International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 10 (2017), pp. 2297-2308

http://www.ripublication.com

Feedback based Root-Cause Identification Model with User Profiler for Effective Content Correlation

S.P. Victor [1] and S. Charles Britto [2]

[1]*Associate Professor, Department of Computer Science, St Xavier's College, Tamilnadu, India.*

[2]*Research Scholar, Bharathiar University, Tamilnadu, India.*

Abstract

Massive amounts of data available online has led to a lot of models leveraging the available knowledge for decision making. However, identifying the appropriate data for analysis seems to be a daunting task. This paper proposes a framework that can be used to effectively identify and retrieve appropriate content from large repositories. The proposed model builds the user's profile based on context, sentiment pairs and frequency of usage of information pertaining to the pair. This leads to the building up of an effective user profile that can be utilized to retrieve content appropriate to the user when a query is presented to the system. Further, incorporation of the final feedback helps in continuous improvement of the user's profile such that frequent usages will lead to better predictions. Comparisons with the existing sentiment analysis models indicate high efficiency in the prediction levels of the proposed feedback based root cause analysis model in terms of Accuracy, F-Measure, True and False Prediction levels.

Keywords: Root Cause Analysis; Sentiment Analysis; User Profiling; Polarity Identification; Significant Term Identification; TF-IDF

1. INTRODUCTION

The exploding growth of social media has led to the huge information sharing habit. The information being shared is a representative of personalized content. Such content are rich in context and semantic orientations. Online media is currently not only famous for its information acceptance level, but also serves as a reliable repository of

information that can also be effectively leveraged for use. Leveraging information is one of the critical aspects, due to the huge size of data and the possibility of a large number of matches from the repository. Although the search terms aims to reduce the search domain to a large extent, there still exists several subdomains with huge amounts of data that matches the current search. This in turn leads to overwhelming number of results returned to the user. This issue can be effectively solved if the search term can be effectively associated with an additional semantic and contextual information, which can be used to narrow down the domain further.

User profiling is one of the major models currently on the raise. With the spread of interconnected devices, it could be observed that every user has their own distinct digital footprint. This can be assimilated to form a behavior pattern that can be effectively used to predict the likes and dislikes of a user. User profiling is a model that has been currently under research for its powerful nature of operation and the high accuracy level associated with it in terms of information association. This paper presents another usability domain for user's profile data. It deals with identifying the content appropriate to the user's requirements during the process of information retrieval.

2. RELATED WORKS

Root cause identification using sentiment analysis and information extraction has been one of the major research domains currently on the raise due to the improved platforms for information sharing. This section discusses some of the recent contributions in the domain of sentiment analysis.

A product rating model using sentiment analysis was presented by Pham et al. in [1]. This model proposes a multiple layered architecture that is used for representing different sentiment levels. These representations are provided to neural networks and a prediction model is created, to predict overall ratings of product ratings. Other machine learning based prediction models include, Latent Rating Regression (LRR) [2,3] that analyzes the aspect ratings and aspect weights for prediction, deep learning models using convolutional neural networks [4], deep memory neural networks [5] and long short term memory networks [6]. A method using bag-of-words (BOW) model and neural network based learning was proposed by Zhao et al. in [7]. This provides an aggregated approach to identifying the best model for sentiment analysis. A survey of models used for document annotation was proposed by Koncz et al. in [8]. This model analyzes all the available active learning models to facilitate their usage in annotation and sentiment analysis.

The rise of microblogging sites has led to the usage of text in microblogs to perform sentiment analysis. A cross media public sentiment analysis model was proposed by Cao et al. in [9]. This model fuses text and image sentiment levels to provide an aggregated fusion based model for sentiment identification. A similar model using

Twitter data for sentiment analysis was proposed by Pandarachalil et al. in [10]. The three sentiment lexicons, SenticNet, SentiWordNet and SentislangNet are used to identify polarity levels for the process of prediction. The proposed unsupervised sentiment analysis model is proposed to provide effective F-Scores. Other similar microblog based sentiment analysis frameworks include a multimodal learning model by Huang et al. in [11] and a cuckoo search based sentiment analysis model by Pandey et al. in [12].

Health based social media sentiment analytics have also been on the raise with the evolution of social media content. A framework to analyze health based data from user communities was proposed by Yang et al. in [13]. This model proposes three operational phases namely; medical term extraction model, virtual document clustering and cluster analysis for sentiments. Another similar model that performs sentiment analysis of HPV vaccine related tweets was proposed by Du et al in [14]. Some of the very recent contributions in the domain of sentiment analysis includes multi-lingual sentiment analysis. This includes a technique comparison model by Dashtipour et al. in [15] and Chinese online review analysis by Zheng et al. in [16]. Other similar sentiment analysis models include an emoticon based sentiment analysis model by Nirmal et al. in [17], fuzzy sequential sentiment analysis model by Charles Britto et al. in [18] and a parallel sentiment analysis model by Charles Britto et al. in [19].

3. FEEDBACK BASED ROOT-CAUSE IDENTIFICATION MODEL

Root cause analysis is the process of identifying the core or base of the given query. The identification process is usually performed by matching the context and the semantic correlation between the query and the document. This paper proposes a feedback based root cause analysis model that creates a user's profile to identify results not just based on query, but also based on the user's general requirements. The proposed architecture has two major components; the profiler and the semantic correlation identifier. The profiler gathers and organizes data corresponding to the user, which can be utilized while ranking the results. The semantic correlation identifier analyses the results, identifies the semantic correlation between the result, the input query and user's profile and ranks the results. The overall architecture is presented in figure 1.

3.1. Profiling

Profiling is the process of identifying the user's behavior in terms of their interests. Interests vary in terms of information context and the sentiment associated with the information. Hence user profiles are built based on context and sentiment of the input query. Building a user's profile requires continuous monitoring of the user's searches and their selections. Hence this is a continuous monitoring module activated on every query and every result selection.

As the user inputs a query, the profiler identifies the significant terms from that query

and records them. A sentiment identification of the terms is also performed. However, queries need not necessarily be oriented towards a sentiment. They can be neutral terms, expecting a result set with a particular sentiment. Hence, if the sentiment was not identified, it is marked as neutral and included in the user's profile. Profile data ranking is also performed by identifying the number of times a particular term with a particular sentiment has been queried for results. This enables better correlation with the user's requirements. Selection of a particular result also creates an impact on the profile, by either appending the context and sentiment or if the context sentiment pair is already present, its significance level is increased.

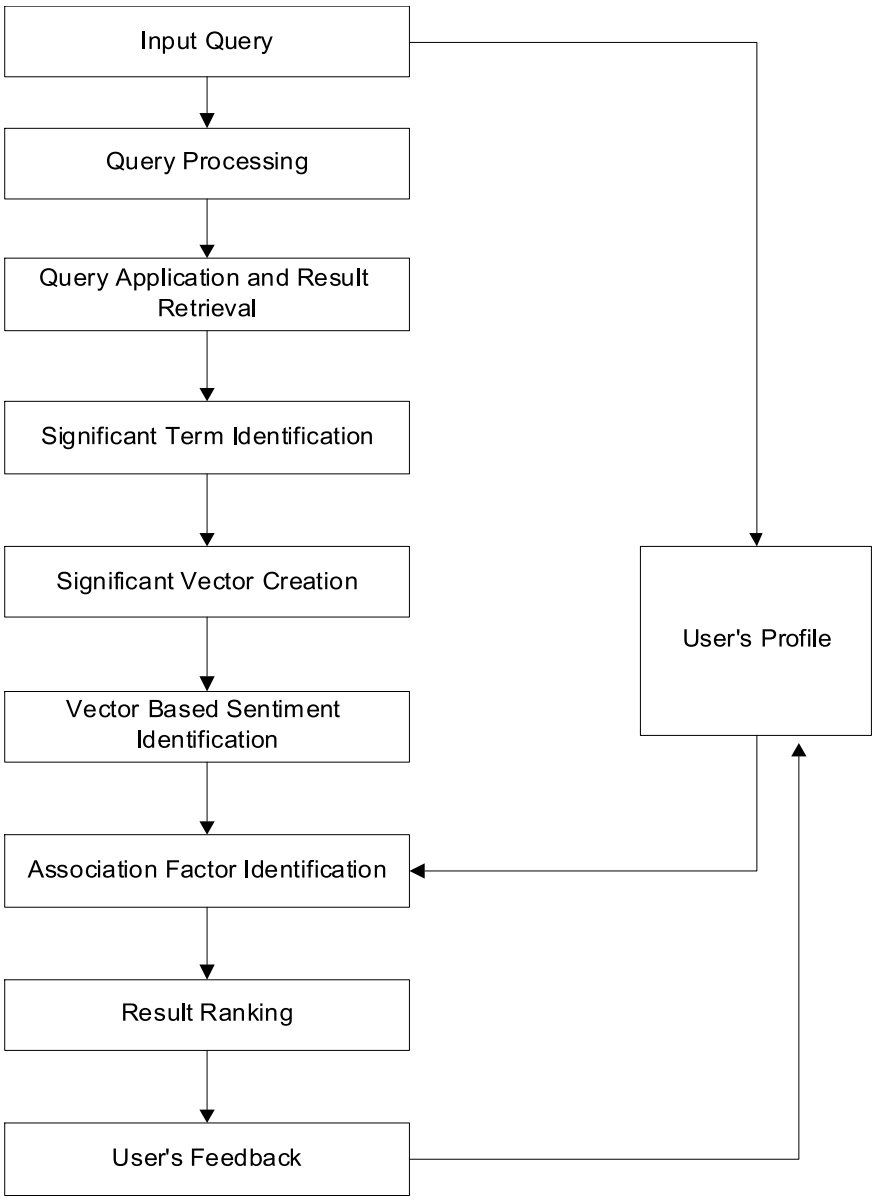

Figure 1: Feedback based Root-Cause Identification Model – Architecture

3.2. Semantic Correlation Identifier

This module retrieves the results from the query and identifies its semantic intensity, matches it with the user's profile, ranks them and finally provides ordered results to the user for analysis.

3.2.1. Query Processing and Result Retrieval

The input query presented by the user is tokenized and the tokens are passed to the profiler to be recorded. Significant terms and their corresponding sentiments are extracted from the tokens, followed by profile building. In parallel to this process, the query modified in accordance with the data source on which it is to be applied using wrappers. Customized query are then applied on the data source and the results are retrieved.

3.2.2. Significant Token Identification and Significant Vector Creation

Every result is treated as a vector and significant terms are identified from each of the vectors and the significant vectors are created. Results are usually in the form of text or documents. Hence they tend to contain several additional content other than the required information. The general vocabulary necessitates the usage of connectors between words to provide a sequence to the sentence. Such connectors are to be eliminated for effective analysis of context and sentiment. In order to simplify this process, significant term identification is performed by identifying the Term Frequency (TF) and the Inverted Document Frequency (IDF) for the tokens [20]. Tokens with higher values of TF-IDF are considered to be significant terms while the tokens with low values of TF-IDF are considered to be terms of low significance.

TF-IDF is calculated using equation 1.

$$tfidf(t, d, D) = tf(t, d) \times idf\ (t, D) \tag{1}$$

Term Frequency (TF) and the Inverted Document Frequency (IDF) are calculated using (2) and (3)

$$tf(t, d) = \frac{f(t,d)}{count(w,d)} \tag{2}$$

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*.

$$idf(t, D) = log \frac{N}{|\{d \in D : t \in d\}|} \quad (3)$$

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*. If the term is not in the corpus, then it will lead to a divide-by-zero error, hence it is also common to adjust the denominator to1+$|\{d \in D : t \in d\}|$.

A user based threshold for the TF-IDF value is used to filter the most appropriate terms from the terms of low significance. These filtered terms are used to create the significant term repository (*STR*). Significant vectors are created by cross verifying the terms in results with *STR* to identify and eliminate terms that are of low significance.

3.2.3. Vector based Sentiment Identification

The major requirement for creating the significant vectors is that sentiment identification is to be performed on per result basis. Each vector created from the results represents a single result entry. Identifying sentiment for each of these vectors will provide record based sentiment levels that can effectively aid result segregation.

Sentiment identification is performed using SentiWordNet polarity repository [21]. SentiWordNet is a human annotated repository that contains terms with their polarity levels from positive and negative dimensions. Every term in the vector is associated with a polarity value. An aggregation of these polarity values provide the final sentiment of the vector. Sentiment of a vector *v* is given by

$$Sentiment_v = \sum_{i=1}^{n} Polarity(t_i)$$

Where $Polarity(t_i)$ refers to the polarity of the i^{th} term in the vector *v*.

Polarity identification is usually performed with the aid of the input query. If the input query has a sentiment related to it, the polarity intensity of the particular sentiment alone is considered for analysis. However, if the input query was identified to be neutral, the polarity dimensions are aggregated and used as the final polarity value. The aggregation is given by,

$$Polarity(t) = Polarity_{(pos,t)} - Polarity_{(neg,t)}$$

Where $Polarity_{(pos,\ t)}$ refers to the positive polarity associated with the term *t* and $Polarity_{(neg,t)}$ refers to the negative polarity associated with the term *t.*

This phase results in the creation of vectors with a sentiment level associated with it, and each vector representing a single result obtained from the input query.

3.2.4. User Profile based Significance Integration

Although sentiment based categorization of results is possible at this stage, they only correspond to the sentiment levels and not the actual user requirements. Hence a profile based significance value is integrated with the vectors to provide appropriate associations with the user's requirements. The user's profile is built with sentiment, context and their frequency of occurrence. Frequency is considered to represent the significance factor for the context related to the particular sentiment. Each term of the sentiment vectors is matched with the user's profile for the frequency level and the frequency factors are aggregated to provide the Association Factor (*AF*) for the vector. Hence every vector now contains the sentiment level factor, representing the intensity of the sentiment levels and the Association Factor that represents the association levels of the user's profile.

3.2.5. Result Ranking and User Feedback Incorporation

Ranking of the vectors is performed in two phases. The first phase ranking is performed using the Association Factor and the second phase ranking is performed using the sentiment levels, within a group of vectors with the same AF value. The results are presented to the user.

Although the results are ranked based of the user's profile, ultimately, it is the user who determines the relevancy level of the presented results. Their selections and order of selection play a vital role in providing improved accuracy levels. Result selections and the order in which the results are selected are provided to the user's profile in terms of feedbacks. These feedbacks enable profile updates for providing precise association levels in the future.

4. RESULTS AND DISCUSSION

Experiments were performed using STS Gold Sentiment Corpus [22], which is a human annotated set of tweets. Comparisons were performed with the emoticon based sentiment analysis model [17], the Sequential Sentiment Analysis model [18] and Parallel Sentiment Analysis model [19]. Analysis is performed in terms of True and False Prediction levels, Accuracy and F-Measure.

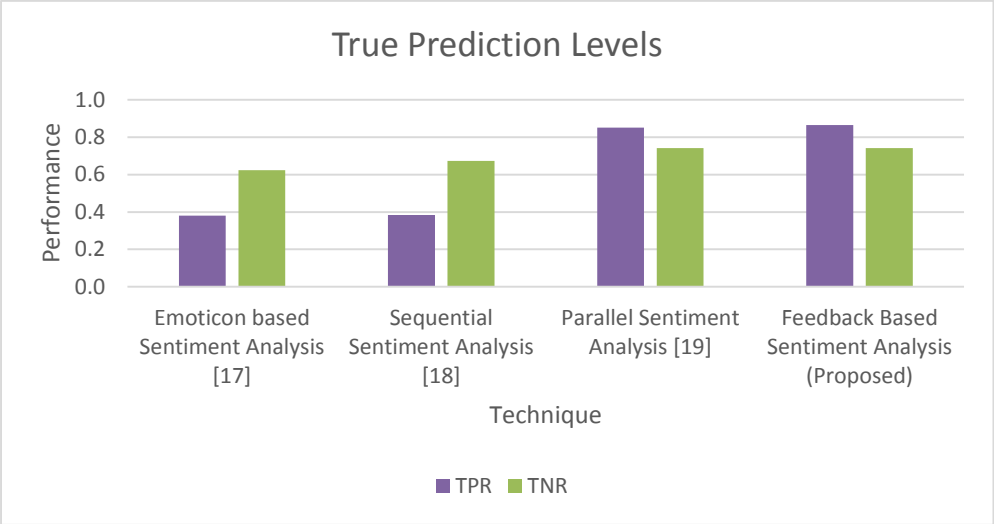

Figure 2: True Prediction Levels (Sentiment Analysis) - STS Gold Sentiment Corpus

The true prediction levels of the proposed algorithm depicting the true positive rates (TPR) and the true negative rates (TNR) are shown in the figure 2. An effective algorithm should exhibit high TPR and TNR levels, depicting effective prediction of both positive and the negative classes. It could be observed from the figure that the proposed feedback based sentiment analysis model exhibits very high TPR and TNR levels showing the efficiency of the proposed architecture.

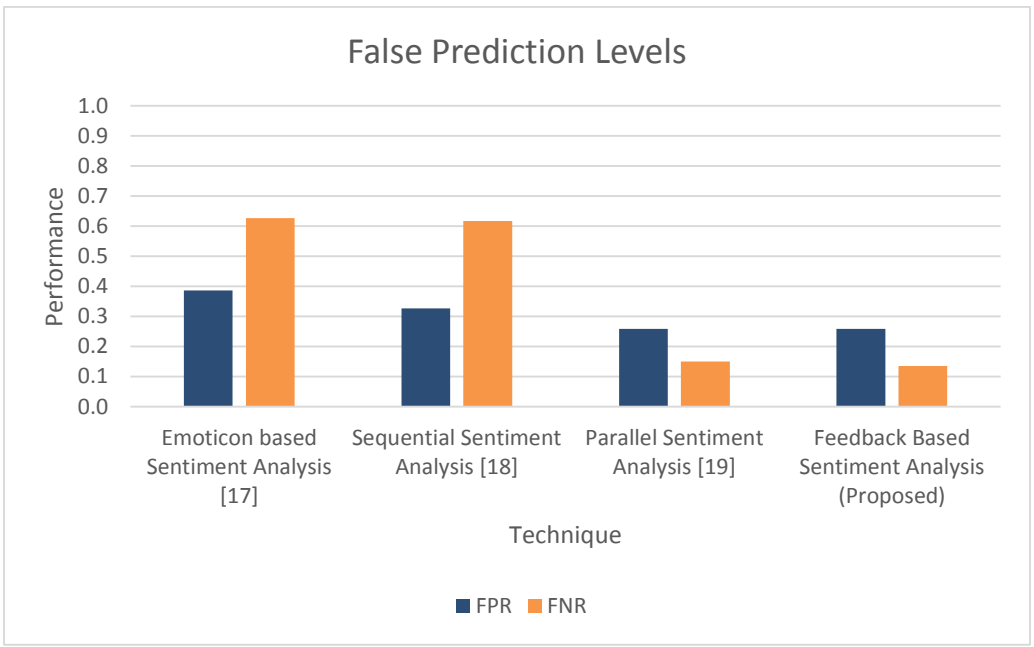

Figure 3: False Prediction Levels (Sentiment Analysis) - STS Gold Sentiment Corpus

The false prediction levels of the proposed algorithm depicting the false positive rates (FPR) and the false negative rates (FNR) are shown in the figure 3. An effective algorithm should exhibit very low FPR and FNR levels. The FPR and FNR levels are inversely proportional to the TNR and TPR levels respectively. It could be observed from the figure that the proposed model exhibits very low false prediction levels, exhibiting the efficiency of the algorithm.

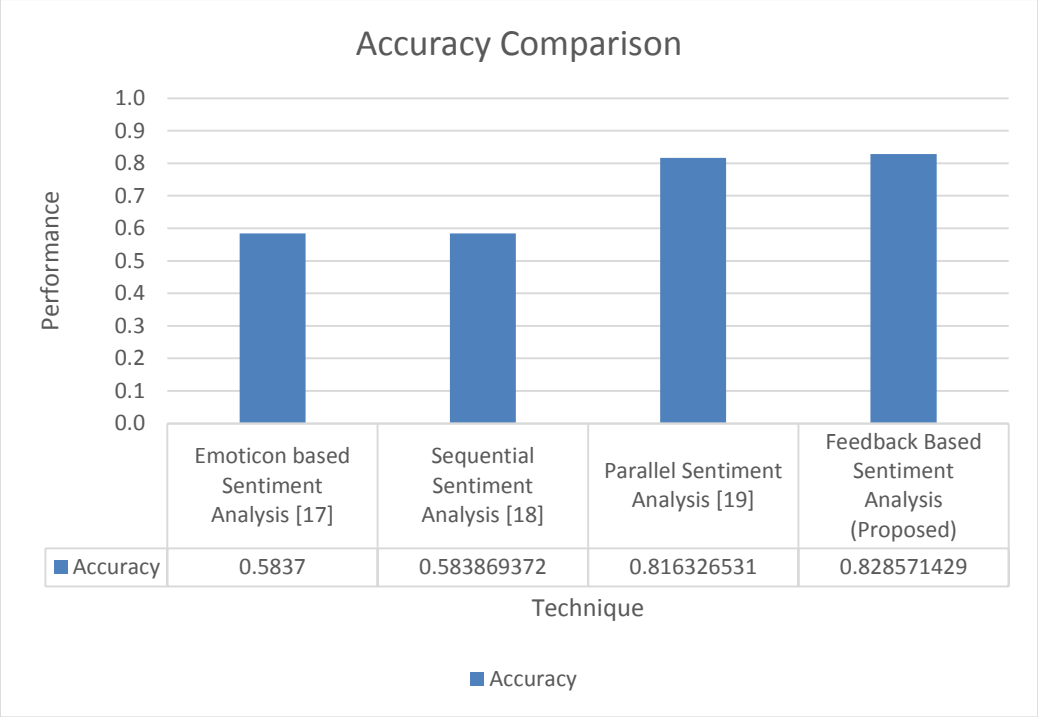

Figure 4: Accuracy Comparison (Sentiment Analysis) - STS Gold Sentiment Corpus

A comparison of the accuracy levels of the proposed feedback based sentiment analysis model is shown in figure 4. It could be observed that the proposed model exhibits 24% higher accuracy compared to emoticon based and sequential sentiment analysis models [17,18] and 1.2% improved accuracy compared to the parallel sentiment analysis model [19]. F-Measure levels of the proposed model is shown in figure 5. It could be observed that a correlation similar to accuracy can be observed in the F-Measure levels.

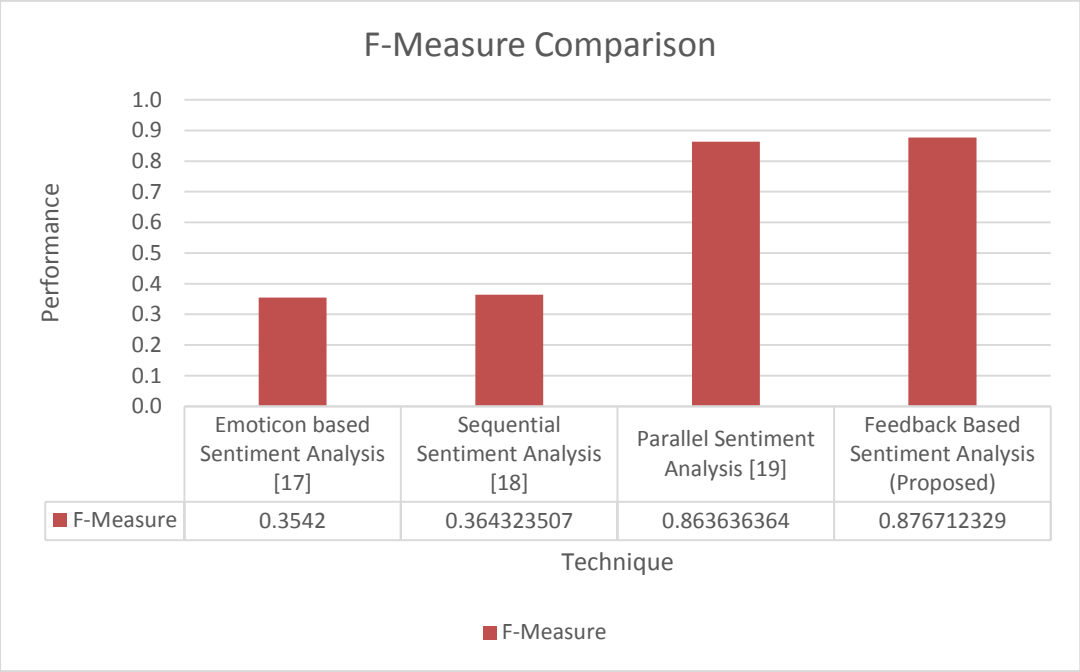

Figure 5: F-Measure Comparison (Sentiment Analysis) - STS Gold Sentiment Corpus

Table 1: Performance Metrics

Metrics	Emoticon based Sentiment Analysis	Sequential Sentiment Analysis	Parallel Sentiment Analysis	Feedback Based Sentiment Analysis
Accuracy	0.5837	0.583869372	0.816326531	0.828571429
F-Measure	0.3542	0.364323507	0.863636364	0.876712329
TNR	0.6240	0.674084709	0.741935484	0.741935484
TPR/ Recall	0.3800	0.383757962	0.850746269	0.864864865
Precision	0.3500	0.34676259	0.876923077	0.888888889
FNR	0.6262	0.616242038	0.149253731	0.135135135
FPR	0.3859	0.325915291	0.258064516	0.258064516

All the performance metrics used for analysis and the results obtained from each of the compared models is shown in Table 1. It could be observed that the proposed model exhibits excellent prediction levels in terms of all the proposed metrics, in comparison to the existing models in literature.

5. CONCLUSION

User requirement based content extraction from a huge repository is a complex task, which can be effectively performed only with accurate domain information. This paper presents an effective feedback based model that can be used to filter appropriate content from a large repository of data. User's queries and selections are used to build user's repositories. Results for the input queries are correlated with the user's profile to identify results that match the user's requirements. Although accurate matches can be achieved, the user might also need to view content that is not related to them. Hence the results are ranked and presented to the user instead of being filtered. Experiments on STS Gold Sentiment Corpus indicates high prediction levels. Comparison with existing models also indicate highly effective performances. The limitations of this method includes moderate False Positive levels. Future extensions of the proposed work will provide effective mechanisms to reduce the FP levels.

REFERENCES

[1] Pham, D.H. and Le, A.C., 2017. Learning Multiple Layers of Knowledge Representation for Aspect Based Sentiment Analysis. Data & Knowledge Engineering.

[2] Wang, H., Lu, Y. and Zhai, C., 2010. Latent aspect rating analysis on review text data: a rating regression approach. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 783-792). ACM.

[3] Wang, H., Lu, Y. and Zhai, C., 2011. Latent aspect rating analysis without aspect keyword supervision. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 618-626). ACM.

[4] Poria, S., Cambria, E. and Gelbukh, A., 2016. Aspect extraction for opinion mining with a deep convolutional neural network. Knowledge-Based Systems, 108, pp.42-49.

[5] Tang, D., Qin, B. and Liu, T., 2016. Aspect level sentiment classification with deep memory network. arXiv preprint arXiv:1605.08900.

[6] Wang, Y., Huang, M., Zhu, X. and Zhao, L., 2016. Attention-based LSTM for Aspect-level Sentiment Classification. In EMNLP (pp. 606-615).

[7] Zhao, Z., Liu, T., Li, S., Li, B. and Du, X., 2017. Guiding the Training of Distributed Text Representation with Supervised Weighting Scheme for Sentiment Analysis. Data Science and Engineering, 2(2), pp.178-186.

[8] Koncz, P. and Paralič, J., 2013, September. Active learning enhanced document annotation for sentiment analysis. In International Conference on Availability, Reliability, and Security (pp. 345-353). Springer, Berlin, Heidelberg.

[9] Cao, D., Ji, R., Lin, D. and Li, S., 2016. A cross-media public sentiment analysis

system for microblog. Multimedia Systems, 22(4), pp.479-486.

[10] Pandarachalil, R., Sendhilkumar, S. and Mahalakshmi, G.S., 2015. Twitter sentiment analysis for large-scale data: an unsupervised approach. Cognitive computation, 7(2), pp.254-262.

[11] Huang, F., Zhang, S., Zhang, J. and Yu, G., 2017. Multimodal learning for topic sentiment analysis in microblogging. Neurocomputing, 253, pp.144-153.

[12] Pandey, A.C., Rajpoot, D.S. and Saraswat, M., 2017. Twitter sentiment analysis using hybrid cuckoo search method. Information Processing & Management, 53(4), pp.764-779.

[13] Yang, F.C., Lee, A.J. and Kuo, S.C., 2016. Mining health social media with sentiment analysis. Journal of medical systems, 40(11), p.236.

[14] Du, J., Xu, J., Song, H., Liu, X. and Tao, C., 2017. Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. Journal of biomedical semantics, 8(1), p.9.

[15] Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A.Y., Gelbukh, A. and Zhou, Q., 2016. Erratum to: Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques. Cognitive Computation, 8(4), pp.772-775.

[16] Zheng, L., Wang, H. and Gao, S., 2015. Sentimental feature selection for sentiment analysis of Chinese online reviews. International journal of machine learning and cybernetics, pp.1-10.

[17] Nirmal, V.J. and Amalarethinam, D.G., 2016. Emoticon based Sentiment Analysis using Parallel Analytics on Hadoop. Indian Journal of Science and Technology, 9(33).

[18] S. Charles Britto and Victor S.P., 2017. Improving Root Cause Analysis Using Fuzzy Polarity Identification and Conflict Resolution Techniques, Journal of Information Technology Research. (Accepted on March 2017)

[19] S. Charles Britto and Victor S.P., Real-time root cause identification on streaming heterogeneous data using spark.

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

[21] 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).

[22] Saif, H., Fernandez, M., He, Y. and Alani, H., 2013. Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold.