ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2020.0294

STUDENT FEEDBACK SENTIMENT ANALYSIS SYSTEM FOR DISTANCE EDUCATION USING ARM WITH K-MEANS CLUSTERING

G.N. Harshini and N. Gobi

Department of Computer Science and Engineering, Dr. Mahalingam College of Engineering and Technology, India

Abstract

The rapid development of Internet has resulted in the boom of evaluations about products and services. For extracting the aspects and determining the opinions from reviews Sentiment Analysis is used. The main challenges faced by Sentiment Analysis system is that, in order to increase or decrease the market value of the product the spammers may post irrelevant or fake reviews and another challenge deals with the classification of both Implicit and Explicit features present among the review sentences in the dataset. The proposed system deals with the identification of fake reviews through fake review Indicators which help in removing the fake reviews. For the better identification of both implicit and explicit features, association rule mining with K-Means Clustering is used. Lexicon method is used for the classification of sentiments into positive and negative polarities. The advantage of the proposed system is that the fake reviews can be detected and eliminated in the dataset and both implicit and explicit attribute extraction from the review sentence can be identified along with its polarities through Lexicon based Method.

Keywords:

Lexicon Based Method, Sentiment Analysis, Spammers, K-Means Clustering

1. INTRODUCTION

Sentiment Analysis is the application of Natural Language Processing and Computation Analysis to identify and extract subjective information through the sources such as blogs, text or standard corpus. It deals with analysing the emotion of a person from a given piece of text written by them. There is a huge demand for sentimental analysis. Opinions, feelings, and emotions can be captured from individual writings, facial expressions, ratings, speech and many other media. People can know about the quality of the product and services through the reviews and ratings given in the websites. It is not only useful for the customers but also to the companies to know about people's opinion on their products and to analyse how successful the project was [12]. The three important classification in sentimental analysis are sentiment classification, feature level identification, opinion summarization. Sentiment classification deals with opinion classification of reviews such as positive and negative opinion towards certain objects. Feature based classification deals with classifying reviews based on features of certain objects. Opinion Summarization is the process used to automatically summarize many opinions that are relevant to the same topic. There are three types in opinion mining namely, document level, sentence level, and aspect level [13].

Document Level: It identifies the document that expresses opinions and classifies the opinion as positive, negative and neutral [14].

Sentence Level: It identifies the sentences that expresses opinions and classifies the opinion as positive, negative and neutral [15].

Phrase Level/*Aspect Level*: It identifies the phrases that are subjected to opinion and opinion orientations.

Machine learning algorithms and lexicon based approaches are mostly used in sentiment analysis. Reviewers may write the reviews about the products either directly by specifying the aspects and their opinion or indirectly through synonyms or phrases about the aspects and its opinion words. Aspect Level SA system classifies the reviews based on the opinion orientation of the aspects namely Explicit Reviews and Implicit Reviews. If the reviewer talks about the positive, negative or neutral aspects of the products directly, then these kinds of reviews are termed as explicit reviews. If the opinion holder indirectly expresses the polarity, then such reviews are known as implicit reviews [16].

2. LITERATURE REVIEW

Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to service. The process of sentiment analysis is to extract the features and identify opinions associated with features. A system that considers temporal sentiment and emotion analysis of multi-lingual student feedback system is developed to classify emotion into anger, anticipation, joy, fear, sadness, surprise, disgust, and trust. Emotion lexicon is used to classify the sentiments into positive and negative and eight basic emotions [1]. Multilingual analysis system which is capable of analyzing different languages is an advantage.

Aspect level sentiment analysis is done to extract the nouns and adverbs from the POS tagging [2]. Sentiment analysis of reviews can be done using Stanford NLP library which contains the algorithm part to process the sentiment analysis to classify the positive and negative percentage of reviews. It extracts the Fine grained sentiment information which helps in analyzing intrinsic aspects.

An effective fake review system which uses a frequent item set mining to find a set of candidate groups [3] is developed. Group spam behaviour indicators and individual spam indicators are used to identify the spam reviews. Group time window, group deviation, group content similarity and group size features are some of the used indicators to identify the group spam reviews. A novel relational model which considers relationship among groups is proposed.

A fake review system which focuses on variety of spam indicators based on product level which is relative to both reviewer and review [4]. Burst pattern recovery technique is used to identify the suspicious time intervals and reviews. It focuses

only on group spam activities which should be extended for finding individual spam reviews also.

To overcome the limitations of K-means clustering, an improved K-means algorithm is developed along with Association Rule Mining which focuses on seeking the minimum rules set and minimum rules covering set [5]. K-Means with Association rule mining algorithm identifies the keywords that are shown frequently in the document and assumes that it must be similar.

A novel two phase co-occurrence Association Rule mining is used in the identification of implicit reviews. A count matrix that satisfies the co-occurrences between the explicit features and opinion words is developed [6]. Rule generation and rule application are done using count matrix. Implicit features are identified by clustering the explicit feature significant rules.

A fake review identification system which focuses on user centric and review centric features is developed [7]. User centric features are divided into four types, namely personal features, social features, trusting features and reviewing activity features. Review centric features focuses only on the text of the review. Better results are obtained when the user centric features are divided into four subsets.

To detect the spam reviews a new technique which focuses on the behavior features of the spammer is developed [8]. Two algorithms are designed to implement the indicators used in finding the spam reviews. The algorithms are used to identify the review similarity between two reviews. The algorithm achieves higher efficiency than the traditional algorithm.

For the cross domain sentiment classification of reviews, a new technique called word alignment based on association rules is proposed which is used to identify the relationship between domain specific and domain shared words [9]. An automated system which identifies the explicit suggestions from the reviews is developed. Data pre-processing, explicit suggestion extraction and visualization is done [10].

3. SYSTEM ARCHITECUTRE

The dataset is obtained from [11] kaggle website with respect to the feedback given by the alumni of the corresponding courses. The reviews in the dataset consist of input text which expresses the opinions of the students about the courses and teachers. The collected dataset may include irrelevant and fake reviews which should be removed by combining review and reviewer characteristics techniques. In order to reduce the size of the dataset, pre-processing is done after the removal of fake reviews. Pre-processing phase includes stop word removal, tokenization and parts of speech tagging. From the pre-processed data, Features are extracted using ARM with K-means clustering. For the identification of sentiment against the features identified from the previous stage, Lexicon method, which classifies the sentiments into positive and negative, is used. The overall process of the proposed system is shown in Fig.1. In the existing system, for the classification of reviews, Association rule mining, which lacks the accuracy of predicting Implicit and Explicit reviews, is used. To overcome the limitations, in the proposed system, Association Rule Mining with K-Means clustering is used.

3.1 FAKE REVIEW INDICATORS

The proposed work of fake review detection is based on,

- Fake review detection by combining number of reviews and reviewers focused techniques.
- Determining reputation of Author through their past history of reviewing process.

Some of the methods used to find fake reviews are

3.1.1 Review Relevancy Rate:

There will be reviews in the dataset that are not needed such as advertisement or link. Review Relevance Rate refers to the degree of relevance that exists between the content of the review and the subject of the product. The formula used in finding review relevancy rate is

$$RRR(r) = e^{(W(s)\cap W(r)/W(s))-1} \quad (1)$$

where $W(s)$ is the set of all segmented words of the product's topic, and $W(r)$ is the set of all segmented words of a review.

3.1.2 Content-Length:

When the review content is short, it shows that the reviewer did not consider it seriously. This kind of reviews do not make much sense for review data analysis.

$$CL(r) = \begin{cases} 1 & r \cdot length \le \lambda \\ 0 & r \cdot length > \lambda \end{cases} \quad (2)$$

where r - length denotes the length of the review r, and λ is a threshold to judge the effectiveness of the length of the review content.

3.1.3 Review Content Similarity:

The most useful technique for detecting fake online reviews is examining the similarity among author's reviews. Spammers tend to reproduce the same written content across their reviews for multiple products because of time constraints, and, sometimes, more sophisticated spammers, who actually make efforts to change the content, tend to use similar vocabulary every time. The review similarity rate between two reviews can be written as:

$$tf_{ij} = f_{ij}/Size \quad (3)$$

3.1.4 Measuring Spam Score:

Overall reviews Spamicity is determined based on results from the combination of several separate reviews scores, author's actions and credibility based on their past activity. Each score is multiplied by an appropriate weight according to its importance, as represented in Table.1.

Table.1. Weights of Spamicity Influencing Factors

Spamicity affecting factors	Purpose	Weight
$CS(a)$	Content similarity	1.5
$RRR(r)$	Review Relevancy Rate	0.25
$CL(r)$	Content Length	0.25

After applying the score for each factor, the spamicity score can be calculated by using the equation,

$$S(r) = 1.5\ CS(a) + 0.25\ RRR(r) + 0.25\ CL(r) \quad (4)$$

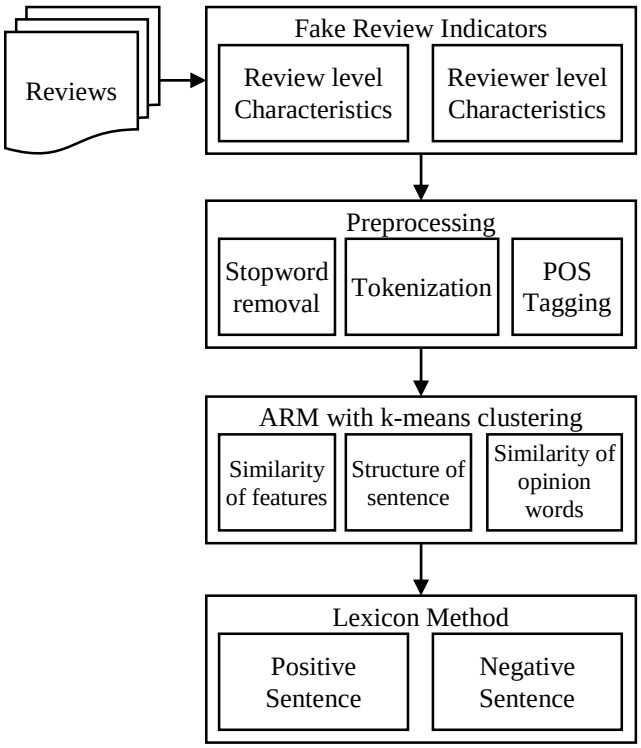

Fig.1. Architecture of proposed system

The spamicity score can be calculated by fixing a threshold. If the spamicity score exceeds 25, then it can be considered as fake review.

3.2 PREPROCESSING

Preprocessing is done to remove the size of the dataset. Stopword removal, Tokenization, and POS tagging is done in pre-processing phase. In the Stopword removal phase the unwanted text like articles, prepositions present in the Review sentences are removed. Tokenization is a process in which collection of sentences in a text document are divided into token by removing white space, comma and other symbols etc. The documents are parsed using Stanford parser through which the system assigns Parts-Of-Speech (POS) tags to every token or word present in the filtered sentence. POS tags are useful to identify the grammatical structure of sentences such as noun, verb, adverb and adjective phrase and their relationship.

Table.2. Output of Pre-processing

Original data	Stop words removal	Tokenization	POS tagging
This is an excellent novel worth to read	Excellent novel worth read	Excellent Novel Worth Read	Adjective Noun Adjective Verb
One of the most useful course on IT management	Useful course IT management	Useful Course IT Management	Adjective Noun Noun Noun

3.3 ASSOSCIATION RULE MINING WITH K-MEANS CLUSTERING

K-Means clustering is a prototype based clustering technique. It is used to identify the features which have no predefined class labels but groups feature using the similarity measures between them. It places most similar features into one class and dissimilar into another class of features. The purpose of clustering is to assign a cluster to each data point based upon its Euclidean distance. In sentiment analysis system K-Means is used to group the features into groups based on their high similarity. Clustering is based on following three observations:

Step 1: Initially it considers the opinion words' similarity which is useful to guide the clustering.

Step 2: It considers the similarity of features in the review sentence. It is used to identify the aspects with the same meaning.

Step 3: It considers the structure of the feature in the comment. It identifies the type of the feature, like whether it is Noun (N), Verb (V), Noun + Verb (NV), Verb + Noun (VN), Noun (N). Based on the inputs provided by the ARM, the clustering model automatically builds the model that can cluster a related class of objects in order to predict the value of the missing attribute.

Initially, to build a cluster, a set of opinion words are placed inside the tuple, where each tuple holds a set of opinion words. This means clustering is used to check whether the opinion words in the review has any of the matching words in the clusters. If a match is found, then it determines the implicit feature for its corresponding opinion word, otherwise, it fires the ARM to predict the implicit feature. A set of association rules are fired for each tuple and the mean confidence score are computed for each set of rules. The finally selected opinion words and implicit feature will be stored in the respective cluster for future references.

3.4 LEXICON METHOD

The lexicon-based approach is used for sentiment analysis of words. The predefined dictionary is used which calculates the semantic score for each word that is used for the classification of words. If the score is in positive, it is classified as positive, and, if the word contains negative score, then it is classified as negative.

K-Means is considered to be more convenient than other methods for the database used in the proposed system since it is simple to implement and consumes less time to process the inputs.

4. RESULTS AND DISCUSSIONS

This section discusses the expected results. Precision, recall and F-Measure are some of the accuracy measures used to evaluate the proposed system's cluster based association rule mining with K-means clustering.

4.1 DATA DESCRIPTION

The dataset is accessed from kaggle. The dataset includes the students' feedback about the class, course, and the staff in distance education. The dataset includes course id, course name, and review of the student. Reviews for each subject is given by

the students. Distance Education College and its courses are reviewed by students.

4.2 EVALUATION METRICS

Several evaluation factors are used in identifying accuracy of the proposed system. Evaluation metrics can be used for classifying the positive and negative words. The terms used in evaluation metrics are shown in Table.3.

Table.3. Factors for evaluation

Factors	Description
True Positive	No of extracted suggestions which are relevant (correct).
False Positive	No of extracted suggestions which are irrelevant (wrong).
True Negative	No of correct suggestions which are not extracted.
False Negative	No of irrelevant suggestions which are not extracted.

Precision: Precision is the ratio of number of correctly extracted suggestions to the sum of correctly extracted suggestions and wrongly extracted suggestions.

Recall: Recall is the ratio of number of irrelevant suggestions extracted to the sum of irrelevant suggestions and relevant suggestions.

F-Measure: The harmonic mean of precision and recall is known as F-Measure.

Human evaluations are conducted against the proposed system and the results shows that the proposed system performs equivalent to the human evaluated system on the sample of 100 reviews.

4.3 RESULTS

The *f*-score of the proposed system has better values than the existing and it can be inferred from the Fig.2 – Fig.4 that the mean of precision and recall values are higher. The classification of implicit and explicit review contributes to the true positive, thus the two metrics precision and recall have gained more score.

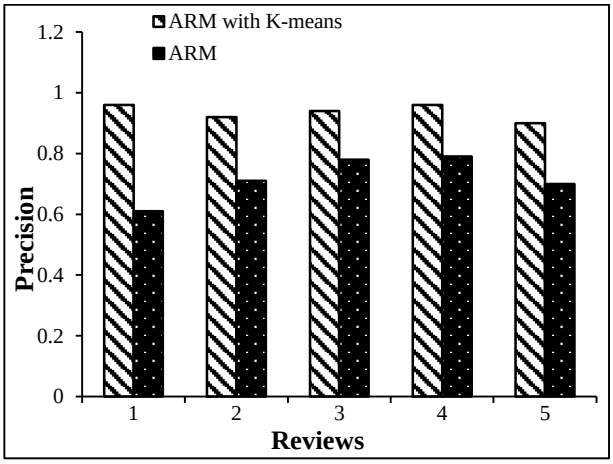

Fig.2. Evaluated Precision

Thus it shows that the proposed approach has identified and eliminated fake reviews in the dataset and also the classification of reviews into both explicit and implicit suggestion is done better than the existing system.

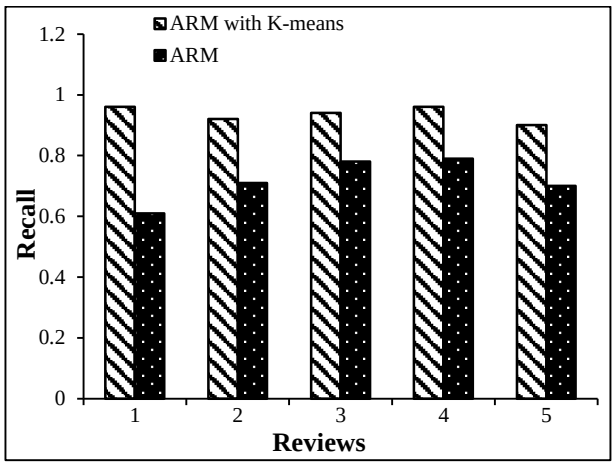

Fig.3. Evaluated Recall

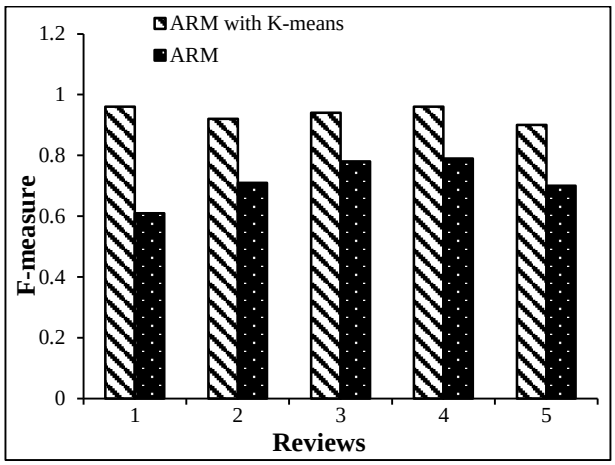

Fig.4. Evaluated F-Measure

5. CONCLUSION

In this paper, fake review indicators are used to identify the fake reviews present in the dataset. Association rule mining K-means clustering is used for identifying the implicit and explicit reviews present in the dataset. The last method in the system is the Lexicon method, which is used to conduct the sentiment analysis of the words as positive and negative. It is expected that, by using fake review indicators, the fake reviews which are irrelevant to the analysis are removed and also the classification of explicit and implicit reviews can be done efficiently.

REFERENCES

[1] Sujata Rani and Parteek Kumar, "A Sentiment Analysis System to Improve Teaching and Learning", *Computer*, Vol. 50, No. 5, pp. 36-43, 2017.

[2] Luxchippiriya Balachandran and Abarnah Kirupananda, "Online Reviews Evaluation System for Higher Education Institution: An Aspect Based Sentiment Analysis Tool", *Proceedings of International Conference on Software,*

Knowledge, Information Management and Applications, pp. 1-7, 2017.

[3] Arun Mukherjee, Bing Liu and Natalie Glance, "Spotting Fake Review Groups in Consumer Reviews", *Proceedings of 21st International Conference on World Wide Web*, pp. 191-200, 2012.

[4] Ioannis Dematis, Eirini Karapistoli and Athena Vakali, "Fake Review Detection via Exploitation of Spam Indicators and Reviewer Behavior Characteristics", *Proceedings of International Conference on Current Trends in Theory and Practice of Informatics*, pp. 581-595, 2017.

[5] Gang Liu, Wray Buntine, Weiping Fu and Yudan Du, "An Association Rules Text Mining Algorithm Fusion with K-Means Improvement", *Proceedings of International Conference on Computer Science and Network Technology*, pp. 781-785, 2015.

[6] Zhen Hai, Kuiyu Chang and Jung Jae Kim, "Implicit Feature Identification via Co-occurrence Association Rule Mining", *Proceedings of International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 393-404, 2011.

[7] R. Barbado, O. Araque and C.A. Iglesias, "A Framework for Identifying Fake Review Detection in Online Consumer Electronics Retailers", *Information Processing and Management*, Vol. 56, No. 4, pp. 1234-1244, 2019.

[8] Neha S. Chowdhary and Anala A. Pandit, "Fake Review Detection using Classification", *International Journal of Computer Applications*, Vol. 180, No. 50, pp. 975-987, 2018.

[9] Xi Bin Jia, Ya Jin, Ning Li, Xing Su, Barry Cardiff and Bir Bhanu, "Words Alignment based on Association Rules for Cross Domain Sentiment Classification", *Frontiers of Information Technology and Electronic Engineering*, Vol. 19, No. 2, pp. 260-272, 2018.

[10] Swapna Gottipati, Venky Shankararaman and Jeff Rongsheng Lin, "Text Analytics Approach to Extract Course Improvement Suggestions from Students Feedback", *Research and Practice in Technology Enhanced Learning*, Vol. 13, No. 6, pp. 1-19, 2018.

[11] Rakibul Hassan and Md. Rabiul Islam, "Detection of Fake Online Reviews using Semi-Supervised and Supervised Learning", *Proceedings of International Conference on Electrical, Computer and Communication Engineering*, pp. 1-5, 2019

[12] Jenifer Jothi Mary, S. Santiago and L. Arockiam, "A Methodological Framework to Identify the Students Opinion using Aspect based Sentiment Analysis", *International Journal of Engineering Research and Technology*, Vol. 5, No. 2, pp. 642-645, 2018.

[13] Vasileios Kagklis, Anthi Karatrantou, Maria Tantoula, Chris T. Panagiotakopoulos and Vassilios S. Verykios, "A Learning Analytics Methodology for Detecting Sentiment in Student: A Case Study in Distance Education", *European Journal of Open, Distance and E-Learning*, Vol. 18, No. 2, pp. 74-94, 2015.

[14] Xinyue Wang, Xianguo Zhang, Chengzhi Jiang and Haihang Liu, "Identification of Fake Reviews using Semantic and Behavioural Feature", *Proceedings of International Conference on Information Management*, pp. 92-97, 2018.

[15] K.C. Ravi Kumar, D. Teja Santosh and B. Vishnu Vardhan, "Extracting Opinion Targets from Product Reviews using Comprehensive Feature Extraction Model in Opinion Mining", *Indian Journal of Science and Technology*, Vol. 10, No. 21, pp. 1-6, 2017.

[16] Z. Kamisli Ozturk, Z.I Erzurum Cicek and Z. Ergul, "Sentiment Analysis: An Application to Anadolu University", *Proceedings of International Conference on Computational and Experimental Science and Engineering*, pp. 752-755, 2017.