Available online at www.worldscientificnews.com

World Scientific News

An International Scientific Journal

WSN 113 (2018) 10-19 EISSN 2392-2192

A Single Substitute Review Creation Using Data Mining Techniques

Rishabh Varma[a] and **Rishabh Saxena**[b]

Department of Information Technology, KIET Group of Institutions, (Accredited by NAAC with "A" Grade), 13-Km Stone, Ghaziabad-Meerut Road, Ghaziabad – 201206, UP, India

[a,b]E-mail address: rishabhvarma22@gmail.com , rishabhsaxena2897@gmail.com

ABSTRACT

E-commerce has been the latest sensational bubble that humankind has witnessed in recent years. Most of the products these days are bought and sold online through the medium of internet. One of the most common things we witness on an E-commerce site is reviews. These reviews are massive and it is almost impossible to go through each and every one of those reviews yet many customers' are influenced by them. Most of these reviews are insignificant and useless; therefore we propose a way to replace all the reviews with a small paragraph which can appropriately summarize all the reviews in an epitome. This includes removing spam reviews and proposing a methodology to appropriately weigh a given review based on its content. Then the final step is creating a single paragraph review containing all the appropriate information about the product.

Keywords**:** E-commerce, reviews, spam reviews, single paragraph review

1. INTRODUCTION

Every E-commerce site hosts a large number of products and for every product; there is a multitude of reviews to analyse at a given time for a normal human being. This makes it really tough to create a clear picture of how a product might function and feel based on the

(Received 02 October 2018; Accepted 23 October 2018; Date of Publication 24 October 2018)

performance of its components. In order to remove this uncertainty, each and every review must be analysed properly to check whether it is a spam and if not then we can use its content to form an association between the characteristics of our product. The first step is spam detection for which we use RLOSD model for detecting spams [1] and then removing them. Text pre-processing is applied to the text obtained from the user.

Text pre-processing involves tokenization, stop-word-removal and stemming and some other techniques. Tokenization involves splitting the text in the form of basic words called tokens. It is used to identify keywords in the stream of texts. Stop-word-removal is the process of removal of words which do not convey a special meaning in the document like the, and, this...etc. Stemming is done to obtain the root word of the data and remove suffixes like -ing, -ion, etc.

After feature extraction, duplicate features are eliminated [18]. Once we have done the pre-processing [19], found the term frequency and inverse document frequency. Now using RLOSD model, we remove the spam reviews. We now have the list of all keywords and all non-spam reviews. Thus a graph is formed between the object and its characteristics. This graph is known as “Characteristics Graph”. Every edge of a graph is given a weight based on the “Characteristic Index Rating Table”, the value of weight for the edges are updated (added) every time a new review is encountered. Finally, the resultant weight of each node and the frequency of each attribute in all reviews are estimated and therefore the paragraph is formed with the ontology of words taken, based on the Characteristic index rating and Frequency estimation.

This paper focuses on creating a single usable review which can replace all the other reviews and is also accurate at the same time. The prediction model comprises graph-based algorithms. This is done by feeding the system with a data set for training the system. This framework can be used in different scenarios regarding other domains such as comparing multiple products at the same time based on reviews of the users. It is highly effective in calculating the results and learning based on the reviews. It can also be used to prioritize the product view based on the reviews as all the products can be reviewed and a list is formed of the products with the highest review to the one with lowest review rating calculated by our framework [4].

Emoticons are a very important part of any review over the internet. It is well known that they are the most expressive part of any text review as they convey the almost real essence of the expression which the user wants to convey [5]. Hence, it is of prime importance to analyse the emoticons used in any review so that the real sentiment of the text is accessible [16].

2. PROPOSED METHODOLOGY

The proposed methodology helps to replace ambiguous and diverse reviews by a single review which covers all the aspects of a product.

The aim is to extract information from the reviews of the user, segregating the spam reviews and use non-spam reviews for different purposes such as product comparison. The model also includes the analysis of emoticons in order to completely parse the statements.

I. Dataset description

The data is obtained by extracting all the reviews related to a product. This can be achieved from multiple sources such as Amazon, Flipkart, etc. All the reviews posted on these e-commerce platforms are stored in a database where we can apply our model and analyse the information in the reviews. The dataset will contain text form of data and emoticons. No other form of data such as images will be analysed through the model. The data set we used is available on kaggle.com. Data mining is then used on this dataset to produce a single substitute review [17].

II. Model Components

A) Information Extraction

The information is extracted from each and every review, spam reviews are removed and then remaining useful information is analysed altogether to converge towards a single idea representing the group of reviews.

- Text Pre-processing [2] & Spam Detection [1]

The processes involved in text pre-processing are.

Tokenization: - Every other review is split into meaningful words called tokens. Example - “Helium is the lightest gas” is converted to “Helium”, “is”, “the”, “lightest”, “gas”.

Data standardization: - It involves converting all words in the review in standard form, converting all words in lower case [12]. Example - “Keep Sodium away from Water” is converted to “keep sodium away from water”.

Emoji conversion: - The emoticons present in the reviews are assigned a keyword based on the expression they convey [13]. Based on the emojis encountered in the reviews, the emojis are replaced by the corresponding words in the ontology of emoji to names.

The emoticons are classified into following two categories: -

Positive emoticons- these are the emoticons which convey positive sentiment and are replaced by positive words with respect to the symbol.

:-) :-] :-} :-)) :-3 8-) :-D :o) :c) :^) 8-D X-D :-> x-D =D =3 B^D :) :] :} :-)) :3 8) :D 8D XD xD :> :-P :-p :P :p X-P x-p XP xp :-b :b d: =p >:P :L =L :S O:-) 0:-3 0:-) 0:^) O:) 0:3 0:) >:-)):-) 3:-) >;) >:)):) 3:)

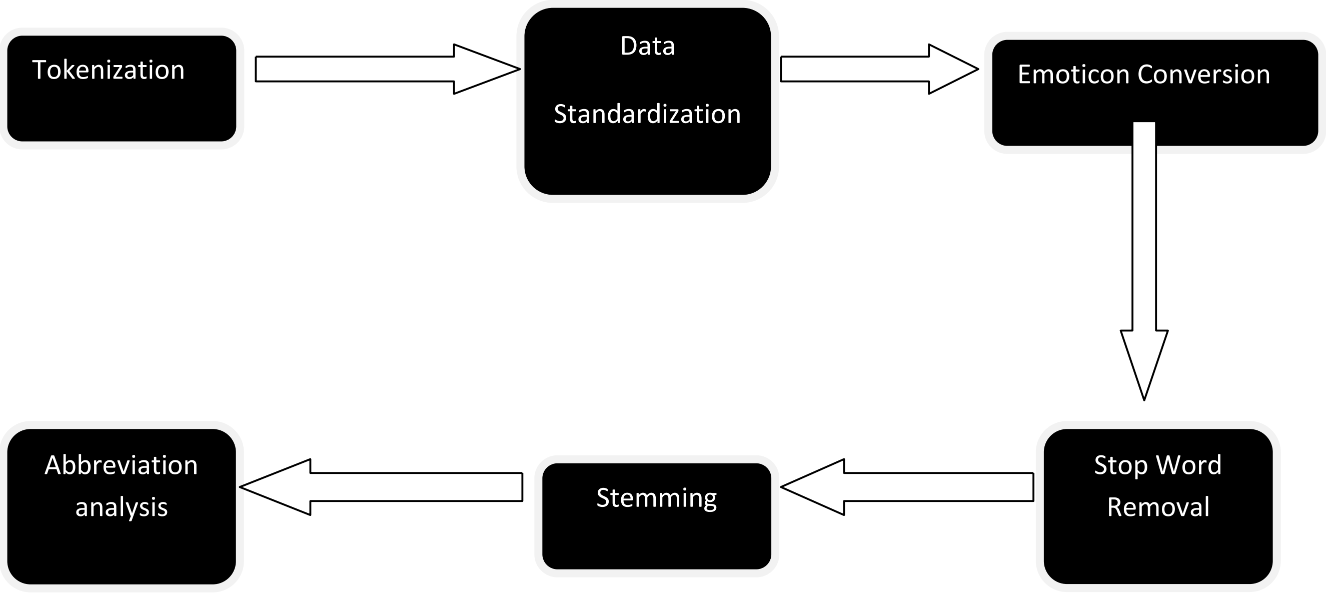
Negative emoticons - these emoticons reflect the sad or disturbed sentiments of the subject and are thus replaced by negative words.

:-(:-c :-< :-[:-|| :’-(:-/ :-. :(:c :< :[:’(
:/ :@ >:[:{ D-’: D:< D: D8 D; D= DX
>:/ :\ =/ =\ :L =L :-###.. :###..

Stop-word-removal: - All the words in the review which do not convey a special meaning are removed like a, the, then etc. [6]. These are irrelevant words in a review.

Stemming: - It involves obtaining the root word corresponding to every word by dropping suffixes ling -ing, -ion etc [3, 11], for example - “running” gives “run”.

Abbreviation analysis: - Replacing the abbreviations present in the review by their full forms. Example: LOL by Laugh Out Loud, UK by United Kingdom, etc.



The process of Text Pre-processing

- N-gram [7]

The next step after data pre-processing is N-gram features extraction. N-gram is a series of n tokens [3]. N-gram is a model very widely used in NLP tasks. The model creates N-grams from the reviews in the data set to extract keyword features from the data set. N-gram makes the text easier to process.

For n = 3, a sequence of three-words for each message is generated. Example - “less expenditure is more profit” is analysed as “less expenditure is” “expenditure is more”, ”is more profit” [8].

- Term Frequency [9]

The number of times a keyword occurs in each data sample is called its term frequency. Words having high frequency have a better relationship with the sample.

- Inverse Document Frequency (idf)

Idf factor is used to diminish the weight of the words that occur very often in the data set and to increase the weight of words that occur rarely [10]. It gives weights based on the frequency of occurrence of keywords.

- Spam Detection using RLOSD Model

To use Representation learning based opinion spam detection (RLOSD) Model for spam detection, we use feature reduction methods to extract informative features which help us in detecting deceptive reviews. In this model, in two successive phases of feature engineering and feature reduction, relevant features are elected and excessive and redundant terms are eliminated from the feature space. Then high frequency terms are chosen from the review and then it employs Principal Component Analysis (PCA) along with Modified Mutual Information (MMI) for finding significant and important terms. MMI considers all possible co-occurrences of a class label and PCA preserves data variance as much as possible. After feature reduction reviews are classified using decision tree methods which employ classifiers using Information Gain to rank features and detect spam reviews [1].

- Objects

The objects are the products which are posted on an e-commerce site for the purpose of selling and buying. Example- books, mobile phones, etc.

- Characteristics

The characteristics of an object are defined as the attributes of the product, which include all of the physical and internal properties of a product. For example- for a mobile phone it will be- hardware like screen, processor, RAM, memory etc.

- Characteristic index rating

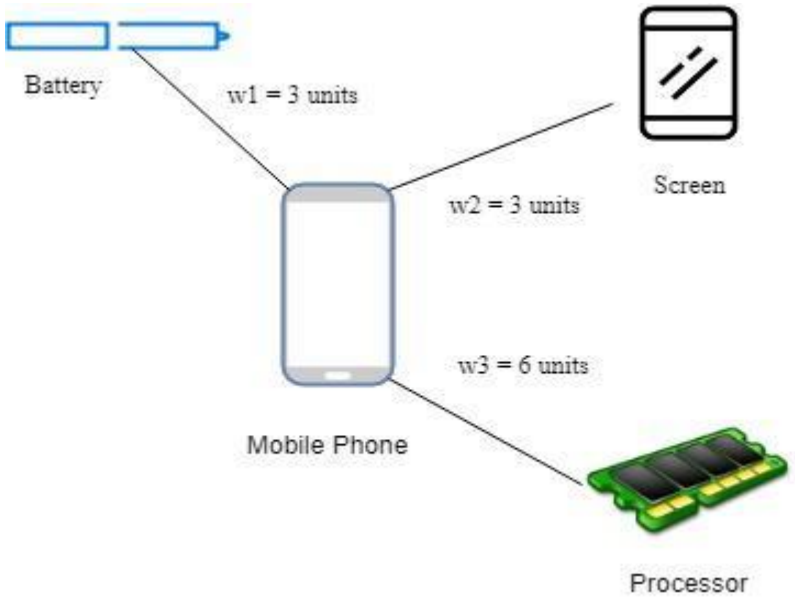
The index rating is defined in different levels starting from lowest level to the highest level. The levels are 1-7. There is ontology of words related to every level. For example: common, normal, general etc. All correspond to the word average and thus the value 4 is given corresponding to every review having same ontology of words.

Weight	**Meaning**
7	Excellent
6	Very good

5	Good
4	Average
3	Bad
2	Very bad
1	Worst

- **Characteristics Graph**

All the characteristics of an object are mapped in the form of a graph with the object as the central node. All the characteristics of the object are connected to the central node with edges carrying weight. The characteristics are also interconnected with each other. The weight of the edges of each attribute is given through characteristic indexing, based on the words used in the review. The weight is added after each and every review is analysed. For example- the screen and battery of the cell phone are bad. This gives the weight of the battery and screen attribute as 3 units. The processor of the mobile phone is very good, so the associated weight of this attribute will be 6 units.



Example Characteristic Graph of Mobile Phone

- **Resultant weight calculation**

The final weight is calculated by dividing the final weight of an attribute by the number of times it is updated in the database [14-16]. Based on the final value of the edges we can get the idea of the characteristic rating of the attribute of the corresponding object. The formula for resultant weight calculation is:

$$W = \sum w_i$$

- **Frequency estimation**

The formation of the final paragraph we require to estimate the frequency of the occurrence of the given characteristics in the review, so that, we can justify the more severe characteristics from the fewer important ones.

The value is given in the form of % of occurrence, which is given by the number of reviews carrying the given characteristic divided by a total number of valid reviews (non-spam reviews).

Occurrence %	**Meaning**
>75%	Most important and very common
>50%	Important and common
>25%	Less important and less common
>0%	Not worth mentioning

- **Paragraph formation**

Based on the Frequency estimation we mention only those characteristics in the resultant paragraph which form more than 25% of the total reviews. Based on the frequency estimation table, we use the ontology of words corresponding to the value of occurrence of characteristics. Now we use the ontology of words from the characteristic index rating table. By using these two different ontologies of words, we can form statements using appropriate connectives. For example: If 65% people review about the battery being bad then the statement formed will be ”In most cases, the battery performance of this device is poor”.

3. RESULT ANALYSIS

The result obtained from the proposed model gives the estimated characteristic prediction of the objects based on the reviews posted by the users. The resulting output can be used in many situations. The ambiguous reviews and semi incomplete reviews are processed and are thus used to predict all the characteristics of the object and their respective indexes. Therefore such predictive models can be used to successfully predict the most accurate summary of a group of reviews. The result of the given methodology is a single paragraph containing the essence of all the important reviews published by the consumers or users. Based on thorough analysis of all the reviews we can accommodate all the reviews containing the highest priorities and cite them in the paragraph that we supposedly form. We form the paragraphs based on the ontology of words used in analysing the objects and their respective characteristics. One of the major aspects is to accept reviews based on the frequency of the characteristics, thus assigning them their level of importance.

Future Scope

The proposed model can be used in situations where information extraction is required to achieve the desired result and use it for various different purposes such as product comparison, most relevant service choice etc.

The e-commerce websites can use this methodology to provide users with unipolar, useful and accurate results using all the characteristics associated with the product. Businesses are very interested in understanding the thoughts of people and how they are responding to all the products and services around them. Companies use sentiment analysis to evaluate their advertisement campaigns and to improve their products. Companies aim to use such sentiment analysis tools in the areas of customer feedback, marketing, CRM, and e-commerce.

This methodological model coupled with other social models can be used for further research areas of stress management by using social platform chat or comment data and then forming a single substitute for that data which can be analysed to find the stress level in a person's life. It can even be used in extrapolating the poll results by analysing a single substitute review formed by all the reviews given by people to political parties. We have suggested a way of meeting the requirements of both, businesses as well as people, which pave a way for further study and research in this area.

4. CONCLUSIONS

The aim here was to create a single meaning paragraph out of a pile of reviews which may show traits like ambiguity, incompleteness, unreadability, etc. using simple data mining techniques. The proposed model takes input from the data set created by accumulating all the text reviews (including emoticons) given by the users. All the reviews may be from various e-commerce platforms such as Amazon, Flipkart, Myntraetc. Nextcomes detecting and removing spam reviews leaving only genuine reviews.

The messages are then pre-processed to obtain the keywords from the data sets. After pre-processing we use probabilistic language models like n-gram. Associating weights to the data set using Term Frequency-Inverse Document Frequency (tf-idf) increases the overall efficiency of predicting algorithms.

The next step is to define indexes which can be used to define the level of fulfilment of expectations of the users. Based on the defined indexes we can paraphrase user satisfaction regarding the characteristics of a product.Emoticons are very common tokens in any review in today's socio-techno world, therefore we must also focus on efficient ways to analyse them. We have converted emoticons to a textual form for our computation processes.

Thus, we propose to give a highly efficient method of finding the most accurate single review of the product by analysing all the available reviews and also processing emoticons of all the reviews posted by the users. As it is not possible to read each and every review, so this single substituted review paragraph will help in more accurate assessment of the given services or products. Thus this model is a requirement and a business strategy in the modern world.

References

[1] Z. Sedighi, H. Ebrahimpour-Komleh and A. Bagheri, RLOSD: Representation learning based opinion spam detection, 2017 3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS), Shahrood, 2017, pp. 74-80

[2] S. Vijayarani et al, (2015). Preprocessing Techniques for Text Mining - An Overview, *International Journal of Computer Science & Communication Networks*, Vol. 5(1),7-16

[3] Hull, D. A., et al. (1996). Stemming algorithms: A case study for detailed evaluation. *JASIS*, 47(1), 70–84.

[4] Pak, Alexander & Paroubek, Patrick. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proceedings of LREC*. 10.

[5] Tanimu Jibril, Ahmed &Hayati Abdullah, Mardziah. (2013). Relevance of Emoticons in Computer-Mediated Communication Contexts: An Overview. *Asian Social Science* 9. 10.5539/ass.v9n4p201.

[6] C. Silva and B. Ribeiro, The importance of stop word removal on recall values in text categorization, Proceedings of the International Joint Conference on Neural Networks, 2003. Portland, OR, 2003, vol. 3. pp. 1661-1666.

[7] Peter F. Brown, Peter V. de Souza, Robert L. Mercer, Vincent J. Della Pietra, Jenifer C. Lai, Class-based *n*-gram models of natural language, *Computational Linguistics*, v. 18 n. 4, p. 467-479, December 1992

[8] D. Lyon and B. Cedex, N-grams based feature selection and text representation for Chinese Text Classification Zhihua WEI, *Int. J. Comput. Intell. Syst*., vol. 2, no. 4, pp. 365–374, 2009.

[9] Trstenjak, Bruno, Sasa Mikac, and Dzenana Donko. KNN with TF-IDF based Framework for Text Categorization. *Procedia Engineering* 69 (2014) 1356-1364.

[10] Cong, Yingnan, Yao-ban Chan, and Mark A. Ragan. A Novel Alignment-Free Method for Detection of Lateral Genetic Transfer Based on TF-IDF. *Scientific Reports* 6 (2016): 30308. PMC. Web. 12 Oct. 2017.

[11] Julie B Lovins. 1968. Development of a stemming algorithm. MIT Information Processing Group, Electronic Systems Laboratory.

[12] Mohamad, Ismail Bin, and Dauda Usman. Standardization and its effects on K-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology* 6.17 (2013) 3299-3303.

[13] Thompson, Dominic, and Ruth Filik. Sarcasm in written communication: Emoticons are efficient markers of intention. *Journal of Computer-Mediated Communication* 21.2 (2016) 105-120.

[14] Jin Lianjing, et al. A Text Classifier of English Movie Reviews Based on Information Gain. Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence (ACIT-CSI), 2015 3rd International Conference on. *IEEE*, 2015.

[15] Salton and Buckley, Term Weighting Approaches in Automatic Text Retrieval, *Inf. Process. Manag.* Vol. 24(5), p. 513–523., 1988.

[16] Liu, B. Sentiment Analysis and Opinion Mining, p. 7. Morgan and Claypool Publishers, USA (2012)

[17] Ming-Syan Chen, Jiawei Han, and Philip S. Yu. 1996. Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering* 8, 6 (1996) 866–883.

[18] Ramya, R. S., et al. Feature Extraction and Duplicate Detection for Text Mining: A Survey. *Global Journal of Computer Science and Technology* 16.5 (2017).

[19] Ahmad, Sartaj & Varma, Rishabh. (2018). Information extraction from text messages using data mining techniques. *Malaya Journal of Matematik.* S. 26-29. 10.26637/MJM0S01/05.