International Journal of Pure and Applied Mathematics
Volume 117 No. 21 2017, 87-92
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

SENTIMENTAL ANALYSIS OF SOCIAL NETWORK SITES FOR CATEGORIZATION OF PRODUCT REVIEWS

Sasikumar A.N [1], Joshila Grace L.K [2]

[1] P.G Student, Department of Computer Science and Engineering, Sathyabama University, Chennai, India.
[2]Assistant Professor, Department of Computer Science and Engineering, Sathyabama University, Chennai, India.

Abstract : Sentiment analysis is concerned with identifying and grouping assumptions articulated in a piece of text to identify whether the user's view towards a particular subject, merchandise, etc. is optimistic, pessimistic, or impartial. Opinions about products influence buying behavior among users since millions of them visit micro blogging websites. In this paper, we propose a scheme based on assortative model to analyze the descriptive reviews about products and to categorize sentiments based on the polarization. We also analyze the classification methods best suited for sentimental categorization. Tweets from twitter are retrieved and experiments are done to evaluate the efficiency of classification methods.

Keywords: Sentiment analysis, Assortative Model, Sentiment Polarization,

1. INTRODUCTION

Sentiment is a stance, reflection, or conclusion encouraged by emotion. Sentiment investigation also referred opinion mining, studies people's beliefs towards a product. Micro blogging and social networking websites provide rich source of information about every possible products people use on a day to day basis. People are willing to share their view about products that give them satisfaction and induce them to share their opinions through various social networking sites such as Flipkart, Twitter and etc. The assortative model focuses on social well being attributes (SWB) such as age, sex, education, no of followers of twitter users to determine whether any kind of homophily exists between them. Happiness factor greatly influences the positive sentiments towards a product or topic.

In this paper, we propose a systematic approach to analyze twitter data. 1) To identify the flaws in the twitter data that potentially deters the development of sentiment analysis. The primary fault is quality of content since users enjoy certain degree of freedom to liberally place their own views hence value of their opinions cannot be assured. We have to identify irrelevant content and spam. 2) Social well being attributes (SWB) are highly influential in determining the polarity of the opinions. 3) A classifier to analyze the sentiments. Data from Twitter extracted using REST API, Search API. We also retrieved

Data from Flipkart.com to understand the similarity between how reviews are posted in twitter and online merchant websites. WordNet version 2.1 is used as a sense warehouse and each word is mapped according Parts-of-Speech Tags and SentiWordNet assigns the sentiment scores. Finally, Weka tool is used to conduct experiments.

This paper addresses a essential difficulty of sentiment analysis, called polarity categorization using assortative model properties. We propose a three phase course of action. In phase 1, tweets are retrieved using APIs. In Phase 2: i) An algorithm is proposed and implemented to identify homophily among users using assortative model; ii) A mathematical approach is proposed to calculate similarity index; In Phase 3: i) Polarity categorization experiments based on SWB characteristic are carried out; ii) Performance analysis of classification methods for assortative model and comparison of results.

2. RELATED WORK

Our work first addresses the basic issue in sentiment analysis: the categorization of polarity - a piece of text, should be categorized into either Positive or negative or neutral. Next important issue is dividing the text into three levels: document, sentence, and aspect level[1]. Xing Fang and Justin Zhan [1] suggested a mathematical model to identify the polarity categorization. Hu and Liu [2] research work focuses on

words that describe positive, negative and neutral words from reviews. Gann et al[3] list out tokens calculated from Twitter review data. A token called a sentiment score TSI (Total Sentiment Index as:

$$TSI = \frac{p - \frac{tp}{tn} * n}{p + \frac{tp}{tn} * n}$$

where p is the number of occurrences of a token in positive tweets and n is the number occurrences of a token in negative tweets. tp/tn is the ratio of total number of positive tweets over total number of negative tweets. Johan Bollen et. al [4] research focuses on creation of Friend Network based on Follower relationships which excludes the occasional users. Lot of research[5-12] has been done on various aggregate latent characteristics of twitter users.

Data collection

We retrieved product reviews from flipkart.com from November 2016 to march 2017 and nearly 4 million tweets were retrieved. From these enormous amounts of tweets only those tweets that contained subjective content were retrieved and tokenized. The friend network model[4] is used as base model for identifying the assortative characteristics of Users. The status information from twitter is useful to identifying them.

Twitter supports GET method and json APIs (https://api.twitter.com/1/friends.json?screen_name=@sasiprof) to track the follower lists. Twitter also supports POST method to retweet. A twitter friend network can be constructed using tweepy a python library.

Preprocessing

In this stage, NLP methods are applied to create tokens and to remove of stop words, special symbols (@,#), emoticons. Finally, Stemming of words[9] reduces the dimensions of a word.

Assortative Model

The assortative characteristics of users viz. age, gender can be used to identify most productive tweets. In this research, we propose a assortative model to classify tweets to determine its usefulness.

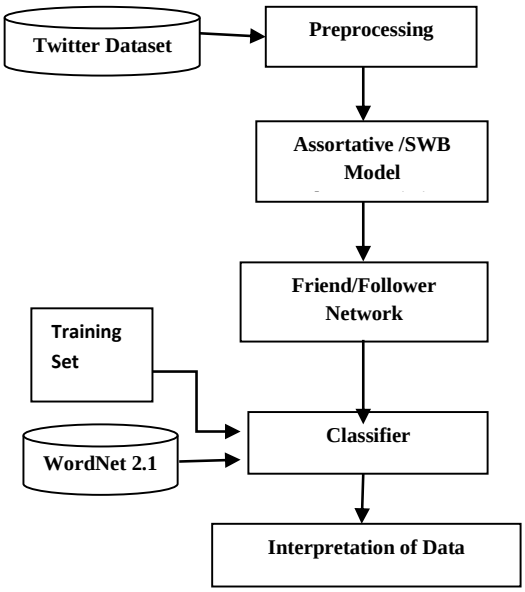


Fig. 1: Architecture of Assortative Model

The Friend-Follower network[4] to identify most productive tweets and to avoid users who occasionally tweet. The assortative model extends this network to assign a weight for each tweet based of their polarity. Let Wi,j be the degree to which users ui,uj have similar sets of friends. The weight Wi,j can be expressed as:

$$W_{i,j} = \frac{|C_i \cap C_j|}{|C_{i,,} C_j|}$$

Where C_i represents neighborhood of friends surrounding user v_i

To measure the Subjective Well Being of user tweets, we used a open source software package for sentiment analysis OpinionFinder 2.0. For each user, the SWB coefficient S(u) can be expressed as

$$S(u) = \frac{N_{pos}(u)}{N} + \frac{N_{neu}(u)}{N} - \frac{N_{neg}(u)}{N}$$

The assortative model enforces a relationship between two groups of users. Group 1 represents users whose tweets have productive tweets and hence share similar SWB values. Group 2 represents users whose tweets are counterproductive. These two groups Assortative values can be denoted as A(G) and their correlation is then given by Pearson Correlation as:

$$A(G) = \rho(S(P), S(N)) = \frac{1}{n-1} \sum_i \left[\left(\frac{S(Pi) - S(\overline{P})}{\sigma(S(P))} \right) \left(\frac{S(Ni) - S(\overline{N})}{\sigma(S(N))} \right) \right]$$

The locale assortative of A(G) expressed as $A_L(G)$ is calculated for each user as:

$$S(L_u) = \frac{1}{|L(u)|} \sum_{l \in L(u)} S(l)$$

Finally, the average of assortivity of these two group of users A(G) and $A_L(G)$ is expressed as:

$$A_{swb(G)} = \rho\left((S(u), S(L_u)\right) = \frac{1}{n-1} \sum \left[\left(\frac{S(P) - S(\overline{P})}{\sigma S(P)}\right)\left(\frac{S(L_u) - S(\overline{L_u})}{\sigma S(P)}\right)\right]$$

Methods

Software used for this study is Weka 3.8, an open source machine learning software package. We've chosen Naïve Bayesian classification model.

Algorithm 1: Psuedo code of Assortativie Method

Input : Twitter Data

Output: Positive Tweets, Negative Tweets, Neutral Tweets

For each Tweet do

Begin

Calculate TSI(Token sentiment Score)

Determine the assortative /SWB factors

Construct a Friend Follower Network

Calculate A(G), $A_L(G)$, $A_{swb}(G)$

end

Naïve Bayesian classifier

The Naïve Bayesian classifier is an accurate and very accurate model. It's one of the most widely used algorithms in Sentimental Analysis. It's based on Bayesian Theorem and can be very suitable if the input's dimension is very high.

The Naïve Bayes Classifier takes input from Positive words list and Negative words list. We also train the NB classifier with training set. The probability of a word belongs to class C is given by the Class Probability P(C) multiplied by the products of the conditional probabilities of each word for that class.

$$P = P(C).\prod_i P(d_i/C) = P(C).\prod_i^n \frac{count(d_i,C)}{\sum_i count(d_i,C)}$$

Here count(d_i , c) is the number of occurrences of word d_i in class C, V_c is the total number of words in class C and n is the number of words in the document we are currently classifying.V_c does not change, so it can be placed outside of the product.

$$P = \frac{P(C)}{V_c^n} \prod_i^n count(d_i, C)$$

3. RESULTS AND DISCUSSION

To measure the performance of the assortative model, the following parameters are calculated.

True Positive (TP) is the set of tweets that are correctly assigned to the category, False Positive (FP) is the set of tweets incorrectly assigned to the category, False Negative (FN) is the set of reviews that are incorrectly not assigned to the category and True Negative (TN) is the set of tweets that are correctly not assigned to the category.

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

$$F_1 = \frac{2 * Recall * Precision}{(Recall + Precision)}$$

Experiments were conducted to measure the efficiency of the model. We used Product review dataset from Twitter and Flipkart.com. We retrieved reviews about Apple I phone and the features considered were: Camera, Display, Battery Life, and Audio.

Table 1: Positive Polarity Measures

Features	Precision	Recall	Accuracy	F_1 Metric
Camera	0.4498	0.6346	0.4765	0.5261
Display	0.5597	0.6534	0.7621	0.6032
Battery Life	0.7149	0.8756	0.6545	0.7871
Audio	0.7631	0.9179	0.7394	0.8333

Fig. 2: Positive Polarity Features

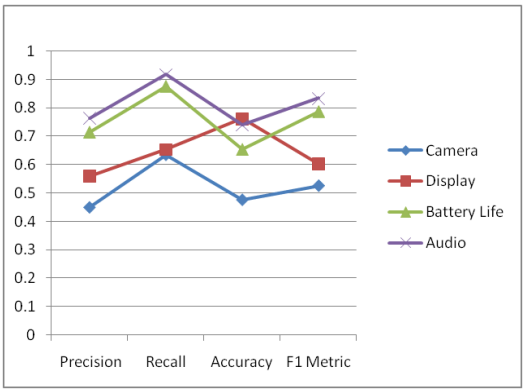
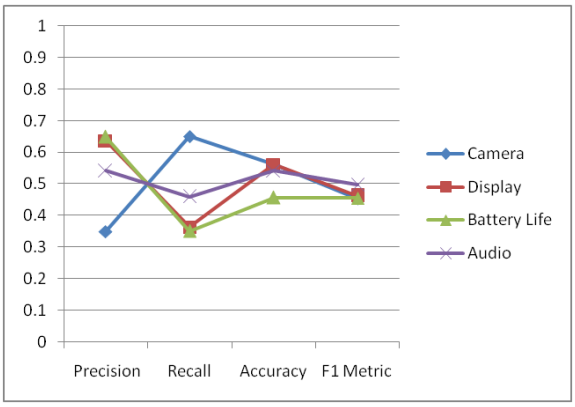


Table 2: Negative Polarity Measures

Features	Precision	Recall	Accuracy	F_1 Metric
Camera	0.3496	0.6504	0.5615	0.4547
Display	0.6359	0.3641	0.5613	0.4630
Battery Life	0.6495	0.3505	0.4561	0.4552
Audio	0.5411	0.4589	0.5424	0.4966

Fig. 3: Negative Polarity Features



4. CONCLUSION

In this research, a new assortative model for sentimental analysis has been proposed. This model gives importance to SWB characteristics of twitter users. The Friend-Follower network[4] of twitter users is used as a model to classify tweets into two groups. Further, we associated each group of twitter users with their assortative characteristics to better understand why their opinion about a product.

REFERENCES

1. Xing Fang, Justin Zhan Sentiment analysis using product review data In: Journal of Big Data (2015) 2: 5. doi:10.1186/s40537-015-0015-2

2. Hu M, Liu B (2004) Mining and summarizing customer reviews In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168–177.. ACM, New York, NY, USA.

3. Gann W-JK, Day J, Zhou S (2014) Twitter analytics for insider trading fraud detection system In: Proceedings of the second ASE international conference on Big Data. ASE.

4. Johan Bollen, Bruno Gonçalves, Guangchen Ruan, & Huina Mao Artificial Life Vol: 17 issue 3, 2011 237-251

5. Pennacchiotti, M., and Popescu, A. 2011. A machine learning approach to twitter user classification. In Proceedings of the International Conference on Weblogs and Social Media.

6.Teleimersion" Research Journal of Pharmaceutical, Biological and Chemical Sciences on March – April 2016 issue

7.A Human Computer Interfacing Application ", International Journal of pharma and bio sciences.

8. Rasheed M. Elawady, Sherif Barakat, Nora M.Elrashidy,"Different Feature Selection for Sentiment Classification, "International Journal of Information Science and Intelligent System, 3(1): 137-150, 2014

9. Albert Mayan .J, Lakshmi Priya .K, Yovan Felix .A," Empirical Study Of GUI To Reuse Unusable Test Cases", Journal of Theoretical and Applied Information Technology,Vol:75, Issue:2,pp:186-192,2014

10. Liu B. In: Web data mining; exploring hyperlinks, contents and usage data. Carey MJ, Ceri S, editors. Berlag Berlin Heidelberg: Springer; 2006

11. Agarwal, N.; Liu, H.; Murthy, S.; Sen, A.; and Wang, X. 2009. A social identity approach to identify familiar strangers in a social network. Proceedings of International Conference on Weblogs and Social Media (ICWSM 2009) 1–8

12. N. Barbieri, G. Manco, and F. Bonchi. Who to follow and why: Link prediction with explanations. In Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD 2014), New York City, USA, 2014.

13. Sudhakar.M, Mayan J.A., Srinivasan.N, "Intelligent data prediction system using data mining and neural networks", Advances in Intelligent Systems and Computing, March 2015
14. Stefano Faralli, Giovanni Stilo and Paola Velardi, Large Scale Homophily Analysis in Twitter Using a Twixonomy. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)
15. Mary Posonia .A, V.L.Jyothi,"Efficient XML Keyword search using H-Reduction factor and Interactive Algorithm" , International Review on Computers and Software(IRECOS), Vol-9,N-12,pp-2022-2030,2014
16. D. Usha Nandini, Dr. A. Ezil Sam Leni, "Shadow identification using ant colony optimization", journal of theoretical and applied information technology, August 2015. Vol.78. No.2, pp.195-200
17. R. Aroul canessane and S. Srinivasan, "A Framework for Analyzing the System Quality", International Conference on Circuit, Power and Computing Technology, IEEE transactions, pp.1111-1115, March 2013.
18. B.Bharathi and Aathilakshmi (2014), "Sans Douleur continous glucose scrutinizer system", International Journal of Applied Engineering Research, Paper code:26376,Vol9, no.20, pg-7221- 7225.