Feature Extraction And Analysis of Online Reviews For The Recommendation Of Books Using Opinion Mining Techniques

S Sumathi[1], D Rajesh Kumar[2], K Rubiya[3], N Vivek[4]
[1]Assistant Professor, Dept of CSE
[2, 3, 4]Dept of CSE
[1, 2, 3, 4] Sri Ramakrishna Institute of Technology, Coimbatore, Tamilnadu, India

Abstract- *Customers prefer to get the opinion of other customers by observing their opinion through online product reviews, blogs, and social networking sites, etc. Consumer reviews play an important role in determining buying behavior for online shopping because customers prefer to get the opinion of other customers by observing their opinion through online product reviews, blogs, and social networking sites, etc. The customer's opinions have a great influence on the items being sold online, as well as on digital devices, household appliances, movies, and books. Hence, extracting the precise product characteristics necessitates extensive analysis of feedback. we're using human intelligence to extract book-related data from these online reviews We've suggested a system to categorizes the book features from customer feedback. It can help in deciding on which books to recommend to readers. The final aim of the work is to provide users with their wanted books. Thus, we have tested our categorization approach through actual users and professors.*

Keywords- Consumers, Online book review,

I. INTRODUCTION

The advancement of web-based technology has resulted in the storage of a large amount of data on the internet. In terms of e-commerce, all big items are now accessible on the internet via a variety of reputable websites such as Amazon, Flipkart, and Google, among others. These websites provide a thorough overview of the goods, including price, availability, features, customer reviews, comments, and blogs, among other things. These digital words of mouth may be used by a new customer to form an opinion about a particular product. A novel, for example, receives reviews from a variety of customers who have already purchased and read it. These reviews can be used to shape a strong opinion on how good the book is, what genre of people like it or don't like it, whether it's better than its predecessor, and so on. For example, a given book XYZ could have 500 reviews, with review stars ranging from one to five and textual content ranging from "Very good" to "Not Worth Reading." One choice is for a reader to go through all of the available feedback and determine whether or not she should read the book based on the research performed through those reviews based on consumer opinion. This process of forming an opinion is time-consuming and exhausting, particularly when the number of reviews is large and diverse. Alternatively, to form opinions and decide the sentiment direction of online review text given on the website, automated machine learning computational techniques may be used for better resultants. Frequently, new methods are applied to the current ones. People can now actively use information technology to search for and comprehend the views of others, thanks to the growing prevalence and availability of opinion-rich tools such as blogs, feedbacks, and online review sites on internet. Sentiment analysis and Opinion mining have exploded in resulting in the popularity of development of modern analytical techniques that deal with both the textual and sentimental contexts at the same time on analysis. Putting together vast amounts of data to extract what people think has become a fascinating and difficult part of information retrieval. Furthermore, recommending goods based on the study brings a new dimension to sentimental clustering and opinion mining. Recommendation systems are used to suggest a product to a specific consumer based on his prior preferences and dislikes. Such systems use the user's interests and previous search history to recommend products for purchase or to scrutinise a product. They're becoming increasingly popular for providing recommendations that effectively snip vast information spaces, directing users to the things that best suit their needs, interests, and location. Because of their individualised or personalised design, such systems are desirable candidate for a wide range of applications. To test the functionality, a major challenge is to minimise the distance between automatically computed text and the user's semantic control [1]. To effectively mine the data and improve the analysis performance, we used word frequency in conjunction with the star-based popularity of a book in this article. Our work is primarily concerned with effectively using data from a book review dataset. Various websites, such as Goodreads, Amazon, Flipkart, Google Books, and others, offer product ratings as well as customer feedback for a specific book. A consumer gives a book a star rating to give it a review. The

success of a book is determined by its star rating. Along with the stars, a consumer also leaves some feedback about how he or she felt about the novel. These comments can provide a wide range of opinions such as Excellent, Loved it, Worth Reading Once, Waste of Time, and so on, and thus include a valuable amount of knowledge about a specific book. We can see comments on any website, but they can be used to represent the polarity of a particular book to make their use more fruitful. Blend of comments, star-based reviews and user profile can be used to infer various non-trivial facts about any given book. This paper proposes a mining technique that blends the conventional method of star-based rating on content-based text mining with a sentimental classification-based opinion-based approach for better results. Using statistical methods, method is used to provide a better vision of a book review, which could be used by recommendation systems in the future. Our research aims to conduct predictive analysis of online book reviews gathered from the internet. The vast amount of information available on the internet includes reviews, comments, ratings, and blogs, among other things, all of which can help us determine the quality of a product that are evolving around the net. However, for the sake of experimentation, our data collection only contains a small number of book reviews. We aim to use content written in conjunction with star-based ratings to predict desired outcomes more accurately and efficiently, allowing us to present a large volume of useful and serviceable data.

II. RELATED WORKS

The method of extracting interesting information or non-trivial patterns from unstructured text documents is referred to as text mining or knowledge discovery from textual databases. It can be thought of as a development of data mining or information discovery from (structured) databases. [8] The weight of individual words is measured using the TF-IDF (Term Frequency -Inverse Document Frequency) formula in the majority of text mining techniques. This weight is a statistical metric for determining the significance of a word in a list or corpus of documents. [5,6] We used TF-IDF to reliably extract commonly occurring terms in a document from raw data. Classification methods may be used to further separate this data into categories and provide a better image for retrieving information from a corpus. Han and Camber [7] proposed an impressive collection of classification techniques for clustering data using various mining techniques. These strategies differ depending on the individual and the data's intended use. One of the techniques used to learn human profiles from descriptions of examples is content-based text mining. Text mining provides subjective information about a person, from which different conclusions about a specific profile can be drawn. A machine learning algorithm for text categorization [9] was previously discussed for content-based recommendation, in which the content of data is considered while additional parameters are taken into account under data mining. According to studies, reviews with strong opinions are more positive than reviews with mixed or neutral opinions. [4] Furthermore, with the increase in popularity and availability of numerous opinion outlets such as blogs, comments, and reviews, new avenues for extracting meaningful knowledge from data clouds which are emerged. With the and availability of opinion-based tools, such as personal profile-based feedback, new challenges have arisen in extracting valuable information from such data. Opinions and emotions are taking on a computational dimension as a result of this eruption. Pang and Lee used machine learning techniques to categorise movie reviews based on their emotional content. For classification, they used Naive Bayes, Maximum Entropy, and SVM classifiers, and found that machine learnin techniques outperformed human-produced baselines [10]. Data mining can be done with classification methods to detect interesting pattern with the polarity of data to determine whether a product is on the positive or negative side. This classification helps to create the basic implication of a specific product and indicates whether it is acceptable or inappropriate for the interests of a particular person. [No. 11]

III. EXISTING SYSTEM

The majority of current recommender systems employ collaborative filtering techniques, content-based techniques, or hybrid filtering techniques that combine the two. Recommendations from collaborative filtering methods are based on the expectations of other users. Content-driven approaches, on the other hand, allow recommendations based on facts about the item itself. Approaches to collaborative filtering generate recommendations by comparing a consumer's previous selections with those of other consumers who have made similar choices. Many of the shortcomings of content-based systems are addressed by collaborative filtering systems. As input, these systems use a compilation of historical rating data from m users on n items, which is gathered by asking users to enter product ratings as numerical values.

IV. PROPOSED SYSTEM

In this paper, a method for extracting features from online reviews are designed and analysing those features to provide a forum for users to buy books online. We took the customer feedback and divided the book's features into seven (7) diverse categories. These features are focused on customer feedback and are available at a variety of online book stores.

Users will benefit from the categorised functionality, which will assist them in finding the right books for them.

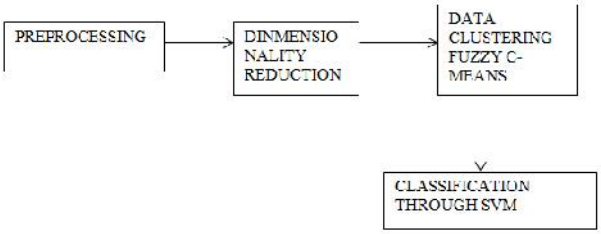

Fig 1. Block Diagram

MODULES

DATA COLLECTION

Major adjectives, adverbs and other powerful terms are filtered out of the training data collection [12]. This data set is then used to extract attributes and the attributes extracted lead to results. The process is described in detail below.

Data pre-processing and filtering

In the above stage, the data collected include polarity information based on English words and scale data based on book reviews' star ratings. This corpus is then transformed to a vector form to conduct the following further operations.

i. **Tokenizing and stemming:** The tokenization process involves splitting a particular document into a series of words. The phrases are broken and transformed into a series of words. Next, the above collection will be generated where all overlapping or duplicate words will be omitted. Adjectives, nouns, orthographing mistakes etc. are considered as one object. For instance, wonderful, wonderful etc. are treated as one word "wonder."
Next, the string data is translated to algorithmic dispensing vectors.

ii. **Conversion Word to Vector:** Until now, we have conducted stemming words to filter out contending words. The significant terms that contribute to feeling analysis are separately filtered out and the vector transformation of the collection is carried out. Further measures are taken to count the word frequency and classify above vectors.

PRE – PROCESSING

Data preprocessing is a technique for converting raw data into a widely understandable format. Data manipulation processes the raw data. Data preprocessing is used in customer relationship management and rule-based applications (like neural networks). Data goes through a number of processes during preprocessing. Data is cleansed by filling in missing values or, removing corrupted or inconsistent data, smoothing the noisy data, or fixing the inconsistency in data. It is most crucial for ML algorithms to smoothly deal with noisy results. Data can be "filtered" by breaking it up into equal parts, or by using a "linear regression" or a "cluster analysis" (clustering). There are data anomalies due to human error (the information was stored in a wrong field). Never duplicate database values to avoid giving the value an advantage (bias). Data Integration: putting together data from various sources with different formats The data are normalized and updated. Data normalization ensure all data is stored in one location and all relationships between data are logical. Queries will become slower as the amount of data become bigger. Data mining phase aims to get the information out of the systems. Data reduction is different processes. For example, if a value with two observations is essential, then anything less than two is discarded. Encoding mechanisms can help to reduce the file size of the data. After decompression, the original data remains undamaged. If any information is lost, it's called lossy compression Aggregation may be used, for example, to compute a single value based on multiple transactions recorded over a shorter time period. Data may also be discretized for evaluating statistical measurements. This analysis includes the reduction of values of continuous attribute from interval. Often, a dataset is too large or too complicated to be work with. Sampling techniques may be used to pick and function with only a subset of the data, given that they have the same properties.

DIMENSIONALIT REDUCTION

The higher the number of features, the harder the training set can be visualised and worked on. Most of these features are sometimes connected and often redundant. This is where algorithms for dimensionality reduction come into play. Dimensionality reduction is a procedure of acquiring a collection of major variables to reduce the number of random variables under consideration. It can be divided into the collection and extraction of functions.

There are two-dimensional reduction components:

Selection of the feature: here we try to find a subset from the original set of variables or features to get a smaller subset that can be used to model the problem. It normally consists of three ways:

Embedded Filter Wrapper

Feature extraction: This reduces the data to a smaller space in a large area, i.e. a space with less dimensional no.

The following measures are included:

- Construct the data covariance matrix.
- Calculate the matrix's own vectors.
- Eigenvectors that represent the largest values of their own are used to recreate a large fraction of the variance of original results.

Advantages of reduced dimensionality

- It helps compact data and also reduces storage space.
- It decreases the time of calculation.
- It also helps to eliminate unnecessary functionality, if applicable.

CLUSTERING C- MEANS

Ensemble machines gaining knowledge of algorithms usually make for better overall predictive efficiency than a single model. The market for machine learning, where the dominant solution was turned into a model for hand radiographs, is taken into account. The C – means for authentication by hand .In order to correct for the effects of background illumination, skin colors and noise, the parameter estimation technology is used to calculate grey size along an axis. C- Means is exactly the same as k-approach, a common simple to combine

C-approach, which is a common simple clustering approach, is exactly identical to C-approach. The only difference is that it should provide some sort of fluidity or overlap between clusters rather than awarding a point totally to the most successful of one cluster. The important aspects are described below by C-manner

CLASSIFICATION

We have used a supervised training technique called SVM Text Classification Algorithm to classify the text in the analysis database. A SVM classification is a Vector Machines supports is like a sharp knife - it works on smaller datasets, but in complex ones, machine learning models can be much strengthened and strong. It uses the highest probability approach to determine the polarity of any given word[14]. This classification technique is used to train the data polarity classification based on string input vectors. Two types of datasets apply the algorithm. The first dataset only contains the summary remarks, the user identity and the book id, while the first dataset contains the star ratings. The classifier will then be trained on the training dataset after the appropriate features have been selected. The training process is an iterative process, with a ratio of 70-30 for training-testing results. We constantly implemented stop word exclusion, word stemming and input features in an attempt to achieve a better performance and further improvise the results. The efficiency of the qualified classifier is then evaluated on the test dataset. The results achieved after classification mean that the use of star-based polarity rating provides a more reliable result than the use of individual polarity measurements.

V. EXPERIMENTAL RESULT

Our Experimental Result describe more efficient percentage compare to existing methods.

VI. CONCLUSION

This article gives an overview of the various algorithms in the sentiment analysis book recommendation scheme. We use human intelligence and categorised book features from online reviews that can help users find the books to choose from. It explains how various classification algorithms contribute to better results in the book recommendation method. Sentiment Analysis is used for the prediction of positive negative outcomes for the book reviews. Similarity Calculation is used to determine the similarity of the objects with users. Compared to existing ones, the SVM delivers productive performance.

REFERENCES

[1] Aciar, S., et al.,2006. Recommender system based on consumer product reviews. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. IEEE Computer Society.

[2] Fang, Y., et al.,2012. Mining contrastive opinions on political texts using cross-perspective topic model. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. ACM.

[3] Hu, M., Liu, B., 2004. Mining and summarizing customer reviews. In: KDD'04.
[4] Huang, W., et al., 2008. Analysis of the user behavior and opinionclassification based on the BBS. Appl. Math. Comput. 205 (2),668—676.
[5] Shani, G., Gunawardana, A., 2011. Evaluating Recommenda-tion Systems. Recommender Systems Handbook. Springer US,pp. 257—297.
[6] Sohail, S.S., Siddiqui, J., Ali, R.,2013. Book recommendation sys-tem using opinion mining technique. In: 2013 InternationalConference on Advances in Computing, Communications andInformatics (ICACCI). IEEE.
[7] Sohail, S.S., Siddiqui, J., Ali, R.,2014a. Ordered ranked weightedaggregation based book recommendation technique: a link min-ing approach. In: 2014 14th International Conference on HybridIntelligent Systems (HIS). IEEE.
[8] Sohail, S.S., Siddiqui, J., Ali, R., 2014b. User feedback scoring andevaluation of a product recommendation system. In: Proceedingof IC3, pp. 525—530.
[9] Sohail, S.S., Siddiqui, J., Ali, R., 2014c. User feedback based evalua-tion of a product recommendation system using rank aggregationmethod. Advances in Intelligent Informatics, Advances in Intel-ligent Systems and Computing, vol. 320., pp. 349—358.
[10] Sohail, S.S., Siddiqui, J., Ali, R., 2015. OWA based book recommen-dation technique. Procedia Comput. Sci. 62, 126—133.
[11] Burke, R.: Hybrid Recommender Systems: Survey and Experiments 12(4), 331–370 (2002)
[12] Adomavicius, G., Tuzhilin, A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering 17(6), 734–749 (2005)
[13] Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender Systems. Cambridge University Press (2010) 71.6 636 Harvinder, D. Soni, and S. Madan
[14] Caoa, Q., Duanb, W., Gana, Q.: Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. Decision Support Systems 50(2), 511–521 (2011)
[15] Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval in Information Processing and Management 24(5), 513–523 (1988)
[16] Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer Reading. Addison-Wesley (1989)
[17] Han, J., Kamber, M.: Data Mining, Concepts and Techniques 3rd edn. (2011)