MODELLING SURVEILLANCE OF PRODUCT AND POLICIES USING SENTIMENT ANALYSIS

Isha Soni[1], Joshita Agarwal[2], Priya Agarwal[3]

[1]Student, Computer Science & Engineering, IMS Engineering College, Ghaziabad, UP, India
[2]Student, Computer Science & Engineering, IMS Engineering College, Ghaziabad, UP, India
[3] Student, Computer Science & Engineering, IMS Engineering College, Ghaziabad, UP, India

ABSTRACT

Sentiment Analysis or opinion mining is one of the major tasks of NLP (Natural Processing Language). This sentiment analysis has gain importance in recent years. In this paper, we aim to deal with how the sentiments can be analyzed and letting the customers know easily about the product. The general process of the working of the same with algorithms is proposed in the methodologies and research design. Data used in this paper are collected from online product reviews on amazon.com.. At last, we also gave an incite of the future works that can be done on sentiment analysis.

Keywords: *Sentiment analysis, sentiment, polarity classification, feature selection*.

1. INTRODUCTION

Sentiment is a thought, opinion, idea or an attitude that is brought about by a feeling about a situation. Sentiment Analysis also known as opinion mining, is a process of determining whether the piece of article or writing is positive, negative or neutral. It derives the attitude or behavior of the speaker and also studies people's sentiments towards certain entities. Internet is a resourceful place full of sentiment information. People post their opinion through various social media, such as blogs, social networking sites, or web sites. These data are collected and further analyzed by the researchers and developers. The data collected will then tell what and why the people think that the product is good or bad indicating to the rate of the positive or the negative or the neutral comments left by the customers on the social media.

While doing this analysis there are various flaws that are encountered. Many people post irrelevant comment or fake comment about the product on the basis of which the opinion of the product cannot be guaranteed. Some spammers also post spam comment on the comment box which are totally meaningless and not even related to the topic or product. This flaw can be easily overcome by two ways: Firstly proper inspection is done on the product review by checking the customer id before it is posted. Secondly each review must have a rating on it which will help to determine the quality of comment provided that is either positive, negative or neutral. [11]

In order to determine the sentiment of the overall document firstly identification of sentiment phrase like “good”, ”nice”, ”very good”, “average”, “not very good” is done. Then syntax matrix is used to determine the syntactic effect of the ordering of words. [10]

This paper deals with the way how the sentiments can be analyzed by reviewing and collecting the contents of the post and letting the customers know easily about the product by applying the required algorithms.

The rest of the paper is organized as follows: In the section ‘Background and Literary Review’ we provide a brief review on some related works on the sentiment analysis. Software packages and the classification models for the categorization and the algorithms used are presented in the section ‘Research Design and Methodology’. ‘Conclusion and Future Works’ section concludes the paper with some future scope.

2. BACKGROUND AND LITERATURE REVIEW

With the growth of online social networking sites, for example, forums, review sites, blogs, and micro blogs, the enthusiasm towards opinion mining has expanded essentially.

(Jeonghee Yi, 2003), has stated that extraction of sentiment (or opinion) about a subject from online text documents is done by sentiment Analyzer (SA). [12] (Wilson T, 2005) have pointed out that nature of sentiment expressions are not necessarily subjective, it can be short sentence or short text. . Categorizing the view or emotion with respect to the particular features of bodies is the focus or point of the Aspect-level SA. [2]

(Yu Liang-Chih Yu, 2013), has shown their research that extraction of people's opinions on features of an entity is the important task of opinion mining. There is a need to assemble these words and phrases, which are domain synonyms, into the same feature group to produce a useful synopsis [4]

(Singh & Piryani, 2013), has stated a fact-finding effort work on a new type of field particular feature-based investigatory for aspect-level sentiment analysis of picture analysis (reviews). A Senti Word Net established plan with two dissimilar semantic attribute choices consisting of adjectives, adverbs and verbs and n-gram attribute withdrawal is utilized.[5]

(Varghese, 2004), has used the concept where the word level extent attribute extraction is done utilizing Naive Bayesian Classifier. . The semantic alignment of the separate sentences is recovered from the contextual data or information. This machine learning viewpoint on average normally affirms a precision rate of 83%. [6]

The most interesting work from the point of view of the SA techniques is (Medhat et al., 2014), which presents a refined categorization of well-known SA techniques including new trends such as Emotion Detection (Rao et al., 2014), Building Resources and Transfer Learning.[7]

Kim and Hovy investigated the sentiment of the text and its holder regarding a given topic. Authors of the research paper have applied several classifiers. The first classifier was applied to each word in the sentence to get its polarity. The second classifier defined the polarity of the entire sentence expressed by opinion holder. In addition, the authors introduced the use of small initial list of seed words.

Medina and Lagos describe that aspect level (feature level) allows to extract opinions towards aspects of entities.[8]

The first technique that can be used for SA is the lexicon-based method. It uses a lexicon that consists of terms with respective sentiment scores to each term. The term can be associated with a single word, phrase or idiom [10].

3. RESEARCH DESIGN AND METHODOLOGY

3.1 Data Collection

Data used in this paper is a set of Patanjali product reviews collected from amazon.in. We collected, in total, over 500+ of Patanjali products.

Each review includes the customer id, star rating given by the customer for the product and review written by the customer for the product.

3.2 Proposed approach

Our methodology includes the following steps:

1. Reviews collected from www,amazon.in and 5 data set is being created.

- ♦ Data set 1 - Reviews
- ♦ Data set 2 – Positive words (ex- nice, great, excellent, good etc.)
- ♦ Data set 3 – Negative words (ex – bad, worst, dark etc.)
- ♦ Data set 4 – Neutral words (ex- average, unbiased, awaited, etc.)
- ♦ Data set 5 - Stop words (ex – but, how, or etc.)

These all segregations are done manually

2. Data cleaning is done by removing all the stop words
3. An array of frequent words is being prepared
4. Reviews are been classified as positive, negative and neutral by using Naïve Bayes classification algorithm.
5. Results are represented on pie chart and bar graph.

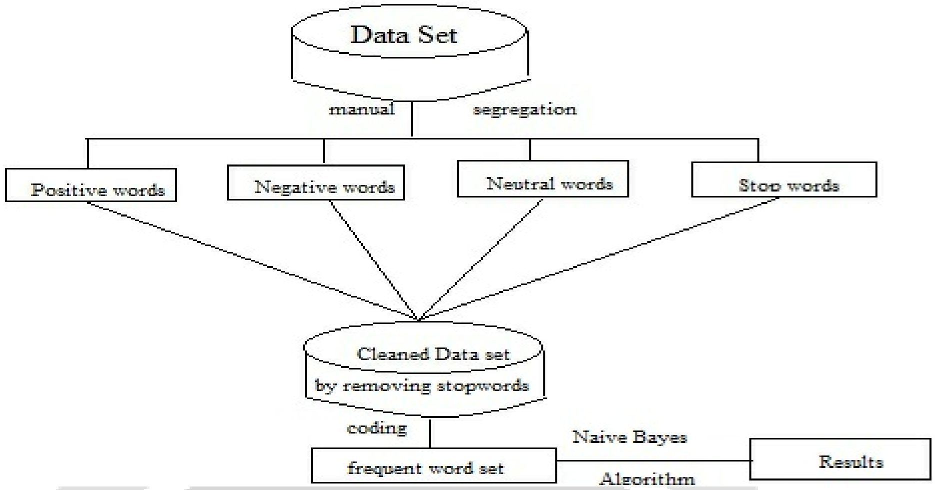


Fig.1 Proposed approach

3.3 Naive Bayes classification Algorithm

Naive Bayes algorithm is the algorithm that learns the probability of an object with certain features belonging to a particular group/class. In short, it is a probabilistic classifier. The Naive Bayes algorithm is called "naive" because it makes the assumption that the occurrence of a certain feature is independent of the occurrence of other features.

3.3.1 The Mathematics of the Naive Bayes Algorithm

The basis of Naive Bayes algorithm is Bayes' theorem or alternatively known as Bayes' rule or Bayes' law. It gives us a method to calculate the conditional probability, i.e., the probability of an event based on previous knowledge available on the events. More formally, Bayes' Theorem is stated as the following equation

$$\mathbf{P\ (A/B)} = \frac{P(\frac{B}{A})P(A)}{P(B)}$$

Let us understand the statement first and then we will look at the proof of the statement. The components of the above statement are:

- P (A/B):Probability (conditional probability) of occurrence of event A given the event B is true
- P(A) and P(B): Probabilities of the occurrence of event A and B respectively
- P(B/A) : Probability of the occurrence of event B given the event A is true

The terminology in the Bayesian method of probability (more commonly used) is as follows:

- A is called the **proposition** and B is called the **evidence.**
- P(A) is called the **prior** probability of proposition and P(B) is called the **prior** probability of evidence.
- P(A/B) is called the **posterior.**
- P(B/A) is the **likelihood**.

This sums the Bayes' theorem as:

$$\mathbf{Posterior} = \frac{(Likeli\ hood\ )(Proposition\ \ prior\ \ probability\ \ )}{Evidence\ \ prior\ \ probability}$$

3.3.2 Bayes' Theorem for Naive Bayes Algorithm

In a machine learning classification problem, there are multiple features and classes, say, C_1, C_2,..........C_k. The main aim in the Naive Bayes algorithm is to calculate the conditional probability of an object with a feature vector x_1, x_2,...........x_n belongs to a particular class C_i.

$$P(C_i / x_{1,} x_2, ...x_n) = \frac{P(x1,x2,.......xnICI)P(Ci)}{P(x1,x2,......xn)} \text{ for } 1 \leq i \leq k$$

Now, the numerator of the fraction on right-hand side of the equation above is

$$P(x_1, x_2, ...x_n / C_i).P(C_i) = P(x_1, x_2, ...x_n, C_i)$$

$$P(x_1, x_2, ...x_n, C_i) = P(x_1 / x_2, ...x_n, C_i).P(x_2, ...x_n, C_i)$$

$$= P(x_1 / x_2, ...x_n, C_i).P(x_2 / x_3, ...x_n, C_i)..\ P(x_1 / x_2, ...x_n, C_i).$$

$$= ......$$

$$= P(x_1 / x_2, ...x_n, C_i).P(x_2 / x_3, ...x_n, C_i)....P(x_{n-1}/x_n, C_i).P(x_n/C_i).P(C_{I)}$$

From the calculation above and the independence assumption, the Bayes theorem boils down to the following easy expression:

$$P(C_i / x_{1,} x_2, ...x_n) = (\prod_{j=1}^{j=n} P(xjICi)) . \frac{P(Ci)}{P(x1,x2,......xn)} \text{ for } 1 \leq i \leq k$$

4. IMPLEMENTATIONS & RESULTS

4.1 Datasets formation

Data used in this paper is a set of product reviews collected from amazon.in. We collected, in total, over 1000 product reviews in which the products belong to 2 major categories: beauty and food products. Those online reviews were posted by a number of reviewers (customers) towards these products. Each review includes the following information: 1) reviewer ID; 2) Review text; 3) rating;

Every rating is based on a 5-star scale, resulting all the ratings to be ranged from 1-star to 5-star with no existence of a half-star or a quarter-star.

Following are some of the reviews :

1. customer id= AH6ZXKWBU6GZNCRZ6ZXDNEOOZNHQ
   Packing was fresh, overall nice. Helps the skin to feel fresh during winters. Also helpful in case of burns. (5 stars)

2. customer id= AHHU2OTYEJP4FXYTUP2TSTR5IV4A
   Worst product ever never found a product like this....got extra pimples after using this product on the first time...waste of time (1 star)

3. customer id= AGX5KTQIHTTRC54HB3GNVTYB5ASA
   Good product but no discounts on patanjali in amazon. If there are discounts people show more interest to buy online (4 star)

Through these dataset of reviews, a total of four different dataset is being formed manually.

1. Positive words dataset
2. Negative words dataset
3. Neutral words dataset
4. Stop words dataset

In the above taken reviews, positive words are : fresh, good etc.
negative words are : worst, waste etc.
neutral words are : product, people etc.
Stop words are : after, but, etc.

Thus, finally we have segregated dataset of words. Now, data cleaning is done by removing all stop words after which an array of frequent words is being prepared.

4.2 Sentiment analysis of reviews

The reviews of different products from Amazon have been taken manually first for the sentiment analysis of the data. After analyzing the sentiment of all the reviews of that product the result is shown with the help of the chat i.e. pie chart and bar graph, showing how good, bad, or average the product is by showing the percentage of positive, negative, and neutral reviews and star ratings.

All the reviews of that product can be taken manually as well as on run time. Web scrapping is done to read the online reviews of the data and to give the analysis. Naïve Bayes classifier is used to give the sentiment analysis of the fetched reviews. Here the URL of the product page having reviews is taken after which all the reviews present on that page are fetched and then classification is done. Based on the classification, analysis is shown in the form of the graph again, showing about how much percentage the product is getting positive, negative, or neutral reviews. On the basis of this percentage the customer can get to know that whether the product is good to busy or not.

4.3 Summarization of result

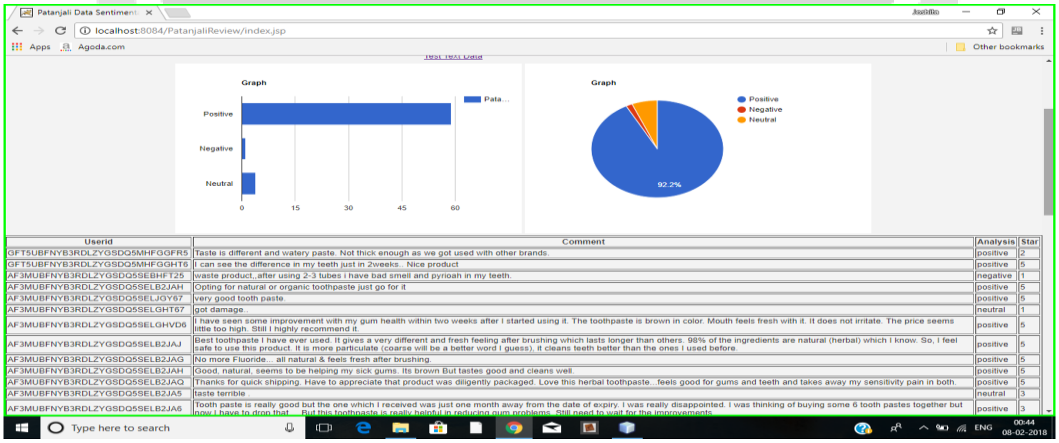
The system that has been developed for mining public opinion for product oriented reviews is implemented successfully .This system is found to provide good performance for varying kinds of reviews.

There are more than 100 spelling mistakes and slang words used in top 10 reviews that effect the performance of any sentiment analysis algorithm. In order to evaluate the effectiveness of the proposed feature extraction approach, we manually read every review and chose the major quality features mentioned in the reviews as the ground truth.

The classification of fetched comments using Naïve Bayes classifier with star rating and total negative, neutral and positive words. There are millions of products and millions of users reviews about products.

The limitation is that since our sentiment analysis scheme proposed in this study relies on the occurrence of sentiment tokens, the scheme may not work well for those reviews that purely contain implicit sentiments. An implicit sentiment is usually conveyed through some neutral words, making judgment of its sentiment polarity difficult. For example, sentence like “Item as described.", which frequently appears in positive reviews, consists of only neutral words. With those limitations in mind, our future work is to focus on solving those issues. Specifically, more features will be extracted.

For the issue of implicit sentiment analysis, our next step is to be able to detect the existence of such sentiment within the scope of a particular product



Userid	Comment	Analysis	Star
GFT5UBFNYB3RDLZYGSDQ5MHFGGFR5	Taste is different and watery paste. Not thick enough as we got used with other brands.	positive	2
GFT5UBFNYB3RDLZYGSDQ5MHFGGHT6	I can see the difference in my teeth just in 2weeks.. Nice product	positive	5
AF3MUBFNYB3RDLZYGSDQ5SEBHFT25	waste product,,after using 2-3 tubes i have bad smell and pyrioah in my teeth.	negative	1
AF3MUBFNYB3RDLZYGSDQ5SELB2JAH	Opting for natural or organic toothpaste just go for it	positive	5
AF3MUBFNYB3RDLZYGSDQ5SELJGY67	very good tooth paste.	positive	5
AF3MUBFNYB3RDLZYGSDQ5SELGHT67	got damage..	neutral	1
AF3MUBFNYB3RDLZYGSDQ5SELGHVD6	I have seen some improvement with my gum health within two weeks after I started using it. The toothpaste is brown in color. Mouth feels fresh with it. It does not irritate. The price seems little too high. Still I highly recommend it.	positive	5
AF3MUBFNYB3RDLZYGSDQ5SELB2JAJ	Best toothpaste I have ever used. It gives a very different and fresh feeling after brushing which lasts longer than others. 98% of the ingredients are natural (herbal) which I know. So, I feel safe to use this product. It is more particulate (coarse will be a better word I guess), it cleans teeth better than the ones I used before.	positive	5
AF3MUBFNYB3RDLZYGSDQ5SELB2JAG	No more Fluoride... all natural & feels fresh after brushing.	positive	5
AF3MUBFNYB3RDLZYGSDQ5SELB2JAH	Good, natural, seems to be helping my sick gums. Its brown But tastes good and cleans well.	positive	5
AF3MUBFNYB3RDLZYGSDQ5SELB2JAQ	Thanks for quick shipping. Have to appreciate that product was diligently packaged. Love this herbal toothpaste...feels good for gums and teeth and takes away my sensitivity pain in both.	positive	5
AF3MUBFNYB3RDLZYGSDQ5SELB2JA5	taste terrible .	neutral	3
AF3MUBFNYB3RDLZYGSDQ5SELB2JA6	Tooth paste is really good but the one which I received was just one month away from the date of expiry. I was really disappointed. I was thinking of buying some 6 tooth pastes together but now I have to drop that.... But this toothpaste is really helpful in reducing gum problems. Still need to wait for the improvements	positive	3

Chart-1 Distribution of reviews

5. CONCLUSION AND FUTURE WORKS

Opinion mining has become a fascinating research area due to the availability of a huge volume of user-generated content in review sites, forums and blogs. Opinion mining has applications in a variety of fields ranging from market research to decision making to advertising. With the help of opinion mining, companies can estimate the extent of product acceptance and can devise strategies to improve their product. Individuals can also use opinion mining tools to make decisions on their buying by comparing competitive products not just based on specifications but also based on user experience and public opinions.

Here, we see a few but important challenges for text analytics tasks like opinion mining. It is very difficult to distinguish between objective and subjective information, generally opinion words also occur in objective sentences, so it is very tough to handle these challenges. Many times we see customers posting the reviews in the blogs or forums with a lot of spelling mistakes which our dictionary cannot catch them and resulting in less accuracy of desired output. Most times of the times we see many spam blogs and spam reviews posted by the users. If we consider these reviews for performing Opinion Mining, we may get deviated from our desired results. So, lot of work has to be done in this field for identifying spam blogs, considering spelling mistakes and for other challenges

6 REFERENCES

[1] Gautam, G., & Yadav, D. (2014, August), "Sentiment analysis of twitter data using machine learning approaches and semantic analysis", In Contemporary computing (IC3), 2014 seventh international conference on (pp. 437-442). IEEE.

[2]Tetsuya Nakukawa, Jeonghee Yi, "Sentiment analysis : Capturing favourability using NLP", 2nd international on knowledge capture, 2003

[3] Theresa Wilson, Janyce Wiebe, Paul Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis", human language technologyand empirical method in NLP, 2005

[4] Liang-ChihYu, Jheng-LongWu, Pei-ChannChang, Hsuan-ShouChu, "Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news", elsevier, knowledge based system 41(2013) 89-07

.

[5] V.K. Singh, R. Piryani, A. Uddin, "Sentiment Analysis of Movie Reviews", Automatic computing Communication control and compressed sensing (iMac 4s), international Multi Conference, 2013

[6].Raisa Varghese , Jayasree M, "A survey on sentiment analysis and opinion mining ", *International Journal of research in Engineering & Technology, 2004*

[7] Jesus Serrano-Guerrero, José A. Olivas, Francisco P. Romero, Enrique Herrera-Viedma," Sentiment analysis: A review and comparative analysis of web service , Elsevier , information sciences, 2015

[8] Salas-Zárate, M. D. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., RodríguezGarcía, M. Á., & Valencia-García, R. (2017), "Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach. Computational and mathematical methods in medicine", *computational and mathematical methods in medicine*, 2017.

[9] Chiavetta, F., Bosco, G. L., & Pilato, G. (2016), "A Lexicon-based Approach for Sentiment Classification of Amazon Books Reviews in Italian Language", *WEBIST 2016, 12th International Conference on Web Information Systems and Technologies*, 2016

[10] https://www.lexalytics.com/technology/sentiment

[11] Xing Fang and Justin Zhan Journal of Big Data (2015), "Sentiment analysis using product review data", *journal of big data*, 2015

[12] http://blog.hackerearth.com/introduction-naive-bayes-algorithm-codes-python-r