TITLE: OPINION SPAM MINING

Ashwin Nair[1], Amit Phapale[2], Vidhi Yagnik[3], Kavita Bathe[4]

[1] Computer Engineering, K.J. Somaiya Institute of Engineering and Information Technology, Maharashtra, India
[2] Computer Engineering, K.J. Somaiya Institute of Engineering and Information Technology, Maharashtra, India
[3] Computer Engineering, K.J. Somaiya Institute of Engineering and Information Technology, Maharashtra, India
[4] Assistant Professor, Computer Engineering, K.J. Somaiya Institute of Engineering and Information Technology, Maharashtra, India

---***---

Abstract – *Nowadays, people are relying more on online shopping since it is having a lot of advantages. Before buying a product the customers go through the reviews posted about that product. This gives them an idea about how good or bad the product is. What others think has always been an important piece of information. These reviews tend to influence the customers. If these reviews are fake the reliability factor on these reviews will fall drastically. In order to prevent the customers from getting cheated it is very necessary to establish the authenticity and truthfulness of these reviews. Also the spammers might be posting these fake reviews to promote or disgrace a product on purpose. Thus to avoid a product from getting a false rating due to the above factors, it becomes essential to have a system to filter out these fake reviews and provide a better rating to the product. Through this paper we aim to propose a system that excludes such reviews through various filters and taking real reviews into consideration to analyze some of the algorithms of classification.*

Key Words: spam, spammers, hyperlink, early time window, frequency, location, Naïve Bayes, SVM.

1. INTRODUCTION

As we know that e commerce is booming around the world on large scale so is the technology. This results in production of various products. All these products are sold in the market with a hype created through advertisements. So nowadays people read reviews from customers who have already bought the product so as to understand what the product really is. But for this purpose the review needs to be genuine otherwise it is done by the manufacturer himself or by people who work for money. There are a few existing systems that deal with problem of identifying fake reviews and also to provide a trustworthy rating for a product using different algorithms. So to overcome the problem we are creating a system in which he filter out fake reviews and spammers so that we obtain a clean set of review data which will help the customers to know the product more closely. Before doing anything we need to segregate the data in number of parent products and further segregate into sub products. This classification will help us to understand the product type. Then we go on to spam. Once we clear the spam we go on to divided the clean data set to development and test set. The two algorithms are then taken into picture they are Naïve Bayes and SVM. These algorithms are then applied on the development set one by one to create a model. This model is then applied to the test set to get a better analysis of the data. This analysis then helps us to create a proper graph and the graph clearly suggest which is a better algorithm which can be used to classify the data that we have collected from different sources. The research focusses on spam detection and comparing the two algorithms Naïve Bayes and SVM.

2. RELATED WORK

2.1 YELP.com

[2]Yelp website allows users to post reviews of different locations of U.S.A. It aims at collecting and filtering the spam reviews. Since the data belonging to this site is highly confidential, the technique followed by them to filter reviews is unknown. Many people have researched about this and have estimated the procedure that yelp might be following.

According to this research, they believe that the classification is done on the basis of SVM2. Based on this classification, feature sets are created for reviews on the yelp website. It checks the past reviews posted by the user. It also determines whether the review is overall positive or negative after determining the authenticity of the user. Later each reviewer and spammers are recognized using different derivation techniques.

2.2 Meta Data Pattern

[3]The concept of meta data pattern was learnt from a business site- Dianping, which provides data about the various hotels of China. The research involves temporal and spatial patterns as well as a combination of both.

The overall idea provided through this research is to identify the percentage of people who change their IP addresses to post fake reviews and to identify a pattern if it exists.

According to Temporal Pattern it is observed that majority of the non spam reviews are posted during weekends and Monday. This is because a hotel has many customers checking-in in that period. But there are people who are actually paid to post fake reviews. They are freelancers and are observed to post during weekdays too.
Spatial Pattern suggests that the probability of people from small towns posting real reviews is far less as compared to that of people from large cities. This may be because the cost of travelling till there from their hometown is high and not every small town person can afford it so frequently.
Temporal and Spatial Pattern involves a combination of both which checks how far the reviewer is from the hotel on which the review is posted.

2.3 Ideological Discussion

[4]This research focuses on the speeches delivered or written by different people. They analyze the different words in the speech and the emotion attached to that word. This helps to decide whether a particular word is used in a positive or a negative context.
The data for this research was obtained from volconvo.com. The JTE-P model is used in this project which discovers the expressions involved in the speech. It also checks the relevance of the phrases used in the content of speech and it is ranked accordingly. It may happen that there are two people, one who speaks for and the other may speak against the same topic. In such a scenario, the researches decide what the correct content is and on that basis they score the context.

2.4 Identifying Noun Feature

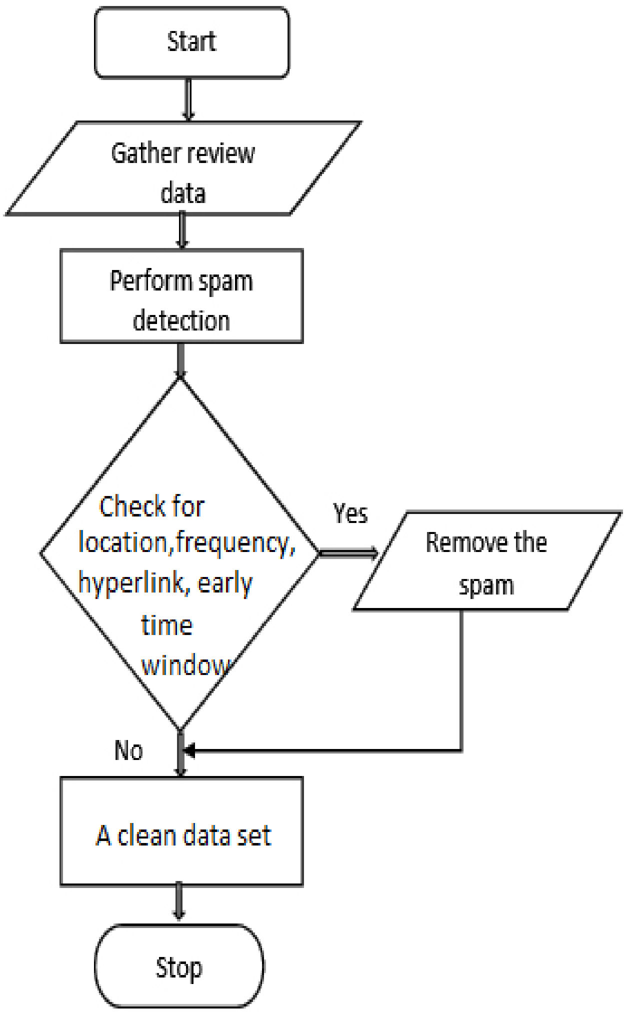
[5]In this research, the feature sets are extracted to detect a noun. These nouns identify the class, which are later used to determine the intensity of the word. If the noun is positive it is scored as +1 else if it is negative it is scored as -1. On the basis of this polarity, the entire score of the sentence is obtained by calculating the total of positive and negative features. This determines whether the sentence is positive or negative. For this they proposed a formula for scoring it was given as:

$$score(f) = \sum_{w_i : w_i \in S \wedge w_i \in L} \frac{w_i.SO}{dis(w_i, f)},$$

The only drawback of this formula is that it only helps to rate the noun or adjective in the sentence and doesn't help to check its intensity.

3. PROPOSED SYSTEM

Our project focuses on the idea to provide best reliable reviews based on the data set which is provided by Amazon to us for testing and programming purposes. The data needs to be refined into classes of different products. This reviews then needs to be checked for spam. The spam are checked using hyperlinks present in the review and many other unique features to detect spam. We then move on to compare the algorithms, Naïve Bayes and SVM with the received clean set of data. This comparison will help us understand which of the algorithm is best suited for classification for the particular case.



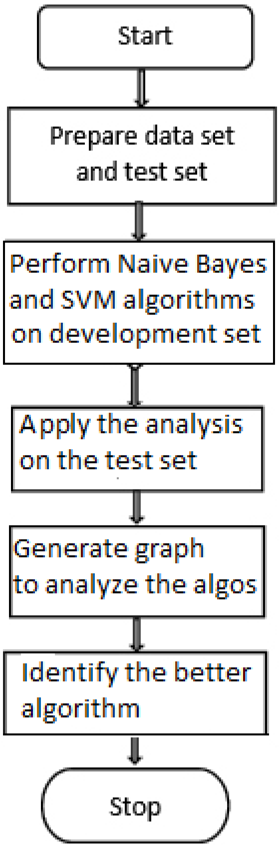
Flowchart -1: Spam removal

A. Data Collection
To score a product we need review data in hand. For this we are collecting the reviews provided by amazon for research and programming purposes.

B. Detecting spam and spammers
We detect the spam through various unique features like location of the user, the frequency in which the user reviews the products, early time framing and also we detect links in reviews which are done to distract users from one product to another. For detecting spams we have first checked these spams on the data set obtained, then we have created a live webpage where these spams can be detected.
[i] Location of the user: This feature plays a major role as studied in the literature review, people not belonging to the region where a product is sold express their view on the product creating a false review just to earn money or to promoted or demote the product. This done through pattern matching.
[ii] Frequency of review by a single reviewer: Normally, a person does not by more than 10 products in a span of a week but we find people writing reviews for more than the 10 products that they have bought to earn money. The system detects such spammers which helps in reducing the spams. The spam is analyzed through date and product differences and its patterns.
[iii] Early time window: In this a product is launched and about 100 pieces of that product is sold, but more than 100 reviews are detected this is early time window. Also a product is just launched and people start expressing the reviews before a product is sold, this is also one kind of spam our research finds such kind of reviews through time and date windows.
[iv] Hyperlinks: Spammers tend to post links to different product to divert the attention of the customers from one product to another. Sometimes links are helpful, but many a times these are done to promote some other product. We have created this detection through pattern matching.

Fig -1: Spam spotted hyperlink.

C. Creating Development and Test Set
Once the spams are removed we obtain a clean set of data. This data is then divided into test set and development set. The development set are those reviews on which the feature of a word being positive or negative will be analyzed and these word features will be then used on the test set to analyze how correctly the algorithms work.



Flowchart -2: Analyzing the data

D. Working with Algorithms
We are analyzing two algorithms on the development set data. The algorithms used are Naïve Bayes and Support Vector Machine (SVM). This algorithms after applying to development set data is then analyzed with the test set data. We get the reading of the correctly classified set of words. In our, research it is found that SVM algorithm has given us a better result than the Naïve Bayes algorithm.
[i] Naïve Bayes algorithm
The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors.
Bayes theorem provides a way of calculating the posterior probability, $P(c/x)$, from $P(c)$, $P(x)$, and $P(x/c)$. Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence

$$\underbrace{P(c\mid x)}_{\text{Posterior Probability}} = \frac{\overbrace{P(x\mid c)}^{\text{Likelihood}}\,\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c\mid \mathbf{X}) = P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c)$$

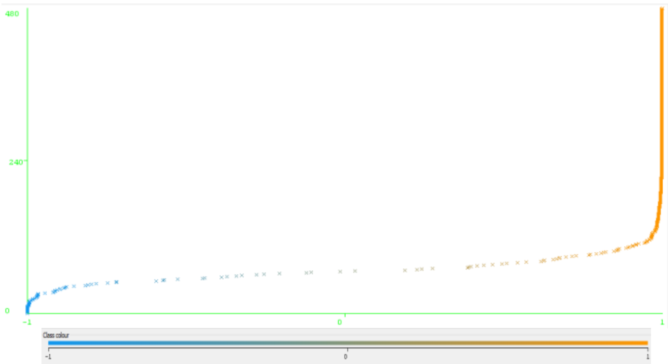


Chart -1: Naïve Bayes marginal curve

The above graph shows that words which have half positive or negative means in the development set are also taken into consideration.

[ii] Support Vector Machine

A Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors (cases) that define the hyperplane are the support vectors.

We are given a training dataset of "n" points of the form (X1, Y1) upto (Xn, Yn). In which Y is either -1 or +1. Each Yi indicates the class to which Xi belongs. Here we find the maximum margin hyperplane that divides the group of points Xi for which Yi = -1 from the group of points with Yi = +1.

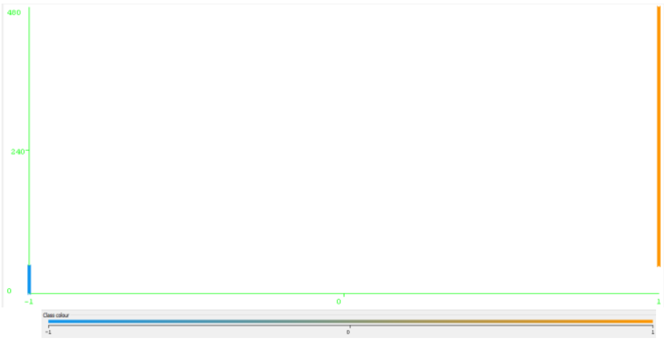


Chart -1: SVM marginal curve

We can see here that all the words are either taken as positive or negative, none are given half positive or negative trademark.

[iii] Comparison between two algorithms:

Naïve Bayes Algorithm	SVM algorithm
Naive Bayes classifiers are computationally fast when making decisions.	The training time for the SVM classifier is comparatively higher than Naïve Bayes classifier.
Naïve Bayes algorithm gives better result with small number of documents and tend to reach best performance on medium size of documents.	SVM gives better prediction with more number of documents. When the whole training set is used the performance of SVM is almost similar to Naïve Bayes.
The cross validation for the development set provided by us took 0.19 seconds	The cross validation for the development set provided by us took 0.27 seconds
The time taken to validate the training model with the test set is 0.17 seconds	The time taken to validate the training model with the test set is 0.29 seconds
The validation after applying this model to the test set it gave us a 86.45% correctly classified rate	The validation after applying this model to the test set it gave us a 92.5% correctly classified rate.

4. CONCLUSION

Through this paper the set of unfiltered review data will undergo a test for detection and removal of various types of fake or spam reviews with the help of natural language processing. We aim to provide a technique to filter all the junk reviews and prepare a clean and reliable data set. After this our objective we applied two algorithms Naïve Bayes and SVM and analyze which of the algorithm is better. According to our analysis SVM classification has a better result in arriving to a correct decision

ACKNOWLEDGEMENT

We are very grateful to Mr. Bing Liu for providing us the data set of amazon for our further research. We also thank Mr. Julian McAuley for the amazon dataset.

REFERENCES

[1] Jing Wang, Clement. T. Yu, Philip S. Yu, Bing Liu, Weiyi Meng. "Diversionary comments under blog posts." Accepted. ACM Transactions on the Web (TWEB), 2015.

[2] Huayi Li, Zhiyuan Chen, Arjun Mukherjee, Bing Liu and Jidong Shao. "Analyzing and Detecting Opinion Spam on a Large-scale Dataset via Temporal and Spatial Patterns." Short paper at *ICWSM-2015*, 2015.

[3] Arjun Mukherjee, Bing Liu, and Natalie Glance "Spotting Fake Reviewer Groups in Consumer Reviews." International World Wide Web Conference (WWW-2012), Lyon, France, April 16-20, 2012. (See media coverage of this work from April 16, 2012)

[4] Arjun Mukherjee, Bing Liu. Discovering "User Interactions in Ideological Discussions." Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013), August 4-9, 2013, Sofia, Bulgaria.

[5] Lei Zhang and Bing Liu. "Identifying Noun Product Features that Imply Opinions." ACL-2011 (short paper), Portland, Oregon, USA, June 19-24, 2011.

[6] Sentiment analysis and opinion mining by Bing Liu

[7] J. McAuley, C. Targett, J. Shi, A. van den Hengel "Image-based recommendations on styles and

substitutes"
SIGIR, 2015
[8] J. McAuley, R. Pandey, J. Leskovec "Inferring networks of substitutable and complementary products" *Knowledge Discovery and Data Mining*, 2015