Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/110618

Product based Business Decision Making for Enhanced Information Retrieval based Root Cause Analysis

S. Subitha[1,*] *and S. Sujatha*[2]

[1] M. A. M College of Engineering, Siraganur, Tiruchirappalli, Tamilnadu, India.
[2] Bharathidasan Institute of Technology, Anna University, Tirchirappalli, Tamilnadu, India.

Received: 10 Apr. 2017, Revised: 17 Oct. 2017, Accepted: 21 Oct. 2017
Published online: 1 Nov. 2017

Abstract: Business decision making has become complex due to increased online e-commerce activities (purchases). Though profitable, this has reduced personalized interactions with buyers, hence leaving the organizations in the leeward side of user's feedbacks and requirements. Social networking sites provide a base platform for users to interact with their peers, also providing a platform for organizations to leverage the missed feedbacks and requirements. However, the huge amount of content (less relevant and more irrelevant) present in such domains makes appropriate data retrieval a complex task. This paper presents an architecture that can be used to effectively retrieve information from heterogeneous data sources based upon product based queries presented by user. Major root causes related to the users query are identified from the retrieved information. The root causes are segregated in terms of their polarity, hence providing results of higher significance. The identified root causes can be used for effective decision making. Efficiency of the retrieval levels and sentiment prediction levels are experimentally evaluated and were found to be effective in terms of scalability, retrieval levels and accuracy levels.

Keywords: Decision Support Systems; Information Retrieval; Sentiment Analysis; Root Cause Analysis; Polarity Identification

1 Introduction

Business decision making is one of the major requirements of the current business environment due to the increasing competitions in the market. Negative reviews by customers cannot be possibly eliminated, however, a review that has not been responded to creates huge negative impact on the organization as a whole [1]. The major issue is that not all buyers provide their feedbacks directly to the organizations involved. Hence identifying the impact of a product (both positive and negative) has become a very difficult task. Effective identification of feedbacks is the first step to identifying the flaws in the products involved. Increase in the number of social networking sites and increase in adoptions by customers have made them the most important areas to look for customers opinions. However, the informal nature of the platform has resulted in requirements for a high level analysis system to appropriately identify the context of the data under analysis. Human analysis is always the best in-terms of decision making. However, the decision making process can be made more accurate and much easier with analysis techniques under the hood. Besides, it is practically impossible for a human user to read and analyze such huge amount of information. Hence, decision making systems come into play. These systems analyze data in appropriate ways, process them and parse them to provide effective result sets, making it easier for decision makers. The result sets provided to decision makers consists of a set of major entities spoken about by the users. These entities are assumed to be major causes for the positive or negative opinions in the reviews [2]. Root-cause analysis is the process of identifying the major reason for an event to occur (mostly an undesirable event or a failure), the removal of which will prevent the event from occurring again. A causal factor is considered to be the one that affects the outcome of an event, however, it is not the root cause for the event. Identifying both the root cause and the causal factors are important, as both play a vital role in determining the desired outcome. In case of positive reviews, the causal factors are identified and improvised and in case of negative reviews, the root cause and the causal factors are identified and eliminated. A good business is the one that

* Corresponding author e-mail: subitha.haran@gmail.com sujathaaut@gmail.com

performs both the processes effectively. Hence root-cause analysis in terms of both positive and negative aspects of a product is mandatory [3]. This paper presents an architecture that can be used to solve the entire workflow from information retrieval to identifying the root cause. The architecture is broadly divided into three sections. The initial section performs context based information retrieval, the second section performs sentiment analysis to segregate reviews and the third section identifies the significant terms in the text to identify the root causes and presents the results to the user. The results presented in this section are made up of components that are the major entities in the text. Analyzing them in terms of their opinion polarities provides an effective mode for decision making. The significant terms are presented to the user in the order of their prominence levels, hence the term/ terms with highest prominence are considered to be root causes, while the other terms are considered to be causal factors. Experiments exhibit high accuracy levels and effective document retrieval levels, suggesting high performance levels. The remainder of this paper is structured as follows; section II presents the related works, section III discusses the proposed information retrieval and root cause analysis architecture, section IV presents the results and discusses them and section V concludes the paper.

2 Related Works

Business process decision making has been a deep-rooted process that has been carried out for many years. However, algorithm aided decision making is fairly new technique and has been on the raise since the increase in social networking. This section discusses some of the recent techniques in the decision making process to identify the root cause analysis. Sentiment analysis and context based information retrieval plays a vital role in identifying the root cause for an aspect. Hence studies regarding sentiment analysis and information retrieval are also discussed. An analysis of various sentiment analysis techniques was presented by Dubey et al. in [4]. This technique stresses the importance of sentiment analysis and analyses the importance of social media in this context. It also discusses several methods that can be used to effectively retrieve information from the web, extract phrases and perform semantic orientation of text. A trust-based microblog analysis on the basis of sentiments was presented by Zhang et al. in [5]. This technique initially evaluates the trust related to various users of the microblogs and then identifies the sentiments related to the reviews presented by them. Past interactions serves as the basis for identifying the users trust levels. Microblog analysis is carried out based on these trust levels. Another technique proposed to evaluate the teaching methods was proposed by Zhao et al. in [6]. This technique operates on the feedback and results providing effective sentiment classifications for teaching techniques. Appropriate information retrieval plays a vital role in the process of root cause analysis. Huge amount of information contained in the Web poses a huge challenge to the analysis techniques. Effective retrieval of information is the key to reducing this complexity level such that further operations can be performed with ease [9]. Retrieving information on the basis of content was the legacy technique used for shortlisting entities. However, the current advancement in techniques have paved way for a much complex filtering technique that identifies entities on the basis of their context, rather than the plain text matching scheme. A novel-context based information retrieval technique that performs ad-hoc retrieval of documents was presented by Zakos et al. in [7]. This technique generates a term significance metric dynamically that can be used as a substitute for the term frequency. This mode of operation is claimed to exhibit increased performance compared to its counterparts. A feedback based information retrieval technique was presented by Golitsyna et al. in [8]. This technique operates on the basis of internal and external feedback and multidimensional quantitative analysis. A context aware mechanism to improve the quality of the information retrieval process in the pervasive computing environment was presented by Loyall et al. in [10]. This paper presents techniques that combine multiple contextual aspects to identify the context. Information, when analyzed in terms of context are usually not analyzed in isolation. They are considered to be a collection of concepts that are aggregated to exhibit a single aspect. This theory serves as the basic aspect for the contribution by Carrillo et al. in [11]. This technique uses random indexing and uses co-occurrence information among words to identify the semantic context related to the word vectors. Significance based information retrieval techniques [**?**] are also on the raise A sustainability and vulnerability based decision making system was presented by Winograd in [12]. This technique operates on the proposed data, identifying them as positive and negative causes. Further operations are carried out by enhancing the positive causes and eliminating the negative causes. A feedback analysis based text classifier that identifies the sentiment related to microblog data was presented by Balusamy et al. in [13]. This technique operates by using conventional classifiers such as SVM and utilizes PCA to perform data reductions.

3 Information Retrieval and Root Cause Analysis

Identifying the major positive or negative root causes for a products success or failure can pave way for effective updates, upgrades or even withdrawals of products. However, the major complexity arising in this design is the unavailability of the appropriate data that can

effectively reveal the root causes. Even though most of the organizations have their own portals to obtain feedbacks, not every satisfied or dissatisfied customer discloses their guileless views at this point of contact. This leads to a communication gap between the producers and consumers of the products. In the current interconnected world, this gap is being bridged by the social networking sites. Though psychologically, a human has several concerns in pointing out the fault of a product to the manufacturer, they do not have such reservations in discussing them with a friend or a colleague. These discussions (actual honest reviews) can be largely found in social networking sites in chats. Such chats tend to reflect the real frame of mind of a user, hence are a valuable resource for the manufacturers. However, millions of such conversations exists online, making it difficult for the organizations to identify the content pertaining to their content of interest. The proposed architecture deals with identifying such text from huge heterogeneous data sources along with the metadata and categorizing them according to the polarity scores to identify the major root causes in both positive and negative reviews.

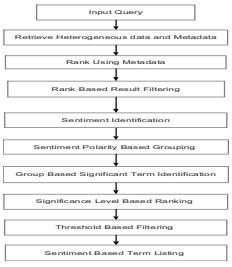

Fig. 1: Information Retrieval and Root Cause Analysis Architecture

The proposed architecture, shown in figure 1 has three major phases; the data retrieval and filtering phase, sentiment identification phase and polarity based root cause identification phase.

3.1 Data Retrieval and Filtering

Data retrieval and filtering forms the initial phase of operations in the proposed architecture. Identifying the appropriate text from the web plays a vital role in identifying the root cause of the issue. A query is constructed with the required content as the retrieval keyword (usually the product name) and is queried on the heterogeneous data sources using their corresponding APIs. Internet, being the storehouse of information contains data in several formats. Heterogeneous data includes databases, HTML/XML pages, multimedia content with tags and plain text documents. Hence queries are built conforming to the types of the data sources, with the constraint imposed by the retrieval keyword. The results obtained are passed to the next phase for ranking. The retrieved results also contain metadata with details about the data provider, credibility of the provider, data source, date of posting, user ranks, likes, comments, unlikes, reposts/ retweets etc. Each data source contains most (if not all) of these information. Rank of a result is calculated using the weighted sum method [18,19]. A default property weight of 1 is initially assigned. The user provides higher weight values for properties that are of importance for their analysis. The rank corresponding to a result is calculated using eq 1.

$$R = \sum_{i=1}^{n} W_i * P_i \tag{1}$$

where R is the rank of the result, n is the total number of available properties, W_i is the weight corresponding to the property and P_i is the actual value of the property. Values of P_i are discrete and categorical calculated according to their interval ranges (in case of timestamp or numerical data) or labelled according to textual data (in case of ordinal or text data). Not all retrieved data contains paired metadata. Some data might contain partial metadata properties, while others do not contain any metadata at all. Rank calculation for data containing partial or complete metadata is performed using eq (1). In case of partial data, properties that do not contain values are ignored. A default rank of 1 (lowest rank) is applied to properties that do not contain any metadata. Threshold based data filtering is performed to obtain the final data. Data falling below the user defined threshold are filtered and the remaining data is retained for sentiment analysis. A defined minimal number of results are required for the next phase, if the number of filtered results fall below this limit, data with lower ranks are incrementally included in the result set until the constraints are satisfied. The shortlisted text is then passed to the sentiment identification phase for polarity identification.

3.2 Sentiment Identification

The sentiment identification phase analyses the individual textual components of the data and identifies their polarity levels. Sentiment of the entire text is determined by aggregating the polarities of each of the textual elements. Sentiments are generated along with their magnitude levels. These results are used to segregate the data into positive and negative classes to identify the root causes. Various stages of the sentiment identification process are shown in figure 2.

3.2.1 Tokenization

Sentiment of the entire text can be determined by aggregating the polarity values of individual elements in

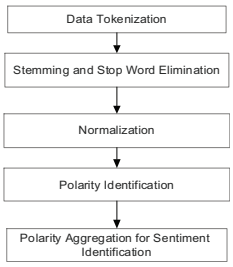

Fig. 2: Sentiment Identification Process Flow

the text. Foremost process of the sentiment identification phase is tokenization. Tokenization is the process of dividing the data stream into meaningful tokens. The meaningful tokens include words, phrases, numbers, symbols or other domain based meaningful elements. The given text is broken into tokens using defined delimiters such as comma, full-stop, space, hyphen, symbols etc. The process of division is usually carried out on the basis of simple heuristics. Several delimiters can be contained in the text, however, domain understanding is required to identify which delimiters are to be considered and which are to be ignored. Heuristics are defined depending on these considerations. The resulting tokens may or may not have the delimiting characters. This depends upon the heuristic being used. This work focuses on the sentiment of the text, hence the proposed heuristic eliminates all the delimiters. As a result, the proposed tokenizer generates a clean set of contiguous and meaningful tokens for every text.

3.2.2 Stop-Word Elimination and Stemming

Stop-Words are connecting entities that do not have any intrinsic meaning associated with them. They act as filler words, providing a glue between works connecting them in a meaningful manner. Due to their neutral behavior, they can be safely eliminated from data without causing any side-effects. This phase eliminates stop-words from text and stems word affixes to obtain root words. The list of tokens corresponding to each document is processed independently. It is checked for stop-word criteria and eliminated if it is a stop-word, else stemming is applied on it. Stop-word elimination is carried in reference to a word repository. This work uses WordNet3.0 (https://wordnet.princeton.edu/) as the stop-word repository. Every word is cross referenced using this repository and if the reference results in a hit, the word is considered as a stop-word, hence is eliminated, otherwise the word is assumed to be a meaningful entity. All tokens passing through the stop-word elimination filter are passed to the stemming module. Stemming is the process of identifying the seed word or the root word of a particular token. Words occurring in a sentence is usually prefixed or suffixed with necessary syllables to make it meaningful. However, polarity analysis requires the basic structure of the word rather than its inflected forms used in the sentence. Stemming eliminates the unnecessary components of the text to provide the seed word. Several algorithms are available for the process of stemming. However, most of the techniques treat the words with the same stem as synonyms. This process is referred to as conflation. The first stemming algorithm was proposed by Lovins in [20]. All the techniques that followed this technique were modifications of the stemmer technique proposed by Lovins. The next remarkable contribution in stemming technique was proposed by Martin Porter in [21]. A second variant of this technique, Porter 2 was proposed by Martin Porter himself and is maintained with appropriate updates [22]. This work uses a variant of the porters stemmer. A modified parallel version of the Porter stemmer is used in this work for the process of stemming. Data parallelism is achieved in the proposed approach, as a single text can contain several tokens that can be processed in-parallel. The tokens that pass through the stop-word filter are processed by the proposed stemming algorithm and the seed words are obtained. Certain words in the text might not contain appropriate stem words, the reason being misspellings or inflections. Such words are returned by the parser as such without any modifications. This work handles only English language, hence words in other languages written with Latin script cannot be recognized by the parser. In such cases, the words are neither eliminated as stop-words, nor given any polarity. The major reason for not eliminating such tokens is that tokens expressed in other languages have high probabilities of containing information of importance. Though specific language parsers cannot understand the vitality of such text, it is much easier for a human to understand it. Hence they still exist in the text with a neutral sentiment associated with them, not affecting the polarity of the text in any manner.

3.2.3 Normalization

Normalization is the process of restoring words to their canonical form such that they get meaningful. Though stemming converts the words to their root form, stemming basically operates on the basis of regular expressions. Hence several words get ripped off, often resulting in an incomplete meaningless word. This mandates the use of normalization. Normalization has two functionalities; it eliminates the unnecessary character recurrences and it applies appropriate inclusions to make complete and meaningful tokens. The major reason for such a process to be incorporated in the workflow is that the root cause analysis is performed on user reviews or on discussions about the product in concern. Users do not usually follow formal modes of writing in discussions with friends, especially on social networking sites. Use of colloquial mode of language is seen in abundance in such context. However, polarity identification is impossible for such

text, making normalizations one of the mandatory process to be carried out prior to polarity analysis. The major advantage of this approach is that it uses both pattern based analysis and a reference repository, making the identification process effective. Normalization is applied on all words existing in their inflected forms or in their ripped off forms. The inflected words are identified by discovering unnecessary additional character repetitions. These repetitions are incrementally reduced by searching for the base word in the repository. The major reason for incremental reduction is that there are several proper words that contain multiple occurrences of a single character. Reduction of repetitions is performed until a meaningful token is identified. This process is carried out by matching patterns with the token under analysis and words in the repository. Initial filtering is performed with partial matches. Word with the best match replaces the token. If the input token does not correspond to any meaningful word, then it is checked for misspellings or swapped characters. Misspellings and character swaps are the mostly occurring mistakes, hence before tagging a word as lexically meaningless, it is mandatory to check for disjoint matches. This is performed by identifying the size of the token (in terms of characters), the number of perfect character matches (number of characters in token that perfectly matches the number of characters in word), number of perfect character group matches (number of word-token subset matches with character count ¿ 1) and the similarity levels in the location of the matched characters. If the token under analysis exhibits high level of correlation with a word in the repository, in terms of all the above mentioned properties, the input token is replaced by the word in the repository. The major advantage of this approach is that this automatically corrects the ripped off tokens by adding the appropriate prefixes and suffixes, making them meaningful and complete. Most of the tokens that exits out of this section are meaningful tokens that could be provided a polarity value. Certain tokens that did not get appropriately processed in any of the above mentioned sections are also contained in the text. Such tokens are considered to be lexically meaningless, and are assumed to correspond to some other language. Such entities are assumed to make sense to a human rather than a parser. Hence they are passed in their true form after redoing all the changes that were applied on them during the normalization phase.

3.2.4 Polarity Identification and Aggregation

Polarity is defined as the level of positive or negative sentiment associated with a token or word. During the initial phases of polarity analysis, a token or word was considered to be associated with a single polarity, either positive or negative or neutral. However, in the current stage, it could be observed that a token is not related to a single polarity alone. It has a level of positivity and negativity associated with it, hence making the single polarity association an inappropriate approach. This work uses the polarity repository, SentiWordNet 3.0 [23]. The SentiWordNet being a human annotated data repository, has very high reliability levels. The normalized tokens are identified from the repository and their corresponding positive and negative polarity levels are retrieved. This process is carried out for all of the text and text based consolidation of data polarity (Polarityd) is calculated using eq 2

$$Polarity_d = \sum_{i=1}^{n} (Polarity_{(pos,i)} - Polarity_{(neg,i)}) \qquad (2)$$

Where n is the number of tokens in the document $Polarity_(pos,i)$ refers to the positive polarity associated with the term i and $Polarity_(neg,i)$ refers to the negative polarity associated with the term i. The result set also contains meaningless tokens that are preserved for analysis by a human reviewer. Such words are provided neutral polarity (0) such that it does not have any effect on the polarity of the data. This process results in the identification of polarity scores for the entire data (reviews). These scores serves as a base in identifying polarity based root causes pertaining to customer reviews.

3.3 Polarity based Root Cause Identification

Identifying the root cause is the major concern of this work, however, analyzing the root cause irrespective of its opinion magnitude (polarity) would be useless. The polarity of the review has a major role to play in the analysis and in the decision making component. Polarity of a review plays a vital role scheming the follow-up actions. Positive root causes will help the decision makers to understand the positive aspects of the product under consideration. Such aspects need to be maintained as such or improved for the betterment of the product. While negative root causes will require immediate attention and solutions. Hence it becomes mandatory to provide root causes in terms of the polarity rather than as a single category. The reviews are first segregated on the basis of their polarity values. Neutral reviews are added to the reviews under positive polarity. The next process is to identify significant terms in the reviews. In general, documents consist of several words, of which certain words are highly significant, while the others will be of moderate to low significance. The major difficulty is that, determining the significance of words based in their frequency of occurrence alone will not be appropriate. For example, it has been identified that The is the mostly used word in English, hence the frequency of the in any document is much higher than any other words. However, the word under study has no significance. Though the proposed work eliminates stop words in the earlier phase, several domain based words also exhibit high level of occurrence and low level of significance. The significant

term identification phase aims at identifying words with high importance eliminating frequent and common domain based words. Significant terms can be identified by a composition of the Term Frequency (TF) and the Inverted Document Frequency (IDF) of terms in a document. Higher value for TF-IDF indicates higher significance of the word in the document. TF-IDF is calculated using eq 3.

$$tfidf(t,d,D) = tf(t,d) * idf(t,D) \quad (3)$$

Term Frequency (TF) and the Inverted Document Frequency (IDF) are calculated using (4) and (5) [24]

$$tf(t,d) = \frac{tf(t,d)}{*} tf(t,d) \quad (4)$$

where, f(t,d) refers to the number of times the word t is contained in the document d and count(w,d) refers to the total number of words contained in the document d.

$$idf(t,D) = log \frac{N}{|d \in D : t \in d|} \quad (5)$$

where, N is the total number of documents in the corpus, and $|d \in D : t \in d|$ is the number of documents that contains word t.

Tokens are ranked on the basis of their significance scores (TF-IDF) in the document. The user can also specify a threshold to eliminate terms of low significance. These final results correspond to the root causes associated with the search query. Segregated results aid users to easily identify the opinions of the user (either positive or negative) corresponding to the token, making it easier for the users to formulate decisions.

4 Results and discussion

Experiments were conducted using the twitter corpus [25] as the base data and queries were applied on the data. This work assumes the presence of a huge amount of data, as the query is applied on heterogeneous data over the internet. Hence the implementation is done on the Spark platform. PySpark is used to implement the parser and HDFS is used as the file system to store the retrieved data. Evaluation of the information retrieval component is performed on the basis of the shortlisting levels and the efficiency of the shortlisting process [26]. Shortlisting levels observed in the selection process and the efficiency of the shortlisting process are presented in figures 3 and 4. The system was presented with ten queries each of varying complexities ranging from very low (1) to very high (10). Query keywords containing single unambiguous terms constitute low complexities, while the number of ambiguous terms leads to an increase in the complexity. A comparison between the actual results (corresponding to the query) contained in the repository vs. the number of results effectively filtered by the parser is presented in figure 3. It could be observed that irrespective of the level of complexity, the proposed technique exhibits varied shortlisting levels. These variations are attributed to the query and the returned result set and its metadata. The shortlisting efficiency is identified by finding the ration between the actual results (corresponding to the query) contained in the repository and the retrieval levels. It could be observed that in certain queries, the efficiency levels are low, while others exhibit high levels of efficiency. The low efficiency levels are due to the presence of query terms as a substring in a higher level term. The proposed technique operates on a unigram model, hence it analyses single tokens independently. Hence it is not possible to retrieve bigram or multi-gram based tokens, leading to low efficiency in certain queries.

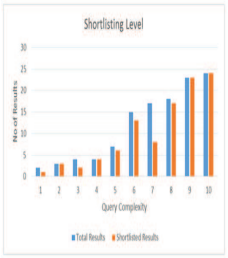

Fig. 3: Shortlisting Leve

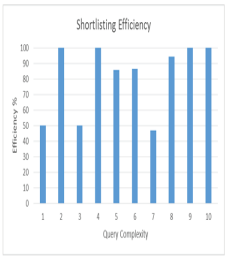

Fig. 4: Shortlisting Efficiency

Scalability is one of the major issues faced by the current information retrieval systems. Enormity of the available data and high levels of complexity associated with the data, time complexity of the processing algorithms tend to escalate. This makes such techniques inefficient when applied on Big Data. Scalability with respect to query complexity and document sizes are presented in figures 5 and 6. Time taken by the algorithms with increase in the complexity of queries is shown in figure 5. It could be observed that as the complexity increases, the time taken for retrieval also increases. However, at a complexity level 9, it is found that the increase in time reaches a saturation limit. This shows that a saturation limit exists for the algorithm, after which,

the processing time satiates to a constant. This property of the proposed algorithm is attributed to the parallelized nature of operations and the default time requirement for storage and retrieval in the Hadoop File System. However, after crossing a threshold in the complexity level, this time stabilizes, making the algorithm scalable.

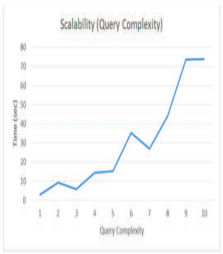

Fig. 5: Scalability (Query Complexity)

Scalability of the algorithm in terms of document size is presented in figure 6. Size of the base repository was increased incrementally from 100 documents to 1000 documents. Query with the highest complexity (complexity level 10) was applied to the repository. It could be observed from figure 6 that as the document size increases, the time taken for processing also increases. The graph corresponds to a linear increase in time, exhibiting a linear scalability of the algorithm with respect to the document size.

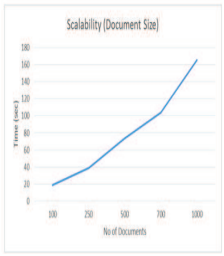

Fig. 6: Scalability (Document Size)

Efficiency of the sentiment analysis module is evaluated using the Kaggle Sentiment analysis data. This is a Twitter based data, designed for supervised learning. The results are recorded and the ROC and PR plots were constructed as in figures 7 and 8. The ROC plot is constructed with the true positive rates and the false positive rates. It could be observed that the points occupy the top left quadrant of the plot, exhibiting very low false positive rates (0) and very high true positive rates (0.89-1). This exhibits that the proposed algorithm performs effectively by correctly classifying 90% of the data.

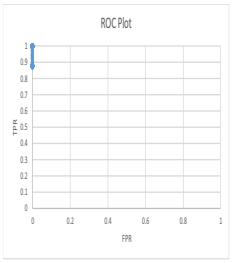

Fig. 7: ROC Plot

Table 1: Additional Performance Metrics.

Metric	Value
TPR	0.950069
FPR	0
Recall	0.950069
Precision	1
Accuracy	0.959966
F-Measure	0.973501
TNR	0.373134
FNR	0.049931

The Precision Recall (PR) plot helps identify the efficiency of the retrieval rates of the algorithm under analysis. High precision and high recall rates are expected of an effective algorithm. It could be observed from figure 8 that the proposed algorithm exhibits high precision levels (1.0) and high recall levels (0.89-1). This exhibits the high performance nature of the algorithm.

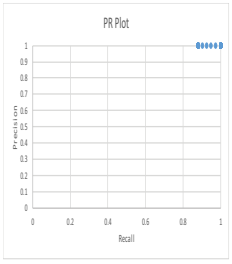

Fig. 8: PR Plot

Several other metrics that are used to measure the performance of a sentiment identification system is presented in Table 1.

Overall average of the performance metrics are presented in Table 1. The values for these metrics range between 0 and 1. It could be observed that true positive rates, precision and recall exhibits very high levels of efficiencies, while the false positive rates exhibit a value of 0. The proposed technique exhibits an accuracy of 0.959 (96%) and an F-Measure of 0.97, both indicating

excellent prediction levels. However, it could be observed that the true negative rate is very low at 0.37.

5 Conclusion

This paper proposes an effective technique for business processes decision making. Decision making in products usually involves customer reviews. However, obtaining the appropriate reviews and processing them is a very complex task, requiring several levels of processing. The proposed technique presents an architecture that effectively performs tasks from retrieval of data to parsing the data, analysis and segregation of results to identify the root cause. A query, if presented to the processing architecture, is processed and the major causes related to the query are presented to the user. The results are segregated in terms of their polarity, simplifying the analysis further. The limitations of this approach includes low true negative rates, making the prediction biased towards positive results. Future works of the authors will incorporate mechanisms such as bigram or n-gram analysis to reduce the true negative rates, making the sentiment prediction section more robust. Though information retrieval is performed in terms of the context of the document, the proposed work does not handle sarcasm. Future works can also be incorporated with components to identify sarcasm for effective data grouping.

References

[1] Bayraktar, A., Uslay, C. and Ndubisi, N.O., 2015. The role of mindfulness in response to product cues and marketing communications. International Journal of Business Environment, 7(4), pp.347-372.

[2] Canavan, K.T. and Martin, A., 2015, January. Delay in referral to hot foot clinic; a root cause analysis and suggestions for service improvement. In BMC Proceedings (Vol. 9, No. Suppl 1, p. A42). BioMed Central Ltd.

[3] Damele, G., Bazzana, G., Andreis, F., Aquilio, F., Arnoldi, S. and Pessi, E., 1996. Process improvement through root cause analysis. In Achieving Quality in Software (pp. 35-47). Springer US.

[4] Dubey, G., Rana, A. and Ranjan, J., 2016. A research study of sentiment analysis and various techniques of sentiment classification. International Journal of Data Analysis Techniques and Strategies, 8(2), pp.122-142.

[5] Zhang, B., Song, Q., Ding, J. and Wang, L., 2015. A trust-based sentiment delivering calculation method in microblog. International Journal of Services Technology and Management, 21(4-6), pp.185-198.

[6] Zhao, H., Ji, X., Zeng, Q. and Jiang, S., 2016. A teaching evaluation method based on sentiment classification. International Journal of Computing Science and Mathematics, 7(1), pp.54-62.

[7] Zakos, J. and Verma, B., 2006. A novel context-based technique for web information retrieval. World Wide Web, 9(4), pp.485-503.

[8] Golitsyna, O.L. and Maksimov, N.V., 2011. Information retrieval models in the context of retrieval tasks. Automatic Documentation and Mathematical Linguistics, 45(1), pp.20-32.

[9] Brown, P.J. and Jones, G.J., 2001. Context-aware retrieval: Exploring a new environment for information retrieval and information filtering. Personal and Ubiquitous Computing, 5(4), pp.253-263.

[10] Loyall, J.P. and Schantz, R.E., 2009, November. Using context awareness to improve quality of information retrieval in pervasive computing. In IFIP International Workshop on Software Technolgies for Embedded and Ubiquitous Systems (pp. 320-331). Springer Berlin Heidelberg.

[11] Carrillo, M. and Lpez-Lpez, A., 2010, October. Concept based representations as complement of bag of words in information retrieval. In IFIP International Conference on Artificial Intelligence Applications and Innovations (pp. 154-161). Springer Berlin Heidelberg.

[12] Winograd, M., 2007. Sustainability and vulnerability indicators for decision making: lessons learned from Honduras. International Journal of Sustainable Development, 10(1-2), pp.93-105.

[13] Balusamy, B., Murali, T., Thangavelu, A. and Krishna, P.V., 2015. A multi-level text classifier for feedback analysis using tweets to enhance product performance. International Journal of Electronic Marketing and Retailing, 6(4), pp.315-338.

[14] Robertson, S.E. and Okapi, S.W., 1999. Keenbow at TREC-8 In Proceedings of the TREC 8 Conference. National Institute of Standards and Technology.

[15] Salton, G. and Yang, C.S., 1973. On the specification of term values in automatic indexing. Journal of documentation, 29(4), pp.351-372.

[16] Salton, G., Wong, A. and Yang, C.S., 1975. A vector space model for automatic indexing. Communications of the ACM, 18(11), pp.613-620.

[17] Walker, S., Robertson, S.E., Boughanem, M., Jones, G.J. and Jones, K.S., 1997, November. Okapi at TREC-6 Automatic ad hoc, VLC, routing, filtering and QSDR. In TREC (pp. 125-136).

[18] Fishburn, P.C., 1967. Letter to the editoradditive utilities with incomplete product sets: application to priorities and assignments. Operations Research, 15(3), pp.537-542.

[19] Triantaphyllou, E., 2000. Multi-Criteria Decision Making: A Comparative Study. Dordrecht, the Netherlands: Kluwer Academic Publishers (now Springer). p. 320.ISBN 0-7923-6607-7.

[20] Lovins, J.B., 1968. Development of a stemming algorithm (p. 65). Cambridge: MIT Information Processing Group, Electronic Systems Laboratory.

[21] Porter, M.F., 1980. An algorithm for suffix stripping. Program, 14(3), pp.130-137.

[22] https://tartarus.org/martin/PorterStemmer/

[23] Baccianella, S., Esuli, A. and Sebastiani, F., 2010, May. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 2200-2204).

[24] Baeza-Yates, R. and Ribeiro-Neto, B., 1999. Modern information retrieval (Vol. 463). New York: ACM press.

[25] http://www.sananalytics.com/lab/twitter-sentiment/

[26] Tamine-Lechani, L., Boughanem, M. and Daoud, M., 2010. Evaluation of contextual information retrieval effectiveness: overview of issues and research. Knowledge and Information Systems, 24(1), pp.1-34.

S. Subitha has received her Master of Philosophy (M.Phil) in Computer Science from Periyar University, India in the year 2008 and also her Post Graduate Degree (MCA) from Bharathidasan University , India in the year 1997. Presently she is a research scholar of Anna University Chennai. She has published 8 papers in national and international conferences and 3 papers in international Journals. She is a keen researcher in web data mining techniques.

S. Sujatha is a Doctorate in Computer Science and having twenty years of experience in teaching with good knowledge in the area of Computer science and Information Technology, currently working as Associate Professor, Bharathidasan Institute of Technology, Anna University, Tirchirappalli, India. She has published 20 research papers in reputed journals, international and national conferences. She has received .Active Researcher Award, Anna University, Chennai in the year 2013. She is a recognized research supervisor in Anna University, Chennai. Her areas of interest include Distributed Computing, Web services, Data Mining, Cloud Computing and its applications