. These steps have been explained in detail through the following subsections. The main methodology and process of this experiment has been depicted in following figure 3-1.

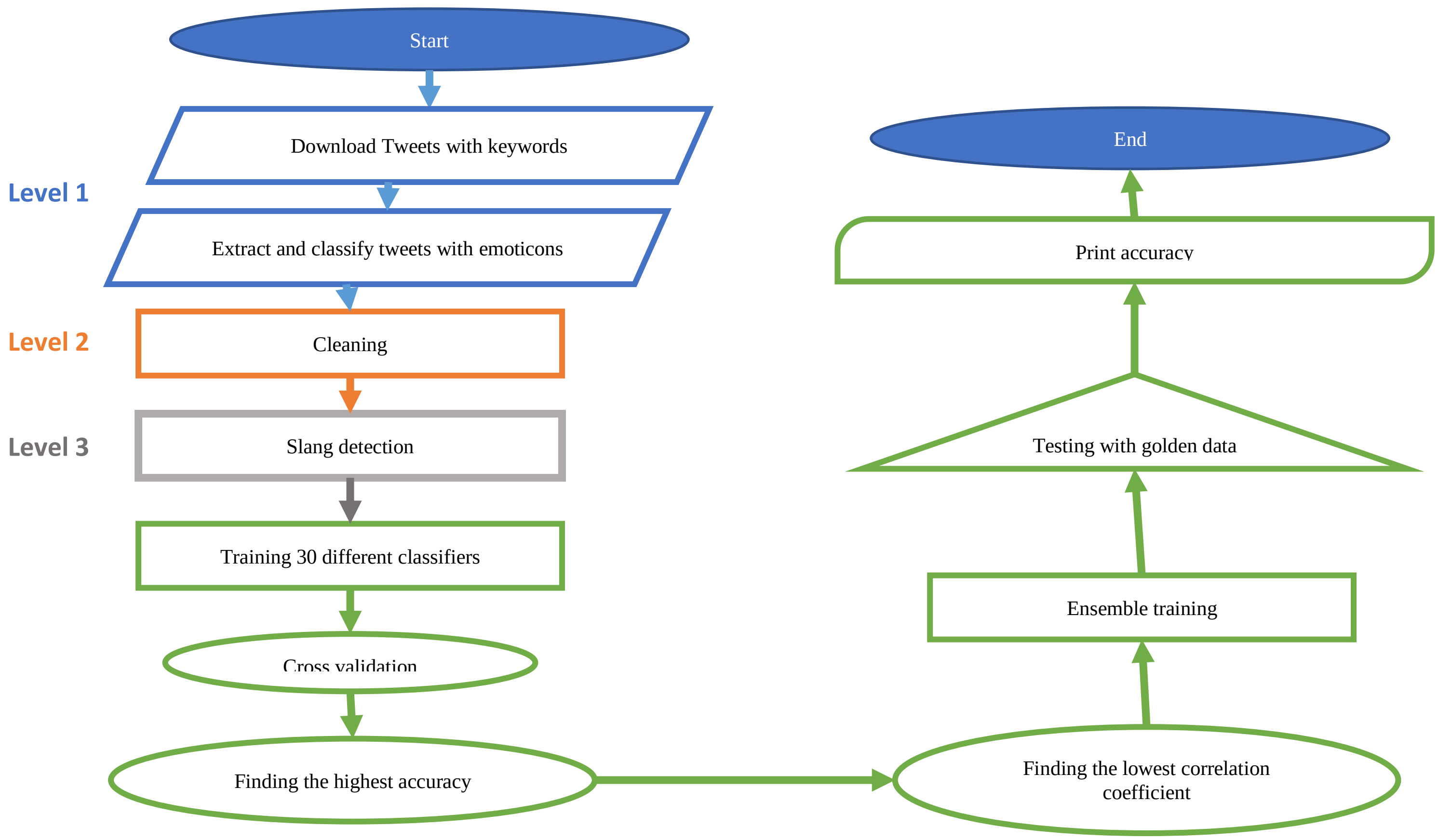

Figure 3-1. Methodology of the experiment

3.1 Data preparation & filtering

Tweets are mostly short and great for sentiment analysis. However there are a lot of problems that data scientists face with processing the data specially the data driven from social networking platform such as twitter. Through this project, we have addressed some of those problems and provided a solution for them.

First part of the project was extracting the tweets using Twitter API. We have chosen “Google, Microsoft, Apple and Tweeter” keywords to only find corporate related tweets. Through our experiment we have downloaded a very large amount (about 10 million) of tweets to apply out emoticon detection which was based on the experiments of Go and Pak [16, 17]. However we have applied a larger emoticon list to cover most of the related tweets. This list has been provided in the table 3-1. Finally we have chosen 1 million tweets with emoticons and applied the labeling according to emoticons.

Table 3-1. List of Emoticons

Positive	**Negative**
:-D	:-(
=D	=(
xD	:’(
(ˆ_ˆ)	- -
<3,	:'<
:->|	:-C
=:-)	<:(
;)	<:-(
(;	'Ω'
:)	:"<
:))	<:'-
:D	:[
;]	<:"(
:^)	:'(
=)	:-{
:-)	-.-
:]	=’(
: ɔ)	(>
;^)	:(
;-)	:-||
*)	>:(
:>	:|
:}	:-|
*-)	;(

After finding the tweets with emoticons and labeling them as "**positive**" or "**negative**" we have performed a 2 step cleaning process to make sure that all of our tweets are legit and they observed the effect of this cleanings on the final classification accuracy. First cleaning step is finding the urls and hashtags and other signs through the text that can reduce the accuracy. Second part is finding the slangs and misspellings through the tweets and replacing them with the correct version. Therefore through this thesis we call unfiltered step as "Level 1", cleaning step as "Level 2" and slang detection step as "Level 3". Also to observe the effect this filtering in the classification and accuracy of the total output we have chosen the blue color for Level 1, orange color for Level 2 and gray color for Level 3.

In the Level 2 we have removed every sign and tags. Therefore we remain with a simple text with no emoticon or question mark or etc. Also in this step we have removed every username and tweet referrals and urls that was provided to pictures and videos. The results of this cleaning has been discussed in the Section 4.

In the Level 3 we have performed a word by word filtering strategy to observe the slangs and misspellings. Tweets are usually short texts and because there is a limited number of characters, people try to abbreviate their words and common dictation mistakes are prevalent. Therefore first we make a bag of words and then iterate through each of them and check if it is a misspelling or not. This part has been managed using Cambridge dictionary API. If the word is misspelled then we replace it with the correct form and if not then we sent it to the next step. Next step is finding weather if the word is a slang or not. We have used a data crawling technique to find the meaning of each word in Urban dictionary. Urban dictionary is a open source dictionary completed by people and is consisted of most of the common slangs that people use. Therefore if the word exists in the Urban dictionary, then we replace it with the right form if not we just remove the word. For example if we have a tweet "Let's meet tmr at librari" then our system replaces "tmr" with "tomorrow" and "librari" with "library". The methodology of our process has been depicted in figure 3-2.

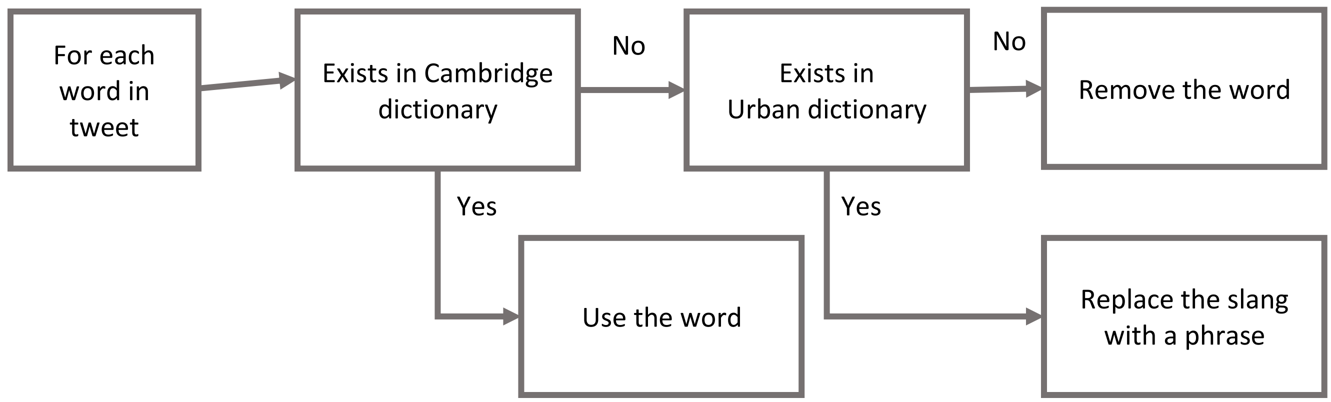

Figure 3-2. Slang detection methodology

3.2 Classification & comparison

This part is the main processing part of this project. We have trained 30 classifiers and optimization techniques to find the best and the most accurate results. We have used the 1 million labeled and filtered tweets and evaluated them through cross validation test. Final results for each Level 1, 2 and 3 showed that random forest classifier has the highest accuracy therefore we chose that as the main classifier and through the next step we have tried to find the optimized selection for ensemble classifier. The results for each classifier is depicted in figure 4-1.

3.3 Ensemble classification

There are many ensemble classifiers available for the machine learning techniques such as voting classifier and Adaboost. We have chosen voting classification because the random forest was showing a great amount of accuracy and with a little bit of correction we could achieve the best results through it. On the other hand Adaboost classifier would have been a good choice but because of time constrains running the random forest classifier would have taken very long time and a large amount of recourse therefore voting classifier was the best choice.

In order to find the best match for ensemble voting classifier we need to use the classifiers with the lowest correlation coefficient. This means that although each of those classifiers have a high accuracy but they are showing different result for each case, which we can choose the most probable one through a voting process. However the random forest classifier had the highest score through the cross validation test and different classifiers seemed to have a different performance. Therefore we have given a weight of their score to each classifier in the voting. This means that if random classifier has a score of 90% accuracy then we gave 9 point on its weight and lower to the ones with lower accuracy. Table 3-2 shows the correlation coefficient of chosen classifiers towards the random forest classifier. Through the 30 trained classifiers chosen the ones that had a correlation coefficient between -3 to +3 to find the optimum points. Final results shows that only 8 classifiers can fit in that range and those are the one we have used in our experiment to build the ensemble model.

Table 3-2. Optimization methods with lowest correlation coefficient

Classifier Name	**Correlation coefficient**
Nearest Centroid	0.28
Perceptron	0.24
SGD_perceptron_l2	0.25
SGD_perceptron_elasticnet	0.24
SGD_squared_hinge_l2	- 0.04
SGD_squared_loss_l2	0.01
SGD_squared_loss_l1	- 0.01
SGD_squared_loss_elasticnet	- 0.03

4. Results & Discussion

First we have tested the impact of cleaning and filtering the data on the systems accuracy. This test was conducted by using a 1000 hand labeled tweets (golden data) by four different people. Using their output we only have selected the tweets that were labeled with the same sentiment. The results show that there is a very significant increase in the accuracy through all of the filtering levels. According to table 4-1 before trimming and filtering the data we only had a 63% of accuracy through the random forest classifier which has been increased to 73% in Level 2 and 78% in Level 3. Therefore we can conclude that our cleaning techniques were efficient and it has resulted in a 14% increase in the accuracy of the whole system.

Table 4-1. Accuracy results comparison

Level	Highest Classifier Accuracy	Ensemble Accuracy
Level 1. Before trimming	63%	69%
Level 2. After Cleaning tags	73%	76%
Level 3. After Slang Detection	78%	83%

Considering the effect of ensemble classification, in the table 4-1 we can observe an average 6% increase in the accuracy for each Level. This increase is even more visible in the figure 4-1 which shows the accuracy of each classifier in different Levels. In figure 4-1 we can confirm that random classifier was the best choice as the target classifier as it has the highest accuracy between the other classifiers. Figure 4-1 also shows that our ensemble approach was effective and it has a higher result than any classifier in all 3 Levels. Among 30 classification and optimization methods that was used in this experiment, we have taken random forest classifier as the main control group.

Through this experiment we can observe the effect of choosing a homogenous and related category of data. Since there are a limited bag of words related to corporates, we can observe a very accurate output even when testing with golden data. The main feature of this data set was not having unwanted words, such as usernames and hashtags, and limiting the words to a certain field. Therefore by limiting the words to a specific field, and finding the replacement for slangs we have decreased unwanted words and increased the accuracy of system.

However in the future experiments we can use other classifiers such as Decision tree or SVM and observe the behavior of ensemble classifier. We expect a similar accuracy but due to the less time consumption in those classification methods, we can save on resources and calculation time which will increase the total efficiency of the system.

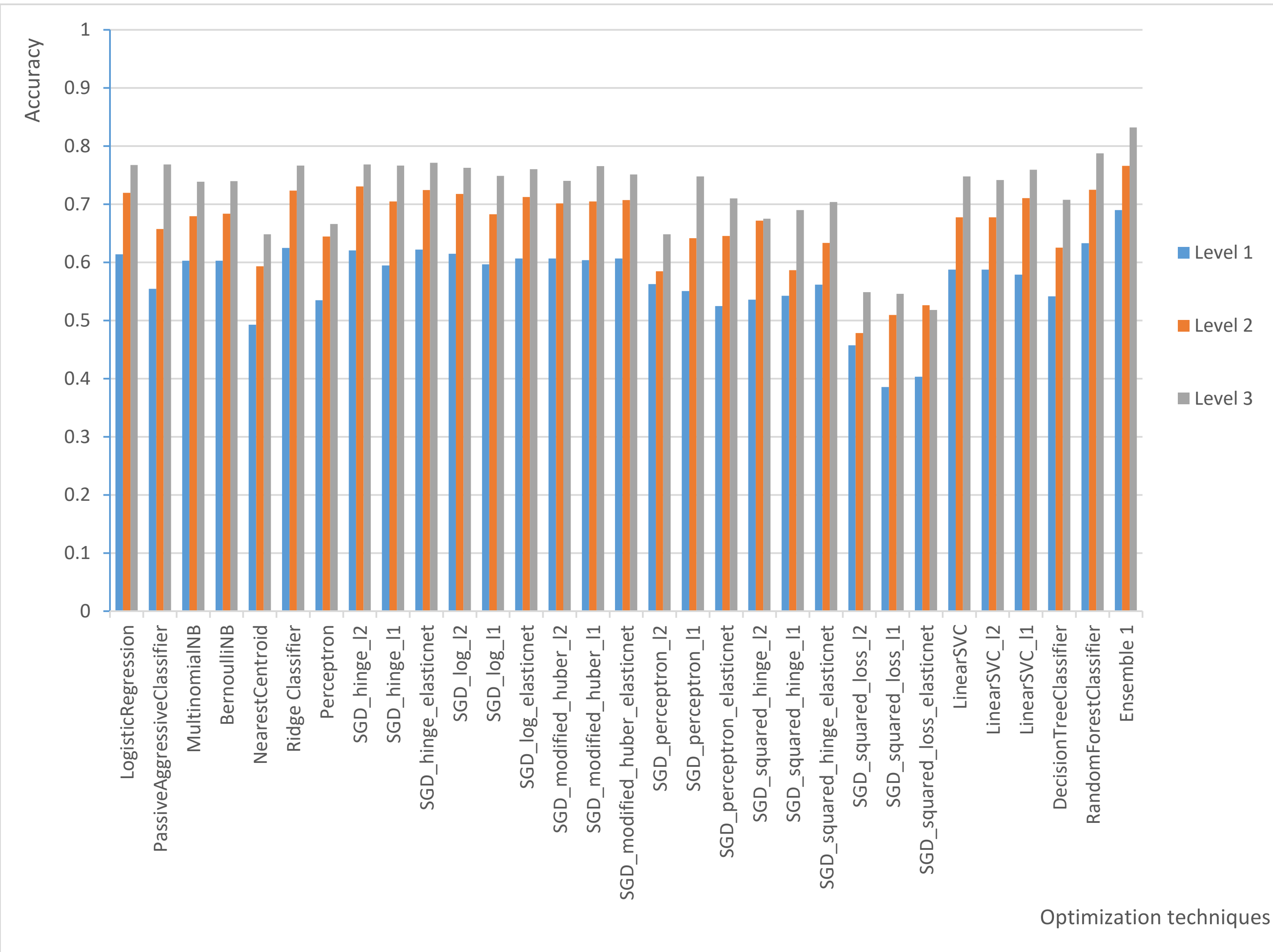

Figure 4-1. Accuracy results for each optimization technique

However the results of this experiment are not only limited to the performance of the system but it also has increased the efficiency of it. Table 4-2 shows that the training time has reduced by 10% which is a saving on resource and gives a faster result.

Table 4-2. Training time for Ensemble classifier

Level 1	Level 2	Level 3
15 Hours	14.5 Hours	13.2 Hours

Using the filtering techniques we have reduced some unwanted words from the processing trail, therefore there are less but more accurately labeled bag of words in each level to train. This reduction in the amount of training data has effected the total training time and reduced the required resources and processing time.

5. Future works

Through this experiment we have tried to implement new methods and the most possible options to observe every kind of possible scenario. However we believe that there are still subjects for improvement in the future works in which we will mention in this section.

Firstly we have chosen English as the main language of this experiment. The main reason we chose English was that our target companies were all placed in an English speaking country therefore most of the valid reviews about those companies were in English. Also English as a global language had a great impact in our selection. However there are many different languages in the world that could have been selected in the future works such as Korean or Japanese. We believe that using those languages would have the best outcome when we are targeting their local companies therefore in the future works we would be targeting Koreans companies such as Samsung and LG and adopt Korean language as the main language.

Through this experiment we have adopted 30 different classifier and optimization techniques. However still there is a room to advance and test the results on the other classifiers and adopting other optimization techniques. This can be more visible when using different kernels for SVM such as polynomial, RBF and etc. Also in ensemble approach we have only tested the voting ensemble classifier, and in the future works we will mainly target Adaboost to observe the changes in the accuracy and efficiency of system.

Sentiment analysis needs a very sensitive approach and we need to have a more advanced studies in our classification method in the future works. Therefore in the future works we will use a 3 classification method of positive, negative and neutral to increase the accuracy while providing more information about each tweet.

Finally as it was mentioned in the Section 4, in the future works we will use different classification methods as the control group to observe the training time reduction and the change in the systems accuracy.

6. Conclusion

Analyzing tweets sentiment brings a promising information for the market and product analyzing purposes but it requires a high amount of accuracy. Through this experiment we have proved that if we narrow the topic to a specific model we can achieve an accurate output. So we have introduced a three step data filtering process to download the data and even detect misspellings and slangs through the tweets. Finally we have proposed an ensemble approach and observed an 83% accuracy which was 14% increase comparing to the scenario of not using this method. Our system is optimized for corporate related tweets and it be used to provide an accurate daily sentiment analysis of a target company. This will help companies to monitor their popularity on Twitter and taking related marketing strategies accordingly. In the future we will broaden the language of this system to languages such as Korean and Japanese to provide this tool to more international corporates.

References

[1] T.Wilson *et al.*, "Recognizing contextual polarity in phrase-level sentiment analysis", Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pp. 347–354, 2005.

[2] D.Davidov *et al*., "Enhanced sentiment learning using Twitter hashtags and smileys", Proceedings of Coling 2010, 2010.

[3] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", Proceedings of LREC, 2010.

[4] R. Parikh and M. Movassate, "Sentiment Analysis of UserGenerated Twitter Updates using Various Classification Techniques", CS224N Final Report, 2009.

[5] A. Go, R. Bhayani, L.Huang. "Twitter Sentiment Classification Using Distant Supervision". CS224N Project Report, Stanford University, 2009.

[6] J. Read. "Using emoticons to reduce dependency in machine learning techniques for sentiment classification", Proceedings of the ACL student research workshop. Association for Computational Linguistics, 2005.

[7] B. Pang *et al*., "Sentiment classification using machine learning techniques", 2002.

[8] A. Hassan *et al.*, "A Bootstrap Ensemble Framework", Social Computing (SocialCom), 2013 International Conference on. IEEE, 2013.

[9] L.Wang, "Support Vector Machines: Theory and Applications", Springer & Business Media, 2005.

[10] Y.Mack, "K-Nearest neighbor estimation", University of California, 1978.

[11] C. Shahabi, “A Road Network Embedding Technique for K-Nearest Neighbor”, Proceeding of the 10th ACM Workshop on Advances on Geographic information system, 2002.

[12] F.Brodley *et al.,* “Decision tree classification of land cover from remotely sensed data. Remote Sensing of Environment”, 61, pp. 399–409, 1997.

[13] L.Bottou, “Large-Scale Machine Learning with Stochastic Gradient Descent”, In International Conference on Computational Statistics, pp. 177-185, 2010.

[14] J.Rodgers, “Thirteen ways to look at the correlation coefficient”, The American Statistician, Taylor & Francis, pp. 59-66, 1988.

[15] B. Lars *et al.*, "API design for machine learning software: experiences in scikit-learn project”, 2013.

[16] A. Go et al., “Twitter sentiment analysis,” CS224N Project Report, Stanford University, 2009.

[17] A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, LREc. Vol. 10, 2010.

초록

소셜 네트워크 서비스가 (SNS) 우리 사회의 불가분한 요소가 되었으며 이를 통해 매일 생산되는 대량 데이타를 사용해 마켓 분석과 금융분석 등을 할 수 있다. 감정분석은 데이타 분석 도구중에서 하나이며 이 도구를 사용해 사용자의 감정을 텍스트를 통해 얻어낼 수 있다. 감정 분석 시스템의 정확도가 높을 수록 더 믿을 만한 마켓 분석 결과를 얻을 수 있다.이 논문은 정확한 기업 관련된 트위트 감정분석을 위한 시스템을 제시하며 이 시스템은 기업에 대한 대중의 사고를 측정할 수 있다. 이 시스템은 안상불 학습을 기반하여 여러 데이타 필터링을 사용하고 있다. 데이타 필터링 방법 중에서 속어와 은어를 찾는 다는 것을 말 할 수 있다. 이 연구를 통해 제시한 필터링 방법으로 시스템의 정확도가 증가 됬음을 확인했다. 그다음에 안상불 학습을 위한 여러 텍스트 분류 방법을 사용해 최적한 모델을 찾았다. 테스를 통해 제시한 시스템의 정확도를 측정하여 결과는 기존에 있는 다른 방법 보다 높게 나왔다. 이 시스템을 통해 목표 회사에 대한 트위트 감정 분석 또는 인기도 분석까지 실행 할 수 있다.

주요어: 감정 분석, 앙상블 학습