Text Mining of VOOT Application Reviews on Google Play Store

Swathi Yadav[1], Shwetha Yadav[2]

[1,2] PG student, Thakur College of Engineering, Mumbai

ABSTRACT: *This survey paper is to classify the vast amount of Voot Application reviews present on google play store is carried out by using text mining. Vocabularies including nice, best, good, well, satisfactory can be classified into the good reviews for this application. And those vocabularies including bad, worst, stupid, slow, time consuming can be classified into the bad reviews. As we classify the reviews into good and bad, the more amount of bad reviews will redirect us that the application needs further improvement. The objective of this paper is to classify the reviews into good and bad. This paper outlined a structured approach of text analysis and for classifying the reviews we will use classification algorithm.*

Keywords: Classification algorithm, Google playstore, Machine learning, Reviews, Support Vector Machine(SVM), Text data mining, Voot Application,

1. INTRODUCTION

In the previous decades the PC equipment innovation has turned out to be capable. This has supported up the database and data industry. Thus a substantial number of databases and data vaults are accessible and the associations put away a lot of information. This has expanded the requirement for capable information investigation which is unrealistic without intense instruments. Information mining devices dissect information from alternate points of view and condense the outcomes as valuable data. They are utilized to work on a lot of information to discover covered up examples and affiliations that can be useful in choice making.

Resent investigation on data mining on observing the reviews on playstore or any online networking goes for learning and example extraction from enormous gathered database is expanding. In addition mining such data is confusing. The information course of action and retrival of such content parts ends up plainly troublesome in light of the fact that they are frequently portrayed in a free format. As of late, interest has increased in text mining since it reveals valuable learning covered in a lot of aggregated documents.Research has begun to apply text mining in numerous regions. For instance, mining text in drug for breaking down patient history, Mining text in gathering reviews.

Coming to our research, google play store is the most used application for downloading apps. There is one benefit in google play store that we can see the reviews before downloading any application. But this is challenging for the app owner because, there will be bad reviews which can harm the reputation of an organization. If they earn a bad reputation, its going to stick with them throughout. That's why, the businesses whether the organization is large or small they are worried about their digital footprint. So we are, retrieving the reviews of the Voot application using text mining which is present in the google play store. And after that using machine learning algorithm the classification of that reviews is performed. To group the immense measure of reviews that are in google play store for voot application , text mining is done utilizing SVM algorithm

2. VOOT APPLICATION

Voot is the free streaming services featuring TV shows produced by Viacom's Indian channels. What it lacks in international content it makes up for with numerous shows and movies in regional languages. The kids section offers optional parental controls. It was created for the India's favourite reality TV shows including Bigg Boss, Splitsvilla, Roadies and more.

3. TEXT MINING

A text analysis issue generally comprises of three critical advances: parsing, search and retrieval, and text mining.

Parsing :

Is the procedure that takes unstructured content and forces a structure for assist investigation. The unstructured content could be a plain content record, a weblog, an Extensible Markup Language (XML) document, a HyperText Markup Language (HTML) document, or a Word report. Parsing deconstructs the given content and renders it in a more organized manner for the resulting steps.

Search and retrieval :

Is the recognizable proof of the archives in a corpus that contain search lists, for example, particular words, expressions, points, or substances like individuals or associations. These search list are for the most part called key terms. Search and retrieval began from the field of library science and is presently utilized widely by web crawlers.

Text mining :

Text mining includes mining through a content record or asset to get significant organized data. This requires modern logical tools that procedure message so as to gather particular catchphrases or key information focuses from what are considered generally crude or unstructured formats.In text mining, built frameworks utilize things like scientific categorizations and lexical investigation to figure out what parts of a content report are important as mined information. Factual models are usually valuable, and frameworks may likewise utilize heuristics, or algorithmic mystery, to endeavor to figure out which parts of a content are essential. Other control frameworks incorporate labeling and catchphrase examination, where tools search for particular formal people, places or things or different labels and key words to make sense of what is being composed about. Text mining includes mining through a content record or asset to get significant organized data. This requires modern logical tools that procedure message so as to gather particular catchphrases or key information focuses from what are considered generally crude or unstructured formats.In text mining, built frameworks utilize things like scientific categorizations and lexical investigation to figure out what parts of a content report are important as mined information. Factual models are usually valuable, and frameworks may likewise utilize heuristics, or algorithmic mystery, to endeavor to figure out which parts of a content are essential. Other control frameworks incorporate labeling and catchphrase examination, where tools search for particular formal people, places or things or different labels and key words to make sense of what is being composed about.

4. LITERATURE REVIEW

S. No.	Paper Nmae	Author Name	Description
1.	Social Media Mining To Analyse Students' Learning Experience	Ms. S. Aswini, Dr. Ilango Krishnamoorthy	We concentrated on students presents on comprehend issues and issues in their education experience. Substantial work load, absence of awareness of social activities, and restlessness are a few issues that students face as they experience circular activities. In light of these outcomes, we began to execute a multi-label classification algorithm to arrange posts reflecting students' issues.
2	A review on text mining	Yu, Zhang, Men	This paper introduces the research status of text mining. Then several general models are described to know text mining in the overall perspective. At last we classify text mining work as text categorization, text clustering, association rule extraction and trend analysis according to applications.
3	Mining online reviews in Indonesia's priority tourist destinations using sentiment analysis and text summarization approach	Puteri Prameswari; Zulkarnain; Isti Surjandari; Enrico Laoh	The main contribution of this research is to combine two techniques in text mining that have never been done before, namely the sentiment analysis and text summarization.

5. PROPOSED METHOD

Going to our exploration, google play store is the most utilized application for downloading applications. There is one advantage in google play store that we can see the reviews previously downloading any application. Yet, this is trying for the application proprietor on the grounds that, there will be awful audits which can hurt the notoriety of an association. On the off chance that they procure a terrible notoriety, it will stay with them all through. That is the reason, the organizations whether the association is substantial or little they are stressed over their computerized impression. So we are, retrieving the reviews of the Voot application utilizing text mining which is available in the google play store. Also, after that utilizing machine learning calculation the order of that surveys is performed. To aggregate the colossal measure of audits that are in google play store for voot application ,text mining is done using SVM algorithm.

5.1.ARCHITECTURE

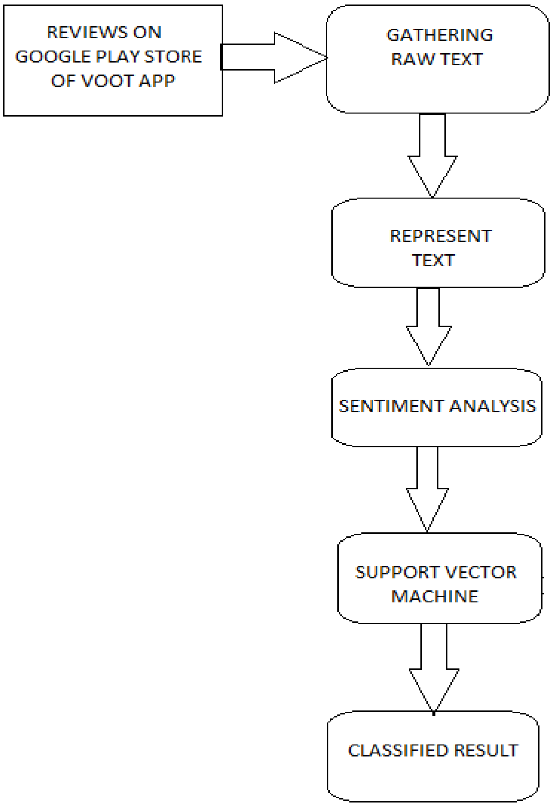

Fig1 : Proposed architecture

5.1.1 GATHERING RAW TEXT

In this step ,we are collecting the raw text from google play store for the reference to voot application.By keeping the application name as the keyword the raw texts are gathered. Generally these raw texts are present on google playstore and it can be easily retrived. The raw text may contain the keyword VOOT . In google play store, from where we download any application we get to see the reviews made by the users. These reviews may or may not contain the keyword 'voot' but then to they are considered as the raw text.

5.1.2 REPRESENT TEXT

After the previous step, we now has some crude content to begin with. In this step, crude content is first changed with content standardization procedures, for example, tokenization and case folding. Now after performing the above techniques the text we get is in more structured format.

Tokenization

Is the task of isolating (additionally called tokenizing) words from the raw text. Raw text is changed over into collection of tokens after the tokenization, where every token is a word. For eg if the text is gud, good will be considered in the same token as good.

Case folding

Is the technique in which all the upper case in a text are converteted into lower cases. But if the words like WHO, General Motors will be tokenized as who, general and motors due to this the meaning of the text would obviously change. So to avoid this issue look up table is generated where those texts are stored which shouldn't be case folded.

5.1.3 SENTIMENT ANALYSIS

After representing text the main goal is to analyze the sentiments of the text. Sentiments are basically the emotions related to the contents. Emotions can be positive or negative. So to analyse whether the emotions are positive or negative sentiment analysis is done. The review may contain positive value or the negative value for the voot application. So by performing sentiment analysis these values are categorized.

5.1.4 SUPPORT VECTOR MACHINE

Support vector machine is the concept of Machine learning which is used for classification of instances. SVM is used to analyse the instances and classify those instances into their respective classes. One of the advantages of SVM is that it allows miss-classification of some of the instances. Miss-classification is nothing but those instances which are wrongly classified. For this problem the SVM introduce the concept of margin. These miss-classified instances are support vectors. And with the help of these support vectors margins are made.

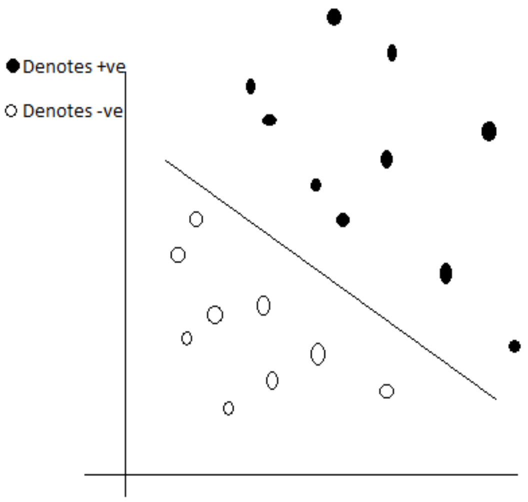

Fig2: classification of linear separable instanc

If the training instances are linearly separable as shown in the above fig2 , then there can be multiple number of classifiers.

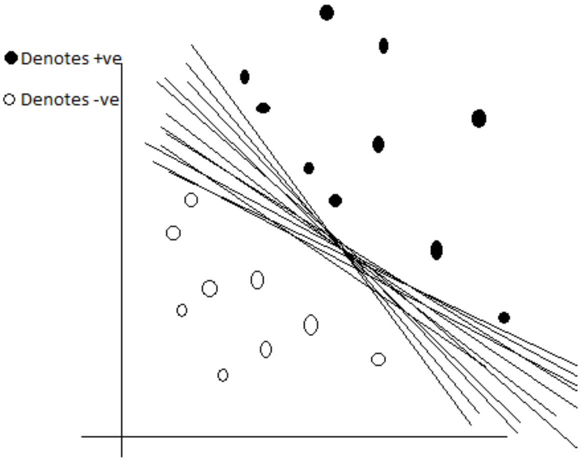

Fig3 : Multiple no. of classifiers for linearly separable instances

If some of the instances belongs to the different class and if it is wrongly classified then miss-classification of that instance is occurring.

The following figure4 depicts the miss-classification of the positive and the negative instance that are falling in different class.

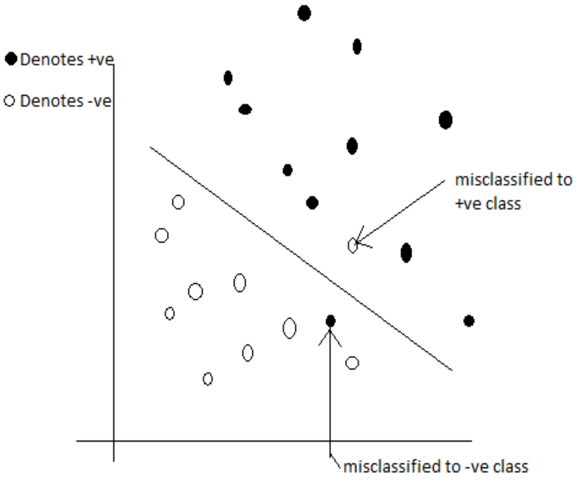

Fig4: miss-classification of training instances

Miss-classification of the instances will lead to problem so SVM is used because SVM allows miss-classificationas depicted in fig5.

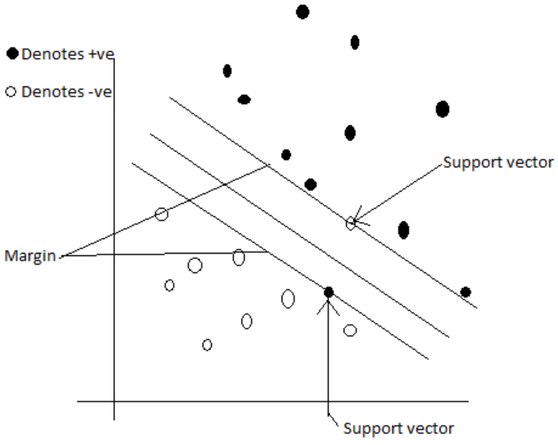

Fig5: Support vector machine

SVM is also called as margin classifier because it takes help of margin to avoid miss-classification.SVM reduces multiple number of binary classifiers by introducing the margin concept.

5.1.5 MATHEMATICAL MODEL

SVM has three main components they are decision boundary , support vectors and margin.

T=w.x ....(decision boundary)……(1)

where T is the some threshold according to which input instances are classified to class {+ve , -ve}

T_i=w.x_i-m ...(for input instance x_i classified to class -ve)..(2) where T_i<T

T_i=w.x_i+m ..(for input instance x_i classified to class +ve)..(3) where T_i>T

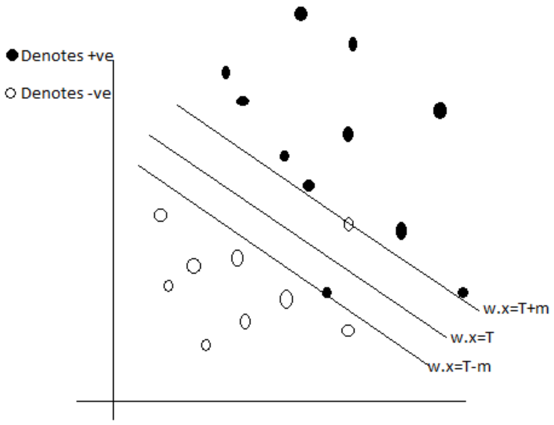

Fig 6: SVM geometry

Case 1: If SVM classifies the input instance x_i to positive class then according to the algorithm

$$T_i=w.x_i+m$$

Let m=1 then $T_i=w.x_i+1$,

T_i-t= $w.x_i+1-w.x=1$ ....$w.x_i$ and w.x is almost same

If actual value of Y_i =+1, that means no errors therefore actual and predicted class are same and therefore no miss-classification.

If actual value of Y_i = -1 that means error has occurred therefore actual and predicted class is not same that means miss-classification has occurred.

Case 2: If SVM classifies the input instance x_i to negative class then according to the algorithm

$$T_i=w.x_i-m$$

Let m=1 then $T_i=w.x_i-1$,

T_i-t= $w.x_i-1-w.x=-1$ ....$w.x_i$ and w.x is almost same

If actual value of Y_i =-1, that means no errors therefore actual and predicted class are same and therefore no miss-classification.

If actual value of Y_i = +1 that means error has occurred therefore actual and predicted class is not same that means miss-classification has occurred.

6. CONCLUSION:

This paper presents our approach towards mining the text from google playstore. The reviews are collected from the playstore of VOOT app and then those reviews are analyzed to find the usefull information from it. This was generally done by the sentimental analysis approach. By sentiment analysis we concluded the reviews into two categories as the good review and the bad review. After sentimental analysis we have used SVM(Support Vector Machine) to finally calssify the n number of reviews into the appropriate class they should belong. This would ultimately help the app owner for get their ratings in the industry.

7. REFERENCES

[1] Xin Chen, Mihaela Vorvoreanu and Krishna Madhavan, " Mining social media data for understanding students learning experience" , IEEE transaction on Learning Technologies, vol.7, no.3, Pp,16-22 July - September 2014

[2] Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, Tom Mitchell, “Text Classification From Labeled And Unlabeled Documents Using EM”, Machine Learning, 39, Kluwer Academic Publishers. Printed In The Netherlands, Pp. 103–134, 2000.

[3] Bo Pang And Lillian Lee,”A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based On Minimum Cuts”, Morgan & Clay Pool Publishers, Pp. 54-58, 2008.

[4] J. Han, M. Kamber, Data mining, Concepts and techniques, Academic Press, 2003.

[5] Tina R. Patil, Mrs. S. S. Sherekar,”Performance Analysis Of Naive Bayes And J48 Classification Algorithm For Data Classification”, J.sci.Education, Vol.86, No.1, Pp.7-15, 2000