INTERNATIONAL EDUCATIONAL APPLIED RESEARCH JOURNAL (IEARJ)
Volume 04, Issue 03, March 2020
E-ISSN: 2456-6713

DETECTING SPAMMER REVIEWS WITH FREQUENT TRANSACTIONS ON ONLINE-ECOMMERCE DATA

Poongodi K[1], Karthick S[2]

[1]*Student, K.S. Rangasamy College of Technology, Tiruchengode, Tamilnadu, India.*

[2]*Assistant Professor, K.S. Rangasamy College of Technology, Tiruchengode, Tamilnadu, India.*

Abstract: *As e-commerce is growing and becoming popular day-by-day, the number of reviews received from customer about any product grows rapidly. People nowadays heavily rely on reviews before buying anything. Product reviews play an important role in deciding the sale of a particular product on the ecommerce websites or applications like Flipkart, Amazon, Snapdeal, etc. In this paper, this project proposes a framework to detect fake product reviews or spam reviews by using Opinion Mining. The Opinion mining is also known as Sentiment Analysis. In sentiment analysis, this project tries to figure out the opinion of a customer through a piece of text. The proposed method called VWNB-FIUT (Value Weighted Naïve Bayes with Frequent Pattern Ultra Metric Tree) automatically classifies users' reviews into "suspicious", "clear" and "hazy" categories by phase-wise processing. The hazy category recursively eliminates elements into suspicious or clear. This results into richer detection and be useful to business organization as well as to customers. Business organization can monitor their product selling by analysing and understanding what the customers are saying about products. This can help customers to purchase valuable product and spend their money on quality products. Finally end users see that each individual review with polarity scores and credibility score annotated on it. This project first takes the review and check if the review is related to the specific product with the help of VWNB. This project use Spam dictionary to identify the spam words in the reviews by using FIUT. In Text Mining this project applies several algorithms and on the basis of these algorithms this project gets the specific results.*

Keywords: *VWNB, FIUT, hPSD, EDI.*

I. INTRODUCTION:

An ecommerce platform is a software application that allows online businesses to manage their website, marketing, sales, and operations. E-commerce is the activity of buying or selling of products on online services or over the Internet. Electronic commerce draws on technologies such as mobile commerce, electronic funds transfer, supply chain management, Internet marketing, online transaction processing, electronic data interchange (EDI), inventory management systems, and automated data collection systems. Product inference mining is a process of tracking the mood of the public about a particular product [13]. Opinions can be essential when it's use to make a decision or choose among multiple option. Information-gathering behaviour has always been to find out what other people think. The availability of opinion-rich resources such as online review sites and personal blogs, and challenges arise, to understand the opinions of others people. Product inference mining is extracting people's opinion from the web. It is also known as sentiment analysis. There are three tasks for opinion document, space, phrase level opinion mining. The area of opinion analysis is to predict the polarity of a piece of opinion text as positive or negative. Here, the tasks related to opinion analysis are Subjectivity Detection, Sentiment Prediction, Aspect Based Sentiment Summarization, Contrastive Viewpoint Summarization, Text Summarization for Opinions, Predicting Helpfulness of Online Comments/Reviews, Product Inference-Based Entity Ranking. It is well-known that many online reviews are not written by genuine users of products, but by spammers who write fake reviews to promote or demote some target products. Although some existing works have been done to detect fake reviews and individual spammers, to our knowledge, no work has been done on detecting spammer groups. The first is the task of learning from labeled and unlabeled examples, which is commonly known as semi supervised learning. In this chapter, this project also calls it LU learning (L and U stand for "labeled" and "unlabeled" respectively). In this learning setting, there is a small set of labeled examples of every class,

Learning from positive and unlabelled data or PU learning is the setting where a learner only has access to positive examples and unlabelled data. The assumption is that the unlabeled data can contain both positive and negative examples. A Reliability Study for Evaluating Information Extraction from Radiology Reports. The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. In statistics, an expectation maximization (EM) algorithm is an iterative method to find maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables.

Data Mining Concept:

Data Mining is an analytic process designed to explore data (usually large amounts of data typically business or market

related) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction and predictive data mining is the most common type of data mining and one that has the most direct business applications. Data Analysis and modeling and it shares with them both some components of its general approaches and specific techniques. The process of the data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification, and (3) deployment.

Stage 1: Exploration:

This stage usually starts with data preparation which may involve cleaning data, data transformations, selecting subsets of records and - in case of data sets with large numbers of variables ("fields") - performing some preliminary feature selection operations to bring the number of variables to a manageable range (depending on the statistical methods which are being considered). Then, depending on the nature of the analytic problem, this first stage of the process of data mining may involve anywhere between a simple choice of straightforward predictors for a regression model, to elaborate exploratory analyses using a wide variety of graphical and statistical methods (Exploratory Data Analysis (EDA)) in order to identify the most relevant variables and determine the complexity and the general nature of models that can be taken into account in the next stage.

Stage 2: Model building and validation:

This stage involves considering various models and choosing the best one based on their predictive performance. This may sound like a simple operation, but in fact, it sometimes involves a very elaborate process.

Stage 3: Deployment:

That final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome.

II. EXISTING SYSTEM:

A partially supervised learning model (VWNB) to detect spammer groups. By labelling some spammer groups as positive instances, VWNB applies positive unlabelled learning (PU-learning) as a classifier to study positive instances from a spammer group detector (labelled spammer groups) and unlabelled instances (unlabelled groups). Specifically, they extract a reliable negative set the positive instances of the terms and the distinctive features. By combining the positive instances, extracted Negative instances and unlabelled instances, they convert the PU-learning problem into a well-known semi supervised learning problem, and then use a Naive Bayesian model and an EM algorithm to train a classifier for spammer group detection a survey performed by a leading site has shown that: More than 80% of the online customers look at the reviews available. 50% base their purchase on the ratings of the products. 30% of the customers compare the ratings of similar products before making their decision. Clearly consumers value the feedback given by other users as do the companies that sell such products.

Drawbacks:

- Identifying reviews in the free text reviews, a straightforward solution is to employ an existing aspect identification approach.
- The spam classification instance grouping may low.
- Less accuracy prediction on opinion analysis.
- User review-based word alignment is cumbersome.
- High in latency to analyse the datasets.
- Naive Bayes theorem is developed on the mathematical Bayes Theorem in probability increase ROC Value that reduces overall accuracy.

III. PROPOSED SYSTEM:

In this proposed system the Value weighted Naïve Bayes with Frequent Pattern Ultra Metric Tree based opinion review analysis reviews possess the following characteristics: (a) they are frequently commented in user reviews; and (b) users' opinions on these reviews greatly influence their overall opinions on the reviews. A straightforward frequency-based solution is to regard the reviews that are frequently commented in user reviews as important. However, users' opinions on the frequent reviews may not influence their overall opinions on the reviews, and would not influence their purchasing decisions. This project am measuring public concern using a two-step sentiment word alignment approach. This work VWNB-FIUT Identifying fake reviews from a large dataset is challenging enough to become an important research problem. Business organizations, specialists and academics are battling to find the best system for opinion spam analysis. A single algorithm cannot solve all the problems' and challenges faced in today's generation with advancements in technologies, though a few are very efficient in analysis. It also improving the performance of the opinion spam analysis, and developing one that is consistently efficient across all categories of data. The opinion reviews obtained from users can be classified into positive or negative reviews, which can be used by a consumer to select a product. This work aims to classify amazon reviews into groups of positive or negative polarity by using machine learning algorithms. In this study, they analyse online amazon reviews using proposed methods in order to detect fake reviews. The text classification methods are applied to a dataset of Amazon or google play store reviews.

Advantages:

- The reviews containing explicit content and with swear words are not taken into consideration and are removed from the dataset.
- Sentiment score for each word is calculated when words are extracted into a form of dictionary or so called 'Bag of Words (BOW)' it first identifies the nouns and noun phrases in the documents. The occurrence frequencies of the nouns and noun phrases are counted, and only the frequent ones are kept as reviews.
- The language model was built on reviews, and used to predict the related scores of the candidate reviews. The candidates with low scores were then filtered out.
- The admin can easily identify related opinion reviews on that session.
- Easily determine reviews quality by using customer reviews.
- I can find Based on the number of reviews classified as Personal Negative; this project computes a Measure of Concern (MOC) and a timeline of the MOC. This project attempt to correlate peaks of the MOC timeline to peaks of the News (Non-Personal) timeline.
- Best accuracy results are achieved.
- Analysis of product after spam removal is done on the basis of their respective features.

IV. MODULES:

A. Domain Specific Sentiment Knowledge and Spam Similarity Module:

For example, an unlabelled review in Kitchen domain is "the food processor is quick and easy for making baby food. Since "quick" and "easy" are used to describe the same target in the same sentence, they probably convey the same sentiment. This is because people usually hold consistent opinions towards the same target in a short period, which is validated by social science theories such as sentiment consistency. If this project can find more cases that these two words co-occur in the same sentence, then this project can infer that they tend to convey similar sentiments in this domain. Thus, in this module this project proposes to extract domain-specific sentiment relations among words from the unlabelled samples based on their co-occurrence patterns.

B. Spam Similiarity Textual Content Based Spam Similiarity:

The textual content-based spam similarity is motivated by the observation that although different topics and opinion targets are discussed in different domains, similar domains may share many common terms. For example, in both Smart Phone and Digital Camera domains, terms like "screen", "battery", and "image" are frequently used. In contrast, the probability of two far different domains such as Smart Phone and Book sharing many common terms is low. Thus, this project proposes to measure the similarity between domains based on their textual content

C. Sentiment Expression Based Spam Similiarity:

The textual content-based Spam Similarity introduced in previous section can measure whether two domains have similar word usage patterns. However, high similarity in textual content does not necessarily mean that sentiment words are used in similar ways in these domains. For example, both CPU and Battery belong to electronic hardware. In CPU domain, the word "fast" is usually positive. For instance, "Intel Core i7 is very fast." However, in Battery domain, the word "fast" is frequently used as a negative word (e.g., "This battery runs out too fast"). Thus, measuring SPAM SIMILIARITY based on sentiment expressions may be more suitable for multi-domain spam review classification task.

D. Similarity Analysis and Spam Review Classification:

In this module, this project proposes to extract prior general sentiment knowledge from general-purpose sentiment lexicons to enhance the learning of the global spam review classification model. Given multiple domains to be analyzed, a small number of labelled samples in these domains, the domain similarities between them, the general sentiment knowledge extracted from general purpose sentiment lexicons, and the domain-specific sentiment knowledge of each domain extracted from both labelled and unlabelled samples, the goal of our approach is to train accurate domain-specific sentiment classifiers for multiple domains in a collaborative way.

V. CONCLUSION:

This work proposes a partially supervised learning-based Model VWNB to detect spammer groups from product reviews. First, the frequent item mining using the VWNB model (FIM) to discover spammer group candidates from the review data. Then, manually labelling some spammer groups as positive instances, the VWNB employs construct to PU-Learning the positive and unlabelled instances of a classier to identify the real candidates from the group of real spammer groups. In particular, the VWNB dense a feature strength functions. The group features of the measure of discriminatory power, and then High discriminative with iteratively removes instances Get a reliable set of unlabelled instances from only non-spammer groups of negative set consisting. By combining the

positive, negative and unlabelled instances. The well-known semi supervised into the PU-Learning problem learning problem, and employer Naive Bayesian Model and EM algorithm to construct a classified as spammer group detector. Experiments on Amazon.cn show that the proposed VWNB model outperforms both supervised and Spammer group detection on unsupervised learning methods. Improvement in the area of our future work of the VWNB model. Beyond the Naive Bayesian model used in VWNB, this project will investigate and incorporate more classification models such as neural network, Semi-Supervised SVM (S3VM) and even ensemble methods. On the positive instances acquisition and RN extraction, this project plan to involve Improving the accuracy and efficiency of active learning data labelling.

REFERENCES:

1. Akoglu. L, Chandy. R, and Faloutsos. C, (2013), ‘‘Opinion fraud detection in online reviews by network effects,’’ in Proc. ICWSM, vol. 13, pp. 2–11.
2. Heydari. A, Tavakoli. M, and Salim. N, (Oct. 2016), ‘‘Detection of fake opinions using time series,’’ Expert Syst. Appl., vol. 58, pp. 83–92.
3. Jindal. N and Liu .B, (2008), ‘‘Opinion spam and analysis,’’ in Proc. Int. Conf. B Search Data Mining, pp. 219–230.
4. Li. F, Huang. M, Yang. Y, and Zhu. X, (2011), ‘‘Learning to identify review spam,’’ in Proc. Int. Joint Conf. Artif. Intell. (IJCAI), vol. 22. no. 3, pp. 219–230.
5. Li .H et al., (2017), ‘‘Bimodal distribution and co-bursting in review spam detection,’’ in Proc. 26th Int. Conf. World Wide b, pp. 1063–1072.
6. Liu .B and Lee. W. S, (2011), ‘‘Partially supervised learning,’’ in B Data Mining. Berlin, Germany: Springer, pp. 171–208.
7. Mukherjee. A et al., (2013) ‘‘Spotting opinion spammers using behavioural footprints,’’ in Proc. 19th ACM SIGKDD Int. Conf. Know. Discovery Data Mining, pp. 632–640.
8. Mukherjee. A, Liu. B, and Glance. N, (2012), ‘‘Spotting fake reviver groups in consumer reviews,’’ in Proc. 21st Int. Conf. World Wide b, pp. 191–200.
9. Santosh. K.C and Mukherjee. A, (2016), ‘‘on the temporal dynamics of opinion spamming: Case studies on yelp,’’ in Proc. 25th Int. Conf. World Wide b, pp. 369–379.
10. Shehnepoor. S, Salehi. M, Farahbakhsh. R, and Crespi. N, (Jul. 2017), ‘‘Net Spam: A network-based spam detection framework for reviews in online social media,’’ IEEE Trans. Inf. Forensics Security, vol. 12, no. 7, pp. 1585–1595.
11. Wang. Y, Wu. Z, Bu .Z, Cao. J, and Yang. D, (2016), ‘‘Discovering shilling groups in a real e-commerce platform,’’ Online Inf. Rev., vol. 40, no. 1, pp. 62–78.
12. Wu. Z, Wang. Y, Wang. Y, Wu. J, Cao. J, and Zhang. L, (Nov. 2015) ‘‘Spammers detection from product reviews: A hybrid model,’’ in Proc. IEEE Int. Conf. Data Mining (ICDM), pp. 1039–1044.
13. Xie. S, Wang. G, Lin. S, and Yu. P.S, (2012), ‘‘Review spam detection via temporal pattern discovery,’’ in Proc. 18th ACM SIGKDD Int. Conf. Know. Discovery Data Mining, pp. 823–831.
14. Xu. C, Zhang. J, Chang. K, and Long. C, (2013), ‘‘Uncovering collusive spammers in Chinese review websites,’’ in Proc. 22nd ACM Int. Conf. Conf. Inf. Know. Manage, pp. 979–988.
15. Zhu. F and Zhang. X, (2010), ‘‘Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics,’’ J. Marketing, vol. 74, no. 2, pp. 133–148.