I J C T A, 9(26) 2016, pp. 343-349
© International Science Press

Smartphone Product Review Sentiment Analysis using Logistic Regression

M. Lovelin Ponn Felciah[1] and R. Anbuselvi[2]

ABSTRACT

Sentiment analysis or opinion mining is one of the vital tasks of Natural Language Processing. The task of automatically retrieving opinions from product reviews is gaining much attention among web mining community. What other people think has become an important piece of information for the customers as well for the decision makers. It makes the researchers to research in this area. This paper performs sentiment analysis on product reviews and focus on the classification of polarity from the customer reviews taken from online for smartphones. The logistic regression is used to find the classify sentiment of the product reviews. For experiments, we have used Samsung mobile phone reviews and evaluated how far the mobile phone users satisfied with the product they bought.

Keywords: Sentiment Analysis, Polarity Classification, Product review, Logistic regression, Opinion mining.

1. INTRODUCTION

The web has greatly change the way that people express themselves and express their views with each other. Opinion are the core for almost all human activities because they influence the human behavior. Social media is a huge resource where there are large numbers of people's review for all types of products and services. It becomes common practice for a customer to know and learn the likes and dislikes of products before buying, even it is very much useful for the manufactures and decision makers to keep track of customer opinion on its product to improve the customer satisfaction as [1]. Each site contains huge volume of opinionated text that is not always easily analyzed in long blogs and forum posting, which will be difficult to summarize, extract the opinion from them.

Sentiment analysis which is also known as opinion mining [2], studies people's thought, judgment views towards certain entities. In past few years there was data in mining opinion in reviews from academic as well as in industry. The opinionated text social media have helped to reshape the business and political system which impacted on the social media. The vital task in sentiment analysis is classifying the polarity of a given text data at the document level, sentence level, phrase or aspect level whether the expressed opinion is positive, negative and neutral. There are basically two types of sentiment opinion and facts. Facts are called as objective expression about events.

Opinion is usually called as subjective expression. With the rapid development of Smart Phones, the end platform has been changed gradually from PC to mobile phones. It's more flexible for users to comment through phones. As predicted, there will be 5.6 billion smart phone users by 2019, which will produce 10 Exabyte (1018 bits) data stream [3]. As the data grows it is difficult to find the exact view from the text and the problem becomes much more complex as

[1] Research Scholar, Department of Computer Science, Bishop Heber College (Autonomous), Tiruchirapalli-620017. *E-mail: lovelinsathya@gmail.com*

[2] Assistant Professor, Department of Computer Science, Bishop Heber College (Autonomous), Tiruchirapalli-620017. *E-mail: r.anbuselvi@gmail.com.*

- A positive and negative sentiment words may have different meaning in different domain
- The sentence contain sentiment word may not express any sentiment
- The sentence may not contain sentiment word but gives opinion.
- Some post may contain irony and sarcasm words are hard summarize the opinion.

In our paper, we study how reviews can be used for product sales. We collected a corpus of 1000 smart phone reviewers evenly spilt manually between two sets. Text which contain positive and negative review of the smart phone.

The contribution of our paper are follows

1. The corpus collected with positive and negative sentiment, a corpus of objective texts. The size of the collected corpora can be arbitrarily large.
2. We use the collected corpora to build a sentiment classification system using logistic regression analysis.
3. We conducted the experimental evaluations on a real set of reviews collected for smartphones.

The paper was organized as follows. In section 2 we discuss the prior work on opinion mining and sentiment analysis and their applications on social media. In section 3, we describe the process of collecting the corpora and the logistic regression in which the collected data is examined. In section 4 the analysis of corpora using logistic regression and the experimental evaluation in section 5. Finally the conclusion about the work in this paper.

2. RELATED WORK

With the dramatic growth in web and social media the sentiment analysis and opinion mining became a field of interest for many user. The board view of the existing sentiment analysis work was presented in (Pang and Lee, 2008). Alexander pak *et. al.* [4] proposed a technique that perform better than the previously proposed method. Using the twitter API they collected the corpus of text posts and formed a dataset that are classified as positive, negative and neutral. The collected corpus trained for a classifier to recognize positive and negative text. Distribution of word frequencies are checked in the collected corpus using zipf‘s law. The collected corpus that train the classifier using tree tagger for POS tagging, the accuracy is better than the previous method

Our work is closely related to [4], where the authors propose a sentence-based analysis, in which it uses association rule mining to extract the most frequent features. The underlying principle of this method is to that, product features will occur most frequently when compared to other words or word phrases in a user reviews of a product (co-occurrence based approach). But, association rule mining has some significant drawback and challenge.

Go *et. al.*, [5] evaluated the Tweets to collect training data and then to perform a sentiment search. The authors construct corpora by using emoticons to obtain “positive” and “negative” samples, and then use various classifiers. The Naive Bayes classifier with a mutual information measure for feature selection. The examine result obtain up to 81% of accuracy on their test set. However, the method showed a bad performance with three classes.

Phillips *et. al.*, [6] examined and evaluate that the logistic regression offers easy interpretability classification process and to ramification of potential misclassification by comparative study based on predictive performance of logistic regression, ANN and SVM. The performance of LR, CCNN, and SVM models have been evaluated for classification on oil samples from mining trucks. The study evaluated that the LR model demonstrated an ability to classify approximately 89% of the oil sample compared with the other.

Felipe *et. al.* [7], analysed the opinion by the combinations of sentiment dimensions provides significant improvement in twitter sentiment classification task such as polarity subjectivity. The objective of this article is to improve two major sentiment classifier, subjectivity and polarity classification. The dimension on their study is strength, emoticons and polarity various lexicon approach were used to calculate the number of features according to the number of matches between the words from the tweets and the words from the lexicon. The sanders data set is used with 10 fold-cross validation in learning algorithm among the SVM, Naive Bayes and logistic regression perform well.

Nit in *et. al.*, [8] investigated the opinion spam based on the analysis of 5.8 million reviews and 2.14 million reviewers, used logistic regression that produces a probability estimate of each review being a spam which is desirable. Techniques used to determine different spam and the result that the logistic regression model is highly effective.

3. SENTIMENT ANALYSIS USING LOGISTIC REGRESSION

This section presents the product review sentiment analysis steps using logistic regression.

3.1 Data annotation and Analysis

The latest research area in opinion mining are sentiment analysis and feature based opinion mining. Sentiment classification analysis each review and classify them as positive or negative. [9, 10] uses a document level sentiment classification. Our work concentrate only in reviews of product which contains feature and opinion pairs rather than the whole review and we perform classification at sentence level. Subjective sentence level classification is studied in [11], which finds whether a sentence is an objective sentence or a subjective one. Sentence level opinion or sentiment categorization is studied in [10, 11]. A review can have many features, and each can have their own opinion polarity. *e.g.*, “picture quality of this phone is awesome and so is the Touch, but the battery life is short.” “Picture Quality”, “Touch” and “Battery life” are featured. The opinion “awesome” on “Picture Quality, “Touch” is positive, and the opinion “short” on “Battery Life” is negative. Before the corpus was classified the data were pre-processed as follows.

3.2 Pre-processing

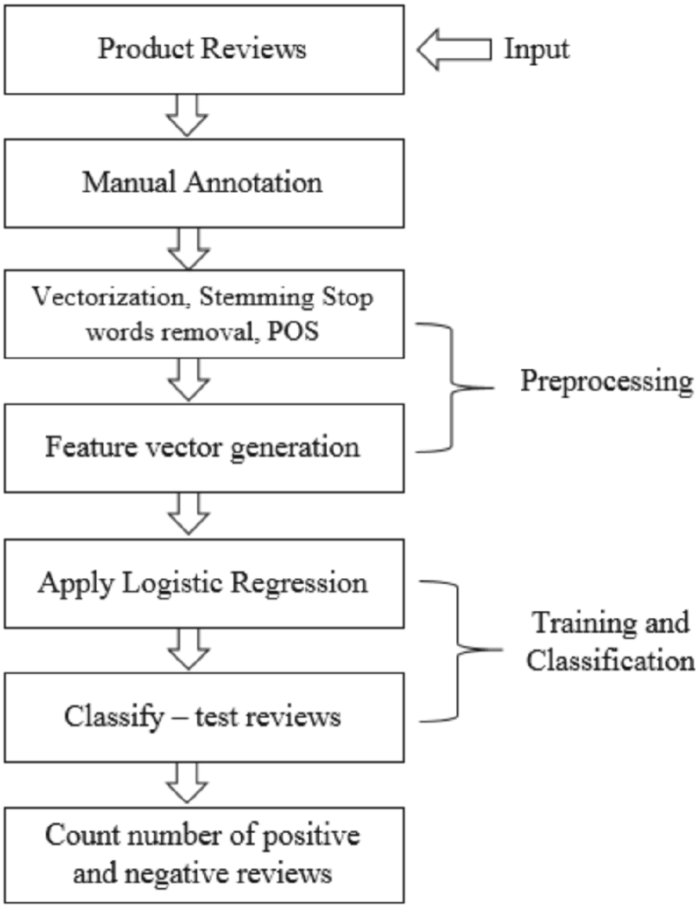
In this section, the following pre-process steps used for analysis the product reviews. There are some basics step involves before classifying the data sets like removing the html tags, non-textual information which is not required for sentiment analysis. The following are some of the text pre-processing steps .

3.2.1. Stemming

Stemming is the process of reducing words to their root that is to its stem, and it can be viewed as a recall-enhancing device or a precision enhancing. A stemmer reduces the number of redundant terms while increasing the matching probability for document comparison.

3.2.2 Part of Speech Tagging

The part-of-speech tagging also called as grammatical tagging, is the process of marking a word in a text as corresponding to a particular part of speech, based on both its definition, as well as its context. It has been found that certain parts of speech such as adjectives and adverbs express polarity more often.


Product Reviews
Input
Manual Annotation
Vectorization, Stemming Stop words removal, POS
Preprocessing
Feature vector generation
Apply Logistic Regression
Training and Classification
Classify – test reviews
Count number of positive and negative reviews

3.2.3 Stop Words

Stop words are used to remove the non-semantic words like articles, prepositions, conjunctions and pronouns. In computing, stop words are words, which are filtered out prior to, or after, processing of natural language data and it is used as one of the pre-processing step to get rid of such words because they do not hold any information. They basically just confuse the classifier and introduce problems.

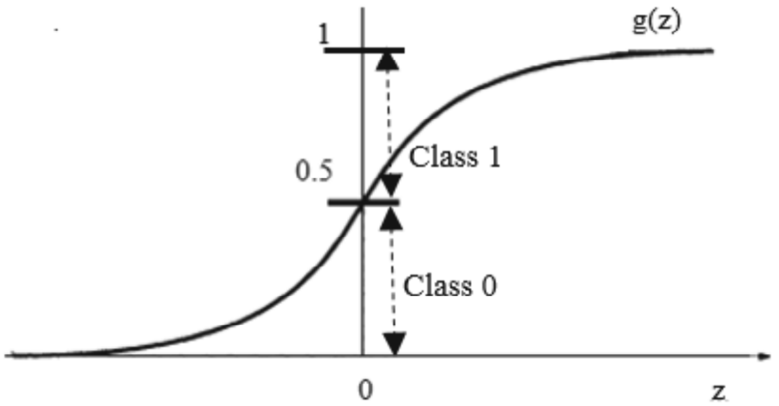
3.2.4 Tokenization

Tokenization is splitting up the systems of text into personal terms or tokens. This procedure can take many types, with regards to the terminology being examined. For English, an uncomplicated and effective tokenization technique is to use white space and punctuation as token delimiters.

3.2.5 Feature vector generation

The 87 most representative features has taken to form a feature vector from the review corpus data. The 87 term features are as follows:

[‘app’, ‘applic’, ‘awesom’, ‘bad’, ‘batter’, ‘batteri’, ‘becaus’, ‘best’, ‘better’, ‘biggest’, ‘black’, ‘bought’, ‘buy’, ‘camera’, ‘charg’, ‘day’, ‘doe’, ‘drain’, ‘dure’, ‘earphon’, ‘edg’, ‘end’, ‘everyth’, ‘face’, ‘fast’, ‘featur’, ‘finger’, ‘free’, ‘galaxi’, ‘gb’, ‘good’, ‘great’, ‘hang’, ‘hardwar’, ‘heat’, ‘iphon’, ‘iss’, ‘just’, ‘life’, ‘like’, ‘long’, ‘look’, ‘lot’, ‘mani’, ‘marshmallow’, ‘mobil’, ‘month’, ‘need’, ‘nice’, ‘onli’, ‘overal’, ‘perform’, ‘phone’, ‘pictur’, ‘powerbank’, ‘practic’, ‘price’, ‘problem’, ‘problemat’, ‘ram’, ‘realli’, ‘run’, ‘samsung’, ‘screen’, ‘simpli’, ‘smartphon’, ‘sometim’, ‘soni’, ‘space’, ‘super’, ‘switch’, ‘tell’, ‘thing’, ‘time’, ‘twice’, ‘’, ‘unlucki’, ‘updat’, ‘use’, ‘veri’, ‘video’, ‘want’, ‘warranti’, ‘went’, ‘work’, ‘worst’, ‘wrong’, ‘year’]



These features are then used to compute feature vectors of all product reviews then they are fed to logistic regression model.

4. LOGISTIC REGRESSION

4.1 Logistic Function

The logistic function or sigmoid function is the heart of the logistic regression technique and is defined as follows:

$$g(z) = \frac{1}{1+e^{-z}}$$

Here, *e* is the numerical constant *i.e.*, Euler's number and z is an input to the logistic function. The figure 1 depicts the logistic function or sigmoid function.

4.2 Logistic Regression Model

The logistic regression is used to fit the prediction model and classified the product reviews into either positive or negative opinion. Here the probability that is greater than 0.5 is treated as positive and the probability that is less than or equal to 0.5 are treated as negative opinion.

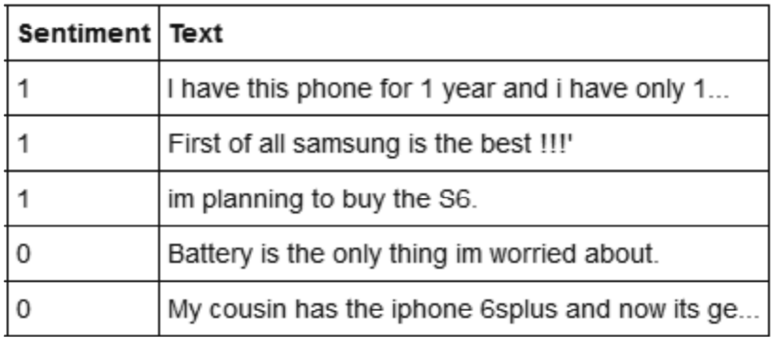
In sentiment analysis polarity of many words is domain and context specific. For example the word "long" is considered positive in "long battery life", however negative in "long shutter lag". Excluding such expressions may lead to poor coverage while tagging them with overall polarity tendency may lead to poor precision. J. Fang and B.Chen [12] showed that identifying domain specific lexicons lead to significant improvement. For each sentence in the review we would try to identify the aspect of the product discussed and then the associated sentiment. The block diagram of the proposed approach is shown in figure 2.

The training corpus used is such that each review is annotated with an aspect as well as associated sentiment with logistic regression. Logistic Regression belongs to the family of classifier known as the exponential or log-linear classifier. Like Naive Bayes, it work by extracting some set of weighted features from the input [13], taking logs and combining them linearly. Technically, logistic regression refers to a classifier that classifies an observation into one of the two classes. The most important difference between naive Bayes and logistic regression is that logistic regression is a discriminative classifier while Naive Bayes is a generative classifier.

5. EXPERIMENTAL RESULTS

In this section, we present the experimental results of the proposed mobile phone review analysis approach.

Table 1
Sample reviews from collected dataset



Sentiment	Text
1	I have this phone for 1 year and i have only 1...
1	First of all samsung is the best !!!'
1	im planning to buy the S6.
0	Battery is the only thing im worried about.
0	My cousin has the iphone 6splus and now its ge...

5.1 Dataset

We collected reviews from the web site: http://www.gsmarena.com for Samsung mobile phone models S5, S6 and S7. It contains 1007, 740 and 1214 reviews respectively. These reviews are manually annotated either as positive or negative review. Sample review annotations are given below:

5.2 Results

We used the collect product reviews and applied the sentiment analysis using logistic regression as explained in section 4. The performance of the logistic regression on Samsung review data is shown in table 2, where precision and recall for sentiment positive (1) and negative (0) are given separately.

Table 2
Logistic Regression performance on dataset

	Precision	*Recall*
0	0.40	0.67
1	0.67	0.40

The consumer opinion on Samsung products are given in table 3. From the table 3, it is understood that 77.06 %, 56.48 % and 68.94 % of users are satisfied on Samsung S5, S6 and S7 mobile models respectively.

Table 3
Product sentiments

Sentiment	*Number of reviews*	*Positive*	*Negative*
Samsung S5	1007	776	231
Samsung S6	740	418	322
Samsung S7	1214	837	377

CONCLUSION

In this paper, we proposed a smart phone review sentiment analysis approach to find out which product got high reputation among customers. We have used Samsung mobile reviews as our dataset and used the logistic regression to find out Samsung mobile phone users' opinion. This approach only give overall opinion of consumers to use the mobile phone. In future, we will perform sentiment analysis considering mobile phone features such as screen resolution, camera, battery, price etc.

REFERENCES

[1] E Kushal Bafna, Durga Toshniwal, Feature Based Summarization of Customers' Reviews of Online Products, Published *by Elsevier B.V.Selection and peer-review under responsibility of KES International*. 1877-0509 © 2013.

[2] Pang, B., and Lee, L,Opinion mining and sentiment analysis, *Foundations and Trends in Information Retrieval*, 2:1-135. 2008.

[3] Lin Zhang ,Kun Hua, Hong gang Wang,Guanqun Qiane,Li Zhang, Sentiment Analysis on Reviews of Mobile Users, *The 11th International Conference on Mobile Systems and Pervasive Computing* (MobiSPC-2014).

[4] Pak, A., and Paroubek, P, Twitter as a corpus for sentiment analysis and opinion mining. *In Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC'10). Valletta, Malta, 2010.

[5] Alec Go, Lei Huang, and Richa Bhayani, Twitter sentiment analysis. Final Projects from CS224N for Spring 2008/2009 at, *the Stanford Natural Language Processing Group.*

[6] J. Phillipsa, E.Cripps, JohnW.Lau, M.R.Hodkiewicz, Classifying machinery condition using oil samples and binary logistic regression, *http://dx.doi.org/10.1016/j.ymssp.2014.12.020 0888-3270/& 2015ElsevierLtd*

[7] Felipe Bravo-Marquez, Marcelo Mendoza, Barbara Poblete, Combining Strengths, Emotions and Polarities for Boosting Twitter Sentiment Analysis, WISDOM'13, August 11 2013, Chicago, *USA.Copyright 2013 ACM 978-1-4503-2332-1/13/08.*

[8] Nitin Jindal and Bing Liu, Opinion Spam and Analysis, *WSDM'08*, February 11-12, 2008, Palo Alto, California, USA.Copyright 2008 *ACM 978-1-59593-927-9/08/0002.*

[9] E. Riloff and J. Wiebe. 2003. Learning extraction patterns for subjective expressions. *EMNLP'2003, 2003.*

[10] M. Gamon, A. Aue, S. Corston-Oliver, and E. K. Ringger. Pulse: Mining customer opinions from free text. IDA'2005.

[11] V. Hatzivassiloglou and J. Wiebe, Effects of adjective orientation and gradability on sentence subjectivity. *COLING'00, 2000.*

[12] Ji fang, Bi Chen. Incorporating Lexicon Knowledge into SVM Learning to Improve Sentiment Classification. *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011*, pages 94–100, Chiang Mai, Thailand, November 13, 2011.

[13] A-M. Popescu and O. Etzioni. Extracting Product Features and Opinions from Reviews. EMNLP-05, 2005.

[14] B. Liu. *Web Data Mining: Exploring hyperlinks, contents and usage data*. Springer, 2007.

[15] Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, Target-dependent twitter sentiment classification. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:* Human Language Technologies, volume 1, pages 151–160, 2011.