Data Mining of Social Media for Analysis of Product Review

Mamatha Kothapalli
College of Science and Engineering,
University of Houston, Clear Lake

Ershad Sharifahmadian
College of Science and Engineering,
University of Houston, Clear Lake

Liwen Shih
College of Science and Engineering,
University of Houston, Clear Lake

ABSTRACT
Social media plays a crucial role in promoting different products. The data collected from the social media helps to improve the quality of products, and helps the customer to select the best product among available products. In this paper, an algorithm is developed based on text mining and TF-IDF (Term Frequency–Inverse Document Frequency) scores. In this paper, it is focused on removing unwanted words such as stop words, stemming words, then the processed data is used for finding sentiment words using NLTK (Natural Language Toolkit). The Stanford POS tagger is also used to tag the words into different categories like positive and negative. The proposed algorithm is implemented using JAVA NetBeans8.2 and achieved desired results. The proposed method can be expanded for the evaluation of different products based on customer reviews provided on the social media.

General Terms
Data Mining; NLTK; Text Tagging; TF-IDF

Keywords
iPhone; Negative Words; Positive Words; Sentiment Words; Side Effect Words; Social Media.

1. INTRODUCTION
Social media is a computer oriented technology that allows sharing and developing ideas in different fields. Social media involves technologies in mobile, computers, smart devices to extract data and make it available to the customers. Social media is one of the causes for changes in communication between individuals and large organizations directly [1]. There is a lot of difference between the paper-based and electronic media reviews from social media information. There is a change in the quality of information, the approach, usability and understanding in the product. Social media provides a dialog system between many sources to receivers. There are many users following social media like Facebook, WhatsApp, Tumblr, Instagram, Twitter, Pinterest, LinkedIn, Gab, Google+, YouTube, Viber, Snapchat, Weibo etc.

There are both positive and negative impacts on people due to social media. It is an effective marketing system to promote a product, organization and etc. Information can be produced, transferred and consumed using online social media. A relation is built between producers and consumers in the form of blog posts, comments, and tweets established in social media. Tracking information from social media gains feedback and methods to improve the product value in the market [2]. It helps customers to make good decisions on the product. There are many different ways to extract the techniques for social media modeling, analytics and optimization. We can understand the market by tracking competitors, monitor competitor prices, get more out of data and make better decisions.

In this modern era, there is a tremendous effect of technology on human lives and their activities. Customers decide about a product from the opinion of other consumers posted in social media. A research has concluded that market value of an electronic device is influenced by the reviews of each brand by the existing customers, evaluating patterns of telephone use. We discover customers who adopt all of the latest services and features of a particular phone company suggesting they will need some changes. The mobile reviews in social media mainly depend on latest features and technologies involved, the cost of the gadget, user experiences on it, battery life, latest market trend. The concept of online shopping helps to order a particular gadget to the desired location and also provide feedbacks on the products and their services, which helps the companies to improve the quality and features of the product by focusing on the needs of the customers. These also help in increasing the popularity of the products like Apple devices. The most trending product on e-commerce websites is the mobile phone. Many customers depend on social media to know the feedback and working of the mobile phones. Therefore, we have considered mobile phone reviews in this paper.

2. RELATED WORK
In this section, the proposed algorithm is implemented with the help of TF-IDF (Term Frequency- Inverse Document Frequency) scores, NLTK (Natural Language Toolkit) and Stanford POS Tagger. These concepts are explained as follows:

2.1.TF-IDF
TF-IDF plays a crucial role in text mining [3]. It helps to identify the strength of the document. Term Frequency tells how many times a word occurs in a group of sentences or a document. Inverse Document Frequency helps in identifying the importance of a word in a document. Some words occur number of times in the document but such words may not have importance, and some word occurs less number of times but have more importance and impact in qualifying the strength of the document. Thus, the TF-IDF helps in identifying such words and their importance in the document by giving some weights or count value to each word. The proposed algorithm is designed and developed in such a way to identify the most important aspects of a document by comparing several types of words from the available dataset and able to rate or rank a document based on TF-IDF methods.

2.2.NLTK Tool Kit

NL toolkit contains text processing libraries for tokenizing, tagging, and classifying the text [4]. In this paper, NLTK is used to tag the words presented in data. NLTK is utilized to evaluate and classify the words to match the meaning of the words within the given data. The tagged words are also used for further classification of the text.

2.3.Stanford POS Tagger

Stanford POS tagger examines the text and assigns parts of speech to each word so that they can be separated based on their importance in the text [5]. Each word in the given text is tagged with some priority and hence the positive, negative and drawback words can be distinguished easily.

3. METHOD

3.1 Data Collection

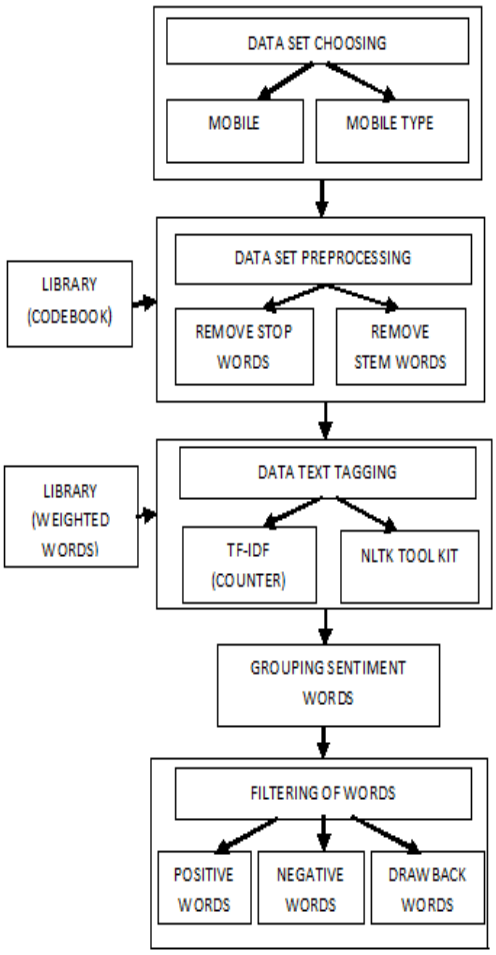

Figure 1: Block diagram of proposed method

For the experiment analysis, the data from various social media websites is considered. The data is searched based on the keywords related to cellphone and the type of the cellphone, e.g. iPhone. The data related to mobile phone is selected because the cellphone is popular. Moreover, the data includes variety of features such as photos, songs, etc. Modern cell phones are capable of accessing to internet, sending and receiving files, having GPS technology which helps users to find their location easily.

Here the iPhone is chosen because it is more popular among cellphone customers. Many people like iPhone due to its outstanding features. Therefore, the reviews on the iPhone are collected. These reviews are provided by many customers from various social media websites. Most of the reviews are taken from gsmarena website [6]. The opinion mining technique presented in [7] are also considered for classifying reviews about the products.

3.2 Initial Text Preprocessing

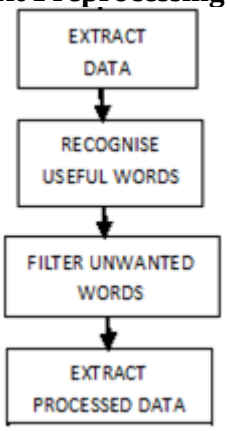

Figure 2: Data Preprocessing Steps

Initially, the data is collected and stored in a database. Then, the data is extracted from the database and useful words in the data are recognized. The data is then processed by removing unwanted words like stop words and stemming words [8]. These kind of words are not necessary for further review analysis, and their presence is not needed to find the strength of the sentence. Therefore, those words are removed in the preprocessing step, so that the other words can be easily measured. The data obtained after removing these words is undergone text tagging in the next phase.

3.3 Text Tagging

The data obtained in the preprocessed phase is used in this stage. Before tagging the text, each word gets a weight based on the following formula [8]:

$$weight_{i,d} \begin{cases} (\log f_{i,d} + 1)\, log\, \frac{n}{x_i}, & if\ f_{i,d} \geq 1 \\ 0, & otherwise \end{cases}$$

In which d represents document, n represents number of documents, $f_{i,d}$ represents the frequency of the word*(i)* and x_i represents number of documents in which i occurs [8].

These words were tagged using NLTK toolkit (www.nltk.org) [8]. NLTK is used to find the level of positivity or negativity of a word in a given text. Therefore, all the words are classified into positive words, negative words, and some words into drawback words. Here, the words appeared less than 10 times were removed. Stanford POS tagger is used which could tag words into different categories like positive, negative, side effect or drawback words along with parts of speech.

3.4 Grouping of Sentiment Words

Sentiment words are the group of positive, negative and drawback words. Sentiment words tell about the sensitiveness of the sentence in terms of positive and negative words [9]. In this stage, all the words obtained after removing unwanted words, stop words and stemming words are grouped together.

After that, the positive, negative and drawback words can be separated easily.

At this phase, influential users can also be identified. Influential users are the users who contribute more in providing the feedback about a product. So these influential users can be used in collecting the data. While searching the data is performed based on the keywords related to cellphone or type of the mobile phone, these influential users could be used as one of the keywords for collecting the data. With the help of influential users, the search process can be simpler.

3.5 Extracting Positive, Negative and Drawback words

From the group of sentiment words, the positive, negative and drawback words are separated based on TF-IDF scores given to each word. A library of default positive and negative words is used here. From that library of words, the presence of those words in our data are need to be checked. If any words match to the default library, they are filtered out and grouped accordingly. If more positive words repeat in our data, then the final review could be stated as positive feedback. If more negative words repeat, then the final review could be stated as negative feedback. Similarly, if more drawback words occur, then the final feedback is considered to have more drawbacks which means the product is not good to buy or use.

4. RESULTS

The method is implemented by JAVA NetBeans8.2 environments. First, the text data which contains customer reviews from various social websites are given to the proposed method.

As an example, the feedback provided by a customer about iPhone is considered; "iPhone is the best phone. It has good display and power backup. It has excellent features when compared to other phones. But it has bad feature for Bluetooth connectivity. Stunning design and build taller, more colorful display".

Removing unwanted words from the sentences (e.g. stop words and stemming words) from the input data is performed by comparing them with the default words provided in the library of stop words and stemming words. The output is shown in Figure 3. Here in this particular case, few stop words are identified which are listed in Figure 3 and because there are no any stemming words in the input data, the corresponding list is shown empty.

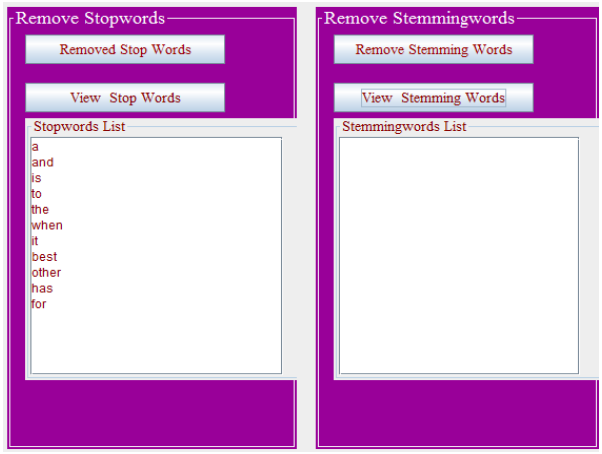

Figure 3: List of Stop words and Stemming words

Later, the sentiment words are selected. Here, four words as sentiment words listed in Figure 4 are found.

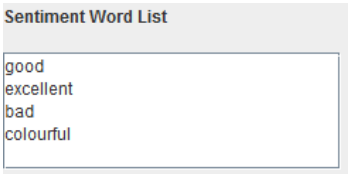

Figure 4: List of Sentiment words

The next step is to filter the words into different categories such as positive and negative. The data obtained from the previous step is transferred to the next step and from the group of sentiment words, the positive and negative words are filtered and listed. From the data that is chosen, two positive words and one negative word which are listed in Figure 5 are found.

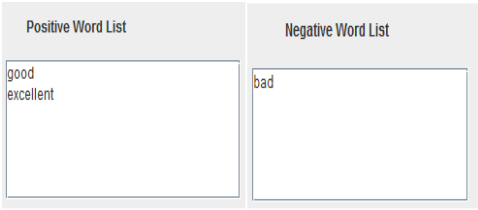

Figure 5: List of Positive and Negative words

Based on results, it is concluded that the reviews about the iPhone are positive.

5. CONCLUSION

In this paper, an algorithm was proposed to analyze reviews of product from social media. The algorithm is developed based on TF-IDF scores and NLTK toolkit. The algorithm is implemented by JAVA NetBeans8.2 environment and desired results were presented. With the help of the proposed method, the product review was able to be evaluated by identifying the positive, negative and drawback words. This algorithm can be implemented not only on the product reviews but also on various kinds of reviews like drug reviews [8], movie reviews and other kinds of service reviews. This kind of review analysis helps customers to choose the best available product or drug, and helps customers in deciding on what to buy and where to buy based on the positive and negative feedback provided. Moreover, customers also can save time by studying a number of reviews available over the internet. It also helps many organizations and companies to increase their turnover by improving the quality of their product and designing products which attract more customers.

6. FUTURE SCOPE

Future research will be done to use the proposed method to evaluate different products available on various websites with the help of advanced text classification techniques. The research can also be extended to process huge documents in the short period of time.

7. REFERENCES

[1] Li, Chun-Wen, Hui-Chi Chuang, and Sheng-Tun Li, 2016, "Hedonic Analysis for Consumer Electronics Using Online Product Reviews." 2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI).

[2] Jin, Jian, and Ping Ji., 2015, "Mining Online Product Reviews to Identify Consumers' Fine-grained Concerns."

12th International Symposium on Operations Research and Its Applications in Engineering, Technology and Management (ISORA 2015).

[3] Simsek, Atakan, and Pinar Karagoz, 2014 "Sentiment Enhanced Hybrid TF-IDF for Microblogs." 2014 IEEE Fourth International Conference on Big Data and Cloud Computing.

[4] Seshathriaathithyan, S., M. V. Sriram, S. Prasanna, and R. Venkatesan 2016 "Affective — Hierarchical Classification of Text — An Approach Using NLP Toolkit." 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT).

[5] Sari, Yunita, Mohd Fadzil Hassan, and Norshuhani Zamin, 2009, "Creating Extraction Pattern by Combining Part of Speech Tagger and Grammatical Parser." 2009 International Conference on Computer Technology and Development, 2009.

[6] http://www.gsmarena.com

[7] Aravindan, Siddharth, and Asif Ekbal, 2014, "Feature Extraction and Opinion Mining in Online Product Reviews." 2014 International Conference on Information Technology.

[8] Akay, Atlug, Andrei Dragomir, and Bjorn-Erik Erlandsson, 2015 "Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care." IEEE Journal of Biomedical and Health Informatics 19, no. 1: 210-218.

[9] Sudhakaran, Periakaruppan, Shanmugasundaram Hariharan, and Joan Lu., 2014, "Classifying Product Reviews from Balanced Datasets for Sentiment Analysis and Opinion Mining." 2014 6th International Conference on Multimedia, Computer Graphics and Broadcasting