ISSN NO. 2320-5407

Journal homepage:http://www.journalijar.com
Journal DOI:10.21474/IJAR01

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

Seed Based Method for Identifying Efficient Online Reviews using Micro-reviews.

Jananee.M and Devi Selvam.
Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Tamilnadu, India.

Manuscript Info

Manuscript History:

Received: 15 April 2016
Final Accepted: 26 May 2016
Published Online: June 2016

Key words:
Supervised learning, class imbalance problem, outlier, random sampling technique

***Corresponding Author**

HanyM. Youssef.

Abstract

Review selection based on supervised learning model is utilized to classify the review based on diversification mechanism. Many contributions can be made on the proposed work in order to improve efficiency of the system; initially unsupervised classification technique can be proposed to cluster the content which is outlier. The class imbalance problem is defined in terms of which the ratio of the majority and minority class cardinalities which is considered as outlier. The main idea is to severely under sample the majority class thus creating a large number of distinct training sets. The outlier of data can be expressed as the data which is inefficient to cluster in either of the classes generated; hence these problems can be resolved using the random sampling technique which improves the performance of the system in terms of classification rate and reduction in the outlier data.

Introduction:-

Reviews regarding a product have an important role in the modern world. It has a large impact on the web users to make the opinions and eventually purchasing the product based on the opinions made. We can find an extremely diversified set of reviewed items, which includes everything from commercial products to holiday packages and even restaurants. A review-hosting website becomes more [popular only if the number of reviews for the available product increases. This means that when people get interested in a product, they post more relevant reviews about it. This method of sharing reviews may improve the online information sharing but it also includes few disadvantages. The large volume of reviews on every single item may lead to the problem of redundancy. Redundancy means the problem of repetitious reviews, reviews expressing the same opinions and reviews exhibiting only little knowledge. There are also reviews which provide false information and may be misleading because it does not have the accurate representation of the attributes of a product.

Some causes of inefficient reviews are

1) Inefficient knowledge:-

The reviewer posts the review without having much knowledge on the product. These reviews are based on partial or irrelevant information.

2) Bogus:-

The reviewer purposely submits wrong information about a particular product in order to harm the product’s reputation.

The main focus of our work is that the customer/users need not manually read all the large volume of redundant and non-informative reviews for identifying the required useful set of reviews. The currently employed search engines do not consider the opinionated text. The bogus and the redundant reviews must be filtered before displaying it to the users. Epinions.com and Amazon.com are some of the examples of review website.

Today, there is a vast increase in the usage of social networks and micro-blogging websites, which has led to the new type of online review content known as the “micro-reviews”. Micro-reviews refer to those reviews whose length does not exceed the length of 200 characters. We have considered an alternative source for selecting the desired reviews which is known as the micro-reviews.

Some of the features of micro-reviews are:-
1) It is compact and comprehensive.
2) Lesser time required to analyze the shorter reviews, since the maximum length does not exceed 200 characters.

Few key considerations for user reviews:-
1) Indexation
To ensure that search engines reach the user generated content, it must appear in text form in the HTML.

2) Non-duplication:-
Though snippets of reviews could appear on multiple pages, such as on category pages, or promoted via the homepage, it is important that the full review has a single page and a single URL.

3) Breadth of content:-
As many pages on the site as possible should feature user reviews to maximise the opportunity to rank well for the review content.

Background:-

The growth of online sites and the review content is tremendous in the present time. The fact is that the reviews are highly diverse and often unnecessarily verbose. Selecting the appropriate reviews is difficult for the users, since there are huge numbers of reviews available on the online sites. Micro-reviews are emerging as a new type of online review content in the social media. Micro-reviews are posted by users of check-in services such as Foursquare. They are concise (up to 200 characters long) and highly focused, in contrast to the comprehensive and verbose reviews. In this paper, a novel mining problem is proposed, which combines the two disparate sources of review content. Specifically, we use coverage of micro-reviews as an objective for selecting a set of reviews that cover efficiently the salient aspects of an entity. This approach consists of a two-step process: matching review sentences to micro-reviews, and selecting a small6- set of reviews that cover as many micro- reviews as possible, with few sentences. The objective is formulated as a combinatorial optimization problem, and shows how to derive an optimal solution using Integer Linear Programming. It also proposes an efficient heuristic algorithm that approximates the optimal solution.[1]

Principle of information retrieval: -

A. Information retrieval: -
It is the process of inferring the information which is relevant to the required information from a large collection of information resources. The information search is related to the full text or metadata indexing. The information overload” an be reduced by using the automated information retrieval system.

Once the user enters the query in the system the process of information retrieval begins. A query is generally known as the formal statement of information needs. A major drawback is that the query in information retrieval does not identify single required object in the collection. A single query can match with several objects with different degrees of relevancy. An object is defined as an entity which is represented by the information available in the database. The database information is again matched with the user queries.

B. Social networking: -
Social networking is a platform which shares similar interests, real life connections, backgrounds and activities among people around the world. A social network service is a collection of each user representation, one’s social link and other additional services. These services allow the users to create a profile which to visible to the public. It also provides the user list with whom one can communicate and share information and view the connections within the system.

Most of the services are web-based and it provides users to share the information via internet. The internet services like e-mail and instant messaging is useful for this type of sharing. There are various other communication tools like mobile connectivity, photo/video sharing. The social network is considered as an individual centered service while the online community is known as the group centered service. Social networking sites allow the users to share information in various formats like pictures, events, activities, ideas, etc.

C. Micro blogging services: -

It is a broadcast medium that can be used in the form of blogging. The difference between the micro blog and traditional blog is that the contents of micro blog are smaller in both the aggregate and the actual file size. The micro blogs allow the users to share very small amount of content like messages, pictures and video links. This is one of the reasons for its popularity.

The messages are also known as the "micro posts". In order to promote the website services and products, the commercial micro blog is used. It is also used to promote the collaboration within the organisation.

D. Micro review: -

It is known as the alternative source of review for the users who are interested about the information based on a product. It generally focuses on a specific property of a product. The maximum length of the micro reviews is about 200 characters.

Some micro blogging services offer features such as privacy settings, which allow users to control who can read their micro blogs, or alternative ways of publishing entries besides the web-based interface. These may include text messaging, instant messaging, E-mail, digital audio or digital video.

Methods: -

A. Bag of words construction: -

The bag of words is the important sentences and tips related to the review. Some mechanism like stop word removal and sentence splitting mechanism are used to extract the bag of words. The extracted data is then used for the selection of efficient review process.

B. Concept and opinion generation: -

There may be same concepts for sentence and tip but it may use different words based on the situation. An approximation bound is set for the variants of the efficiency function in the micro-reviews and reviews. There are many important constraints to determine the positive and negative opinions of the sentence. This is used to extract the sentence related to the set of opinions.

C. Selecting of Subset of Reviews: -

Few reviews which have high coverage may be too lengthy. Reviews containing more sentences take so much time to analyze and may be irrelevant. The efficiency concept is used to avoid the reviews which are too lengthy. If we consider sentence s and tip t are matched, then we say that s covers t.

D. Generating set of the reviews as heuristics: -

The greedy algorithm is used to select the reviews because of the sub-modularity property. It is used to identify the solution with the approximation ratio and which maximizes the coverage. The reviews with the high gain to cost ratio is selected.

E. Applying Greedy algorithm: -

The local optimal choice at each stage is done using the greedy algorithm. It also helps in finding the global optimum solution. In general, this strategy does not produce an optimal solution but using a greedy heuristic we can produce an optimal solution in reasonable time.

F. Modeling Feature selection and class formation using unsupervised learning model: -

Bagging is resampling ensemble method for feature selection based on the decision tree mechanism to the micro reviews. It is a relatively simple idea: given a training set, bagging generates many bootstrap samples or training subsets, generating subsets of samples where each sample which is considered as feature is selected with replacement and equal probability. A bagged ensemble predicts a new sample by having each of its base models classify that example. The final prediction of the class is normally obtained by majority voting.

G. Review Prediction: -

A conventional training of a positive class using a dataset containing representative proportions of Samples from the positive and negative classes will tend to find a solution that will be biased towards the larger class. In other words, the probability of misclassifying samples from the negative class will be lower than the probability of error for the positive class. However, as we wish to retrieve review features of micro reviews to group the major reviews based on the data prediction. We need to reverse the impact of the priors for classifying the positive and negative opinions.

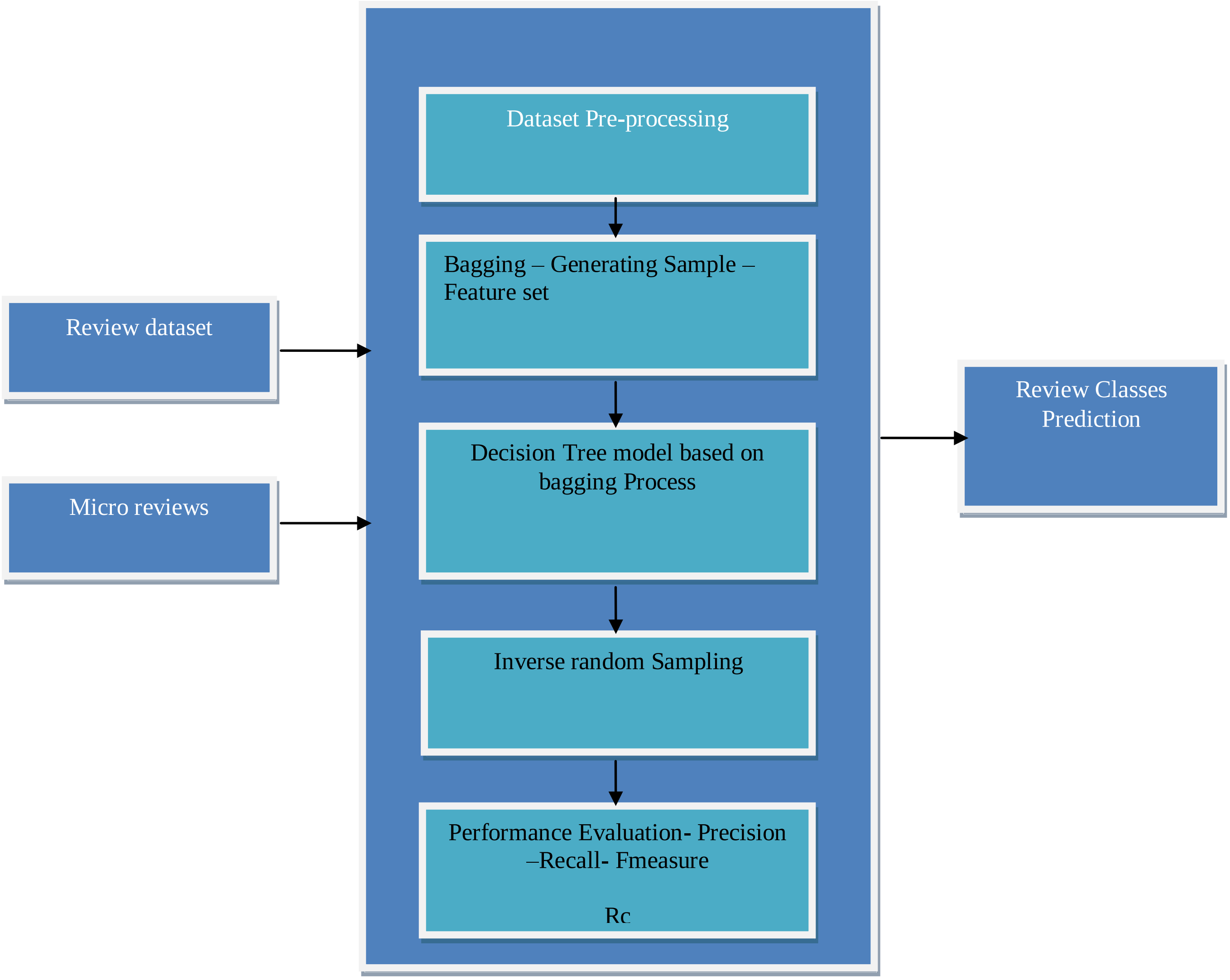

Dataset Pre-processing
Bagging – Generating Sample – Feature set
Decision Tree model based on bagging Process
Inverse random Sampling
Performance Evaluation- Precision –Recall- Fmeasure
Rc
Review dataset
Micro reviews
Review Classes Prediction

Figure 1: Architecture Diagram

Results and discussions: -

A. Data collection and data preprocessing: -

The experiment requires data coming from two different sources (reviews and micro-reviews), concerning the same set of entities. We pick the domain of manufacturing, because it is a domain where there are active platforms for reviews as well as for micro-reviews and it is available in message and detailed appreciation mail or certificate.

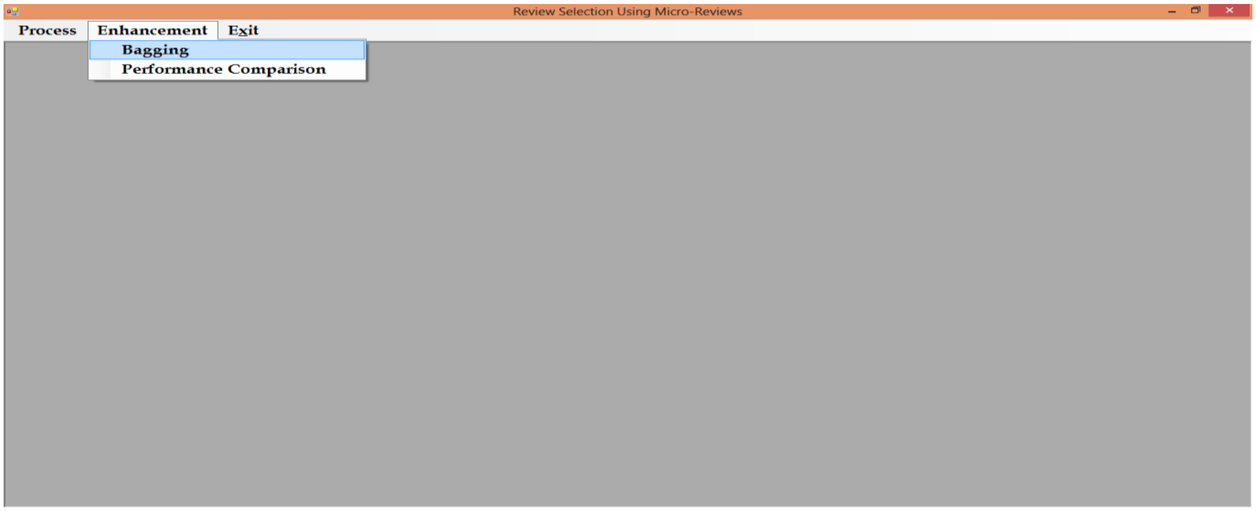


Figure 2: Data preprocessing

B. Review Granularity: -

Review granularity aims in the collection of tips. Generally, a tip is based on a single point and it is short and concise. A review is longer and based on multiple information of an entity. If the tip appears on the text of the review, then we say that the review covers the tip. For making it simple the entire review content is split into sentences which are the semantic units with the similar granularity as that of the tip.

We can match a sentence s with a tip t only if they both convey the same meaning. For making the matching decision, three criteria are to be considered. First, is to consider the sentence and the tip as the bag of words. It is considered to convey a similar meaning if they share a substantial subset of the content. Now, we consider that they have high syntactic similarity. The reviews and the tips express the opinion of the respective authors. Other than having the similar keywords and concepts, the tip and the review must also have the same sentiment which may be either positive or negative. Now, we say that they have high sentiment similarity.

C. Bagging: -

The bagging mechanism is implemented to predict the most relevant reviews to the micro review in effective manner. The effective result gets obtained by training large number of subsets to extract the suitable reviews. The implementation of this methodology helps to achieve better results. Each and every review gets grouped under some related classes with the help of outlier detection. The new class formation gets supported when any review doesn't match with any existing classes helps in improving stability of the system.

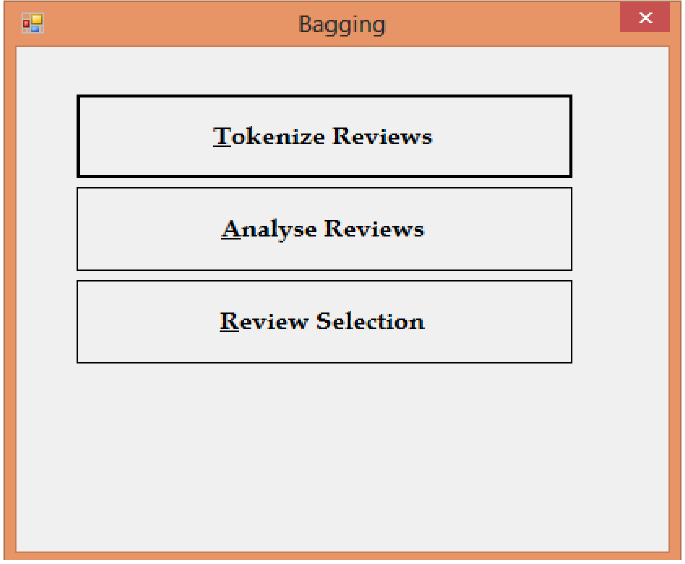
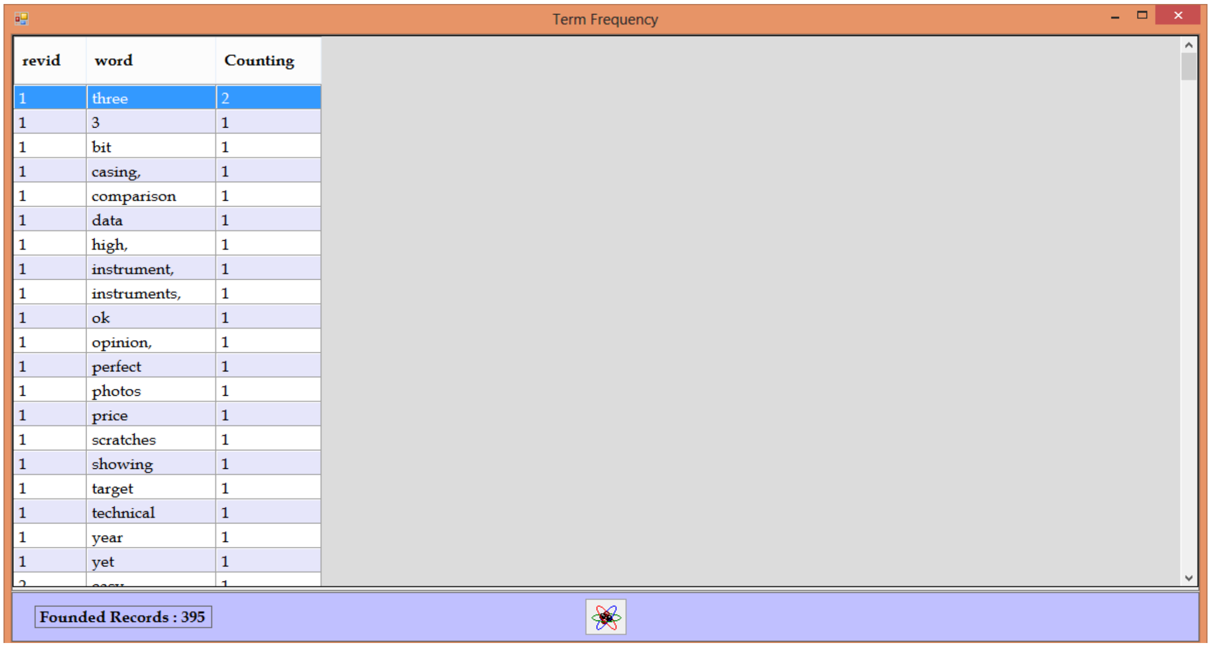


Figure 3: Bagging

D. Identifying term frequency: -

In order to identify the best matching reviews between the micro-reviews and the user reviews the term frequency is to be identified. The term frequency is mothing but the number of times a feature is repeated in the sentence. It also delivers the number of times the feature is available in the adjacent reviews. The feature which has the highest term frequency is the feature which is to be selected. This feature is compared with the other reviews present in the user review.



Term Frequency

revid	word	Counting
1	three	2
1	3	1
1	bit	1
1	casing,	1
1	comparison	1
1	data	1
1	high,	1
1	instrument,	1
1	instruments,	1
1	ok	1
1	opinion,	1
1	perfect	1
1	photos	1
1	price	1
1	scratches	1
1	showing	1
1	target	1
1	technical	1
1	year	1
1	yet	1
2	easy	1

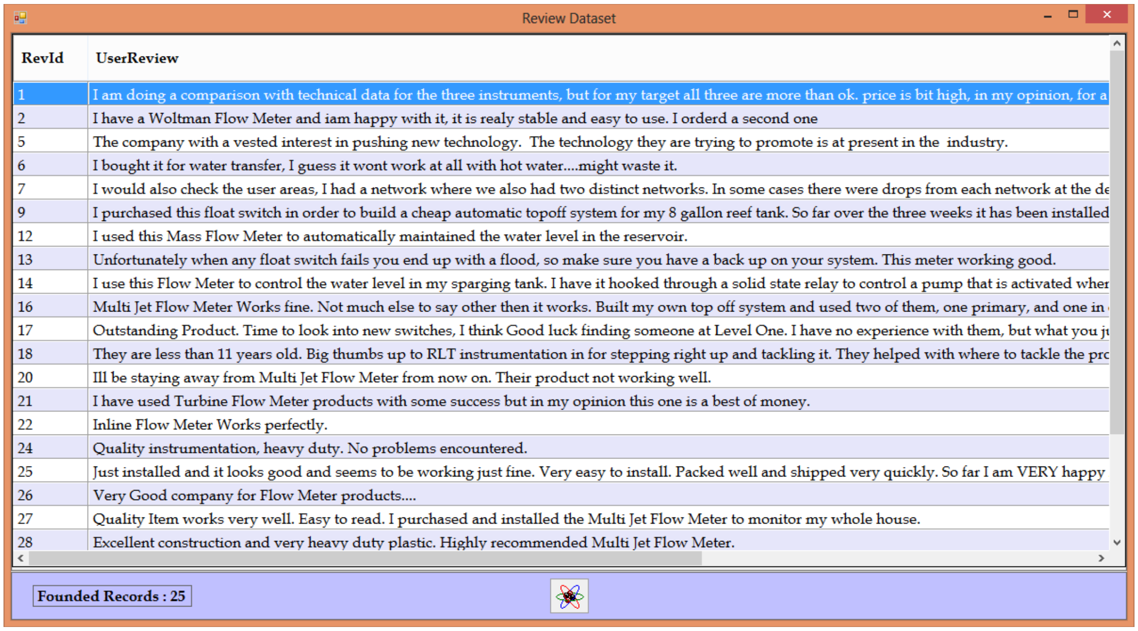
Founded Records : 395

Figure 4: Term Frequency

E. Review Selection: -

The sentiment analysis cannot be easily done. The data can be collected from specific websites using either the open application programming interface or the correspondent crawler. Collecting the data based on the specific topic is more complex. We now design a customized data wrapper for one platform to extract the metadata. The metadata includes the user ID, timestamp, post message and the properties of users. The method helps in the collection of highly accurate data on a specific topic.

The most important sentences are given the higher priority, such as they are used as the title, the first sentence, the last sentence or the emphasis. While calculating the overall polarity, we must consider the location of the sentiment sentence. The weight in the overall polarity calculation represents the importance of a sentence to the document. Hence the weight of the important sentence must be greater than the other sentences in the document.



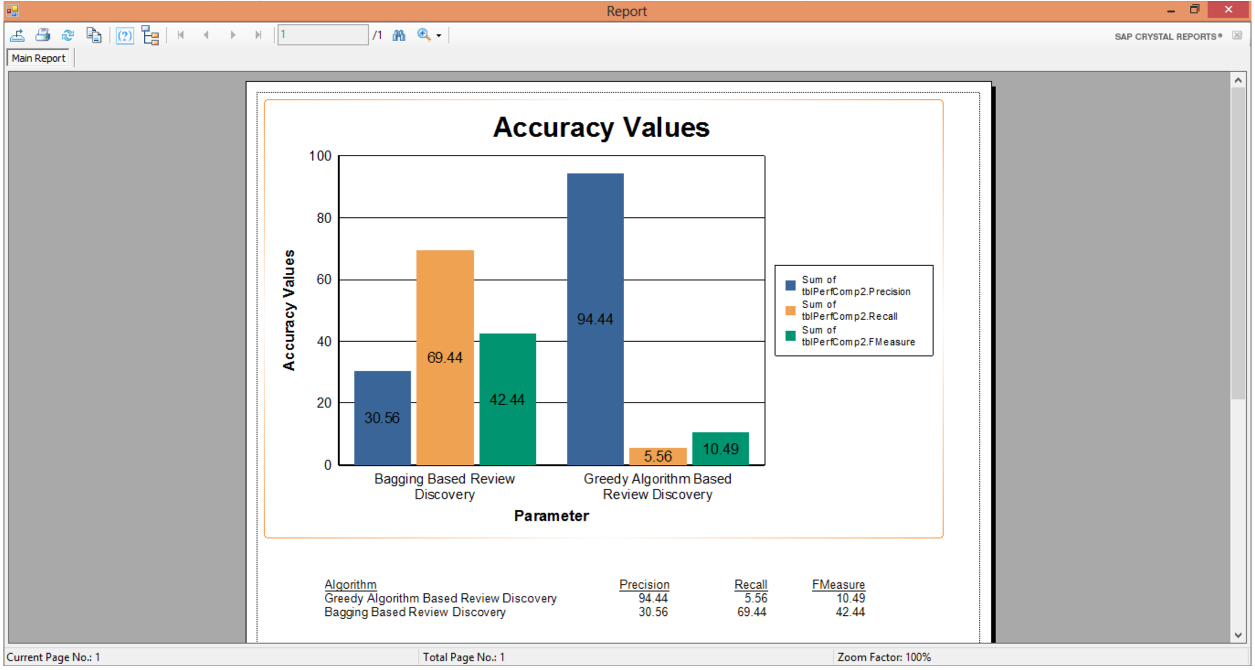
RevId	UserReview
1	I am doing a comparison with technical data for the three instruments, but for my target all three are more than ok. price is bit high, in my opinion, for a
2	I have a Woltman Flow Meter and iam happy with it, it is realy stable and easy to use. I orderd a second one
5	The company with a vested interest in pushing new technology. The technology they are trying to promote is at present in the industry.
6	I bought it for water transfer, I guess it wont work at all with hot water....might waste it.
7	I would also check the user areas, I had a network where we also had two distinct networks. In some cases there were drops from each network at the de
9	I purchased this float switch in order to build a cheap automatic topoff system for my 8 gallon reef tank. So far over the three weeks it has been installed
12	I used this Mass Flow Meter to automatically maintained the water level in the reservoir.
13	Unfortunately when any float switch fails you end up with a flood, so make sure you have a back up on your system. This meter working good.
14	I use this Flow Meter to control the water level in my sparging tank. I have it hooked through a solid state relay to control a pump that is activated wher
16	Multi Jet Flow Meter Works fine. Not much else to say other then it works. Built my own top off system and used two of them, one primary, and one in
17	Outstanding Product. Time to look into new switches, I think Good luck finding someone at Level One. I have no experience with them, but what you ju
18	They are less than 11 years old. Big thumbs up to RLT instrumentation in for stepping right up and tackling it. They helped with where to tackle the pro
20	Ill be staying away from Multi Jet Flow Meter from now on. Their product not working well.
21	I have used Turbine Flow Meter products with some success but in my opinion this one is a best of money.
22	Inline Flow Meter Works perfectly.
24	Quality instrumentation, heavy duty. No problems encountered.
25	Just installed and it looks good and seems to be working just fine. Very easy to install. Packed well and shipped very quickly. So far I am VERY happy
26	Very Good company for Flow Meter products....
27	Quality Item works very well. Easy to read. I purchased and installed the Multi Jet Flow Meter to monitor my whole house.
28	Excellent construction and very heavy duty plastic. Highly recommended Multi Jet Flow Meter.

Founded Records : 25

Figure 5: Review Dataset

F. Performance Evaluation: -

The data contains the reviews and micro reviews for processing the methodology. The set of entities which are helpful to extract the review are get presented in the micro review. The comparison graph between the greedy based review discovery and the bagging based review discovery are shown in the below figure.



Algorithm	Precision	Recall	FMeasure
Greedy Algorithm Based Review Discovery	94.44	5.56	10.49
Bagging Based Review Discovery	30.56	69.44	42.44

Figure 6: Performance comparison chart

The value gets compared between the precision, recall and Fmeasure value between the two algorithms. The enhanced methodology shows improve in the performance of bagging approach. This shows that the performance improvement in the proposed methodology.

The bagging of words helps to match in accurate manner about the micro review and the reviews present in the system. The training of words helps to extract more reviews in effective manner. The outlier detection calculates the class positive and negative with the existing class and matches with the present class otherwise forms a new class.

The graph shows that the bagging values provides high value of accuracy in predicting the system which avoids the confusion for the user. The stability and accuracy of the system shows high performance.

The data contains the reviews and micro reviews for processing the methodology. The set of entities which are helpful to extract the review are get presented in the micro review. The comparison graph between the greedy based review discovery and the bagging based review discovery are shown in the below figure.

Conclusion: -

In this work the effective mechanism to improve the stability of the system by using bagging methodology and every review get in to the analysis by implementing the outlier detection. The bagging approach extracts multiple training sub-samples from the original dataset and combine predictions made based on these samples to form an aggregate final prediction of the review matching for the micro review. By applying this process achieves reducing in computation time further (as compared to global smoothing), while still providing significant stability and accuracy improvements. The experimental results show the improved results get achieved in the proposed system.

References: -

1. Thanh-Son Nguyen, Hady W. Lauw, ”Review selection using Micro-Reviews,” in Knowledge and Data Engineering, IEEE Transaction on, 2015, No. 4, vol 27, pp. 1098-1111.
2. K. Ganesan, C. Zhai, and E. Viegas, “Micropinion generation: An unsupervised approach to generating ultra-concise summaries of opinions,” in Proc. 21st Int. Conf. World Wide Web, 2012, pp. 869–878.
3. Gao, J. Tang, X. Hu, and H. Liu, “Exploring temporal effects for location recommendation on location-based social networks,” in Proc. 7th ACM Conf. Recommender Syst., 2013, pp. 93–100.
4. Kouloumpis, T. Wilson, and J. Moore, “Twitter sentiment analysis: The good the bad and the omg,” in Proc. 5th Int. Conf. Weblogs Social Media, 2011, pp. 538–541.
5. Ghose and P. G. Ipeirotis, “Designing novel review ranking systems: Predicting the usefulness and impact of reviews,” in Proc. 9th Int. Conf. Electron. Commerce, 2007, pp. 303–310.
6. M. Hu and B. Liu, “Mining and summarizing customer reviews,” in Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2004, pp. 168–177.
7. P. Tsaparas, A. Ntoulas, and E. Terzi, “Selecting a comprehensive set of reviews,” in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2011, pp. 168–176.
8. T. Lappas, M. Crovella, and E. Terzi, “Selecting a characteristic set of reviews,” in Proc. 18th ACM SIGKDD Int. Conf. Knowl .Discov. Data Mining, 2012, pp. 832–840.