Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method

H. M. Keerthi Kumar[1], B. S. Harish[2], H. K. Darshan[3] *

[1] JSSRF, JSS TI Campus, Mysuru, Karnataka (India)
[2] Department of Information Science and Engineering, JSS Science and Technology University, Mysuru, Karnataka (India)
[3] Department of Information Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysuru, Karnataka (India)

Received 12 June 2018 | Accepted 17 November 2018 | Published 7 December 2018

unir
LA UNIVERSIDAD EN INTERNET

Abstract

Social Networking sites have become popular and common places for sharing wide range of emotions through short texts. These emotions include happiness, sadness, anxiety, fear, etc. Analyzing short texts helps in identifying the sentiment expressed by the crowd. Sentiment Analysis on IMDb movie reviews identifies the overall sentiment or opinion expressed by a reviewer towards a movie. Many researchers are working on pruning the sentiment analysis model that clearly identifies and distinguishes between a positive review and a negative review. In the proposed work, we show that the use of Hybrid features obtained by concatenating Machine Learning features (TF, TF-IDF) with Lexicon features (Positive-Negative word count, Connotation) gives better results both in terms of accuracy and complexity when tested against classifiers like SVM, Naïve Bayes, KNN and Maximum Entropy. The proposed model clearly differentiates between a positive review and negative review. Since understanding the context of the reviews plays an important role in classification, using hybrid features helps in capturing the context of the movie reviews and hence increases the accuracy of classification.

Keywords

Classification, Hybrid Features, Short Text, Sentiment Analysis.

DOI: 10.9781/ijimai.2018.12.005

I. Introduction

SOCIAL media has become an integral part of human living in recent days. People want to share each and every happening of their life on social media. Nowadays, social media is used for showcasing one's pride or esteem by posting photos, text, video clips, etc. The text plays a vital aspect in information shared, where users share their opinions on trending topics, politics, movie reviews, etc. These opinions which people share on social networking sites are generally known as Short Texts (ST) because of its length [1]. ST have gained its importance over traditional blogging because of their simplicity and effectiveness in influencing the crowd. They are even used by search engines in the form of queries. Apart from their popularity, ST has certain challenges like identification of sarcasm, sentiment, use of slang words, etc. Therefore it becomes important to understand short texts and derive meaningful insights from them, which is generally known as Sentiment Analysis (SA) [2].

SA played an important role in the US Presidential Elections 2016 [3]. People shared their likes and dislikes regarding a particular political party on micro-blogs such as Twitter and Facebook. Those blogs where analyzed and candidates pruned their tweets based on these analyses. Thus, SA helped them to increase their popularity and followers. SA is widely used by most of the companies because of its capacity to analyze a large number of documents at once, which manually would take more time. In the business sector, companies use SA to derive new strategies based on the customer feedback [4].

Reviews are short texts that generally express an opinion about movies or products. These reviews play a vital role in the success of movie or sales of the products [5]. People generally look into blogs, review sites like IMDb to know about movie cast, crew, review and ratings. Hence it is not only the Word of Mouth that brings the audience to the theatres; reviews also play a prominent role in this regard. In other words, SA on movie reviews makes the task of Opinion Summarization [6] easier by extracting the sentiment expressed by the reviewer.

The task of SA on movie reviews mainly include – Preprocessing [7], Feature Extraction followed by Selection [8], Classification [9] and finally the analysis of results. Preprocessing involves removal of stop words, abbreviating short forms, replacing slangs, etc. which are scrutinize for the task of classification. Feature Extraction involves identifying the features that represent the documents in the vector space. Many feature extraction [10] methods that exist will extract the features from the reviews, mainly by statistical based and lexicon based approaches. In statistical feature extraction methods [11], the words present in the review are used as features by calculating various weighing measures like Term Frequency (TF), Inverse Document Frequency (IDF) and Term Frequency-Inverse Document Frequency (TF-IDF) [31]. In Lexicon [12] based feature extraction methods, textual features are extracted by deriving the patterns among the words, deriving from Parts of Speech of the words tagger, using Lexicon Dictionaries, etc. Lexicon based methods generally capture the semantics of the text by considering the ordering of text in the

* Corresponding author.
E-mail addresses: hmkeerthikumar@gmail.com (H. M. Keerthi Kumar), bsharish@jssstuniv.in (B. S. Harish), bharadwajdarshan@gmail.com (H. K. Darshan).

review. Hybrid approaches [13] involving both Statistical and Lexicon based feature extraction methods will increase the overall accuracy of the model. Once the extraction of features is done, relevant features are identified using feature selection methods [14] which eliminate the features that do not contribute towards effective classification. Classification involves identifying the polarity of the review and classifying it as either positive or negative sentiment.

This paper proposes a Hybrid method, where the features are extracted by using both statistical and lexicon methods. In addition, we apply various feature selection methods such as Chi-Square, Correlation, Information Gain and Regularized Locality Preserving Indexing (RLPI) [15] for the features extracted by statistical methods. This maps the higher dimension input space to the lower dimension input space. The Lexicon based feature extraction method extract features based on the Lexicon dictionaries. Features from both methods are combined to form a new feature set which is of lower dimension when compared to the initial dimension of the input space. The new features set is classified using various classifiers such as Support Vector Machines (SVM), Naïve Bayes (NB), K- Nearest Neighbor (KNN) and Maximum Entropy (ME) classifiers on IMDb movie review dataset.

The contents of the paper are divided into five sections. Section II presents an overview of the literature survey of previous works on sentiment analysis. Section III presents the methodology. Section IV shows experimental results and the paper is concluded in section V.

II. Literature Survey

The process of Sentiment Analysis involves the construction of the input vector space from the existing document vector space. Mainly there are two approaches to carry out vector space mapping. The machine learning based or statistical based feature extraction methods are widely used because extraction of features is done by applying statistical measures directly. Earlier works on sentiment classification using machine learning approaches were carried by Pang et al. in 2002 [16]. Sentiment analysis was performed on IMDb movie reviews using n-gram approaches and Bag of Words (BOW) as features. The model was trained using different classifiers like Naïve Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM). The unigram features outperformed when compared to other features [16]. Similar work was done by Tripathy et al. [5], where TF, TF-IDF was used for the conversion of the text file to a numerical vector. Experimentation was done with n-gram approaches and its combination are tried to get the best results.

Apart from the word features which are considered for the classification task, special symbols which are present with words -known as emoticons (☺, ☹, 😐, …) can also be used as features. Neetu et al. [17] used these special features along with the word features. The use of an ensemble classifier which classifies based on the results obtained by different classifiers like NB, ME and SVM is the major highlight of the work. Many researchers have worked on extracting features based on the parts of speech tagger. Geetika et al. [18] used unigram model to extract adjective as a feature which in turn describes the positivity or negativity of the sentence.

Identifying the semantics or the meaning of the text by a machine learning algorithm is a challenging task. Lexicon features are used in this regard to extract the opinions expressed in the text. Sarcasm detection is one of the major advantages of choosing lexicon features. Anukarsh et al. [19] focused on the slangs and emojis which were present in the text to detect sarcasm. Use of slang and emoji dictionaries during preprocessing increased the efficiency of sarcasm detection. Capturing the sentiment orientation of the text towards a topic helps in identifying the overall polarity of the text. Taboda et al., in [12], used dictionaries to calculate the Semantic Orientation (SO) and termed it as Semantic Orientation CALculator (SO-CAL). Various factors such as Parts of Speech (Adjectives, Nouns, Verbs and Adverbs), Intensifiers (Somewhat, Very, Extraordinary etc.,), Negations, etc., were considered to calculate sentiment orientation. Results showed that the Lexicon based sentiment analysis gives better results and can be applied to wide domains. Similarly in [32], Dehkharghani developed lexicon for sentiment analysis.

Melville et al. [20], worked on extracting features using lexicon methods. Positive and negative word counts that are present in the text were used as the background lexicon knowledge and then the probability that a document belongs to a particular class was calculated. Use of pooling multinomial classifiers which incorporate both training examples and the background knowledge is the major contribution. Kolchyana et al., in [21], used both machine learning and lexicon approaches to perform sentiment analysis on Twitter data. Special lexicon features such as N-grams, Lexicon sentiment, Elongated words number, Emoticons, Punctuations, etc., were used. Use of these features increased the overall accuracy of the model. The hybrid method combines the features generated by both machine learning approach and lexicon approach. Use of a hybrid approach reduces the complexity of the overall model by retaining only the important features and thus increases time efficiency. The main advantage of using the lexicon features is that it captures the meaning or the semantics expressed in the reviews thereby contributing to the effective classification. The experimental results showed that the review classification was more accurate because of the use of semantics of the review as a feature and is comparable with the human review classification.

The polarity of a review depends on the intensity of each word present in the review and the context used by the reviewers to express their opinion. Therefore, identifying the features that extract the intensity of words based on context that inclines the polarity either towards positive or negative polarity is a challenging task. The proposed work captures the polarity of a word and determines how important the word is for the classification task. The capturing phase is done through the features generated using Hybrid Feature Extraction Method (HFEM). The HFEM combines the reduced Machine learning features with the Lexicon features to increase the performance of the model.

The major contribution of this paper includes:

- Identifying the lexicon features such as Positive word count, Negative word count, Positive Connotation count, Negative Connotation count, which helps to identify the semantics of the reviews.
- Use of RLPI feature selection method to reduce high dimensional features.
- Comparison of the classification accuracy and F-measure of Machine learning feature, lexicon feature and HFEM using different supervised learning algorithms.

III. Methodology

In the proposed model, sentiment analysis is employed on IMDb Movie Reviews. The input for the proposed model is the set of reviews whose polarity needs to be determined. The output corresponds to reviews with polarity assigned to each of them. The task of sentiment analysis is carried out in the following phases: preprocessing the dataset, feature Extraction (Both Statistical and Lexicon approach), feature selection and finally classification using hybrid features. Fig. 1 gives the overall workflow of the proposed model.

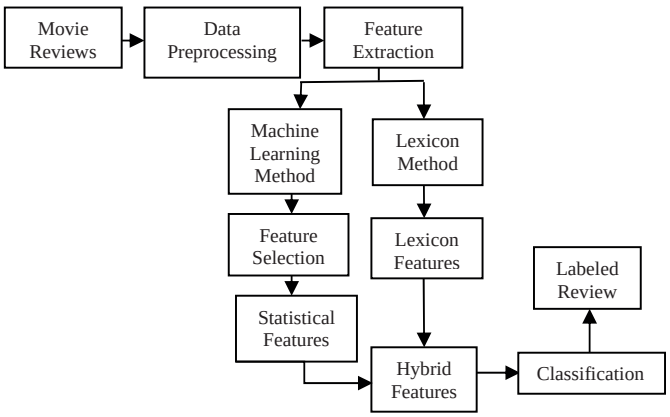

Fig. 1. Proposed model for sentiment classification.

A. Preprocessing

The reviews which need to be analyzed consist of words, numbers, and special symbols as its constituents. Consider the following review, "The great and underrated Marion Davies shows her stuff in this late (1928) silent comedy that also showcases the wonderful William Haines. A must for any serious film buff or for anyone interested in the still-maligned Marion Davies!". The review contains the year 1928 and punctuations like '.', '!', ',' that does not influence on sentiment analysis because of its neutral polarity [33]. Hence the numbers and punctuations are removed. Many words such as 'a', 'an', 'the', 'should', etc., which are commonly known as stopwords are also eliminated. Many words that are present in the reviews will not be in their root forms. For example, words like 'studying', 'studied' belong to same root word 'study'. This process is known as Lemmatization [22] where the ineffectual endings of the words are removed by bringing to the root form with the help of vocabulary. Further, the preprocessed dataset is used for feature extraction in the next phase.

B. Feature Extraction

Feature Extraction identifies the features that have a positive effect towards classification. In this work, feature extraction is carried in two different parallel stages namely- Machine learning based feature extraction and Lexicon based features extraction.

Machine learning based feature extraction method is used to extract the features using popularly known technique Bag of Words, wherein the column corresponds to words and row corresponds to value of weighing measures such as Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF).

Lexicon based feature extraction method which is used in the proposed work extracts 4 different features from the review. They are as follows: Positive word Count (PC), Negative word Count (NC), Positive Connotation Count (PCC) and Negative Connotation Count (NCC).

Positive and negative words present in the reviews are identified by using Positive word dictionary and Negative word dictionary respectively. Connotation refers to the abstract meaning of the word depending on the context. For example, "Desire" is a positive connotation which is absent in the positive dictionary and "Avoid" is a negative connotation which is absent in the negative dictionary. Hence, Positive and Negative connotation lexicon is used in addition to the regular positive and negative words dictionary to identify PCC and NCC.

Combination of features extracted through different feature extraction methods will increase the overall performance of the model. Combining machine learning based features with the features extracted by using a Positive-Negative lexicon and Positive-Negative connotation lexicon is the key idea of the paper. This combination of features helps in identifying the overall polarity of the review more accurately. Use of Hybrid Feature Extraction Method (HFEM) is the major contribution of this paper.

C. Feature Selection

Features extracted in the previous phase may contain hundreds to thousands of features for a small set of reviews. Handling large number of features increases the burden on the classification algorithms. Feature selection methods are generally applied to reduce the dimension of the feature space by selecting only the important features. Reduction in the feature dimension should not affect the classification accuracy. Importance of a feature is calculated by various statistical methods. Various conventional feature selection methods such as Chi-Square [23], Correlation [24], and Information Gain [25] are used in literature to perform sentiment analysis. Many feature selection methods which are used in other domains such as medical image processing, document indexing and clustering etc., can also be used to perform sentiment analysis effectively. RLPI [15] is one such unconventional feature selection method which is generally used for document indexing and representation. RLPI reduces the dimension of the features by performing Eigen vector decomposition on feature space and then selects top Eigen vectors to represent the features. Thus, RLPI helps in handling large number of features which can be further reduced to smaller dimension feature space. Use of RLPI features along with Lexicon features for testing and training the learning algorithms is the major contribution of this paper.

In the proposed approach, we apply feature selection method to the feature set generated by an statistical approach because of its larger dimension. On the other hand, we extracted 4 features using the lexicon approach. The selected feature set from statistical approach is combined with the Lexicon features to form a matrix which is the input for the classification algorithm.

D. Classification

Classification is the process of assigning labels to the reviews whose label is unknown. In the proposed work, supervised learning algorithms such as Naïve Bayes [17], Maximum Entropy [16], Support Vector Machines (SVM) [13] and K- Nearest Neighbor (KNN) [26] [30] are used. Naïve Bayes and Maximum Entropy work on the principles of probability and hence they are known as probabilistic classifiers. Naïve Bayes works on the principle of independence of features and calculates the probability of a review belonging to particular class using Bayes theorem. Maximum Entropy classifies the review by calculating the conditional probability. Maximum Entropy does not assume independence of features. SVM is independent of the number of features in the feature space. SVM uses a hyper plane to separate the samples of two classes. KNN classifies the review with the unknown label by comparing it with the reviews present in the test data. The comparison is done by applying various similarity measures and identifies the most similar review (nearest neighbor). Further, it assigns the label of nearest neighbor to the unlabeled review.

IV. Experimental Setup

A. Dataset

IMDb is the most commonly used website for getting information about a movie throughout the world. Because of its popularity and due to the presence of large number of reviews related to a particular movie, IMDb Movie Review Dataset [27] is used in the proposed work. It is one of the standard benchmark datasets used for Sentiment Analysis on Movie reviews. The dataset contains 25,000 positive and negative reviews each. However, due to the limitations of computational resources, we have randomly chosen 5000 reviews for experimentation.

B. Lexicons

In the proposed work, the following lexicons are used: 1. Opinion Lexicon created by Hu et al., [28] which is used as Positive-Negative Lexicon. It consists of around 6800 words including positive and negative words. 2. Connotation Lexicon created by Feng et al., [29] which contains positive and negative connotations.

C. Experimentation

The experimentation is carried on a machine running on Ubuntu 16.04 operating system with R Studio version 3.4.2 environment. For experimentation, 5-fold validation technique is used i.e., 5000 reviews were randomly selected from the dataset and were again split into 5 batches containing 1000 reviews in each batch. The final result shown in Table I is the average of results obtained in the 5 batches. Initially, 13,346 features are extracted using machine learning feature extraction method. Because of its larger dimension, feature selection methods such as IG, Correlation, Chi-Square were applied. The reduced dimension of the input space was varied from 10% to 60% of the initial number of features. Feature count is varied from 1000 to 8000 features and the best results were obtained for the feature counts 2000, 5000 and 8000. In case of RLPI feature selection method, the reduced dimension of the input space was varied from 50 to 150 because lesser number of features are sufficient to represent feature in terms of Eigen vectors and hence feature count varies around 1% of the initial feature count. Better results were obtained for fewer feature counts like 50, 100 and 150. Four lexicon features namely Positive word count, Negative word count, Positive Connotation count and Negative Connotation count are used along with the reduced machine learning features. The hybrid features are used to train the model using learning algorithms such as Naïve Bayes, SVM, Maximum Entropy and KNN.

D. Result Analysis and Discussion

This section presents the analysis of results obtained using various classifiers with different feature selection methods (FSM). Machine learning features described in this section consists of features generated using different weighing schemes like TF, TF-IDF. When these features are merged with the Lexicon features mentioned previously, it generates the Hybrid features. Fig. 2-4 show the comparison of accuracy obtained using Machine Learning features, Lexicon features and Hybrid features on classifiers like SVM, NB, ME and KNN during 5 batches of experimentation. The accuracies shown in the figures are the highest accuracies in that particular batch using different FSM on different classifiers.

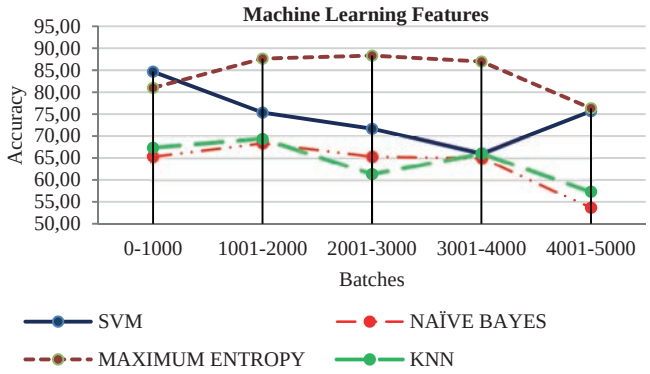

Fig. 2. Comparison of accuracy obtained using Machine Learning features for 5 batches.

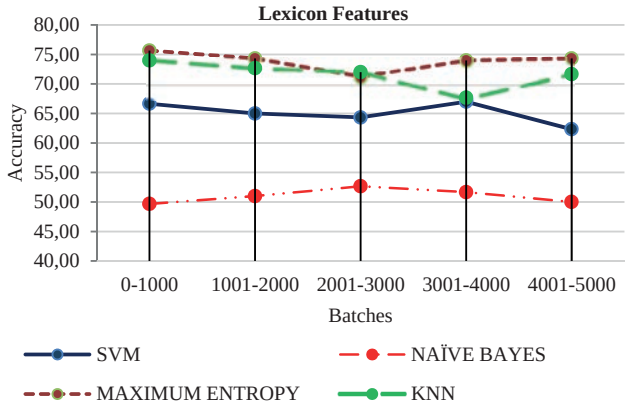

Fig. 3. Comparison of accuracy obtained using Lexicon features for 5 batches.

When compared to various feature selection methods, RLPI gives the best accuracy of 74.66% with less number of features using Maximum Entropy classifier (see Table I). This is because RLPI selects the discriminating features which are smaller in number but are highly sufficient to represent the feature space. This can be considered as the best result since the complexity of the input data is reduced to 0.4% of the original dimension. However, accuracy is the tradeoff when we consider lesser complex input data which contains less number of features. Maximum Entropy is the best performing classifier with the highest accuracy of 83.93% when correlation is used as a feature selection method, because the features selected by correlation are highly correlated with the class and have effective contribution towards classification. Hybrid features outperforms machine learning features and gives the best result in terms of both accuracy and F-measure irrespective of the feature selection method and classifier used. Further, addition to lexicon features along with machine learning features i.e., use of hybrid features increases the accuracy using SVM classifier with Chi-Square feature selection method.

The results summarized in Table I can be analyzed by considering two parameters namely complexity of the input data and highest accuracy achieved. Percentage of improvement in number of features, accuracy and F-measure is presented in Table I. The percentage change corresponds to the percentage increase or decrease in number of features, accuracy and F-measure of Hybrid features when compared to the Machine Learning Features.

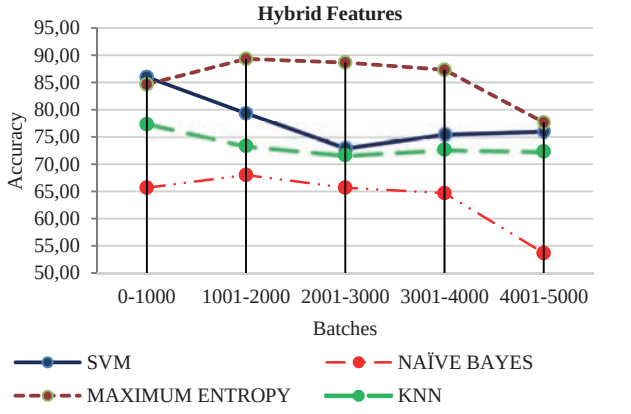

Fig. 4. Comparison of accuracy obtained using Hybrid features for 5 batches.

It is evident from the Table I that the highest percentage of decrease in the input space is about 37.5% and the highest increase in percentage of accuracy is 36.38% and that of F-measure is 78.19%. The results obtained by using Hybrid Feature Extraction Method are promising both in terms of accuracy, F-measure and complexity.

TABLE I. Results Using FSM with Classifiers Using Hybrid Features

FS Method	Classifier	Hybrid Features			% of improvement in no. of features	% of improvement in accuracy	% of improvement in F-measure
		Number of features	Accuracy	F-Measure			
IG	SVM	5000	75.467	0.752	-37.5	32.554	59.660
	Naïve Bayes	8000	54.733	0.608	60	2.62	10.144
	Maximum Entropy	**8000**	**78.333**	**0.780**	**0**	**7.600**	**8.18**
	KNN	5000	72.267	0.723	150	24.454	17.84
Correlation	SVM	5000	76.600	0.764	150	34.859	75.632
	Naïve Bayes	2000	60.667	0.573	0	-0.108	-0.174
	Maximum Entropy	**5000**	**83.933**	**0.837**	**-37.5**	**1.94**	**2.32**
	KNN	2000	72.000	0.721	0	26.021	15.733
Chi Square	SVM	5000	75.467	0.752	-37.5	36.386	78.199
	Naïve Bayes	8000	54.733	0.608	0	0	0
	Maximum Entropy	**8000**	**78.333**	**0.780**	**0**	**7.600**	**8.18**
	KNN	5000	72.267	0.723	150	24.454	21.717
RLPI	SVM	100	73.600	0.734	0	1.939	2.370
	Naïve Bayes	50	63.400	0.567	0	0	-0.176
	Maximum Entropy	**50**	**74.667**	**0.739**	**0**	**2.564**	**2.354**
	KNN	150	71.933	0.720	200	24.022	22.033

V. Conclusion

Sentiment Analysis on short informal text is a challenging task. Due to the limited number of characters, huge dimensional features and sparseness, which increases complication. In this paper, Hybrid Feature Extraction Method (HFEM) is used to extract features from machine learning and lexicon based feature extraction methods. Initially, machine learning features are in high dimensional in nature. The feature selection methods such as Information Gain, Correlation, Chi Square and RLPI are applied on the machine learning features to reduce high dimensional features. On the other hand, lexicon based features such as Positive word Count (PC), Negative word Count (NC), Positive Connotation Count (PCC) and Negative Connotation Count (NCC) are extracted. Combining machine learning features with the lexicon features captures the orientation of words and thus identifies the context of the review. To demonstrate the effectiveness of the proposed work, we used four different classifiers such as SVM, KNN, Maximum Entropy and Naïve Bayes on IMDb movie review dataset. The Maximum Entropy with correlation shows the best results in terms of both accuracy and F-measure when compared to other classifiers. Use of Hybrid Feature Extraction Method (HFEM) makes the model more efficient in terms of accurate classification by adding the advantages of individual feature extraction method. HFEM improves the space complexity by reducing the input space to minimal number of features that are sufficient to represent the review content. Thus, results obtained are highly promising both in terms of space complexity and classification accuracy. In future work, we will include more lexicon features to the feature subset and thereby expect to increase the classification accuracy.

Acknowledgment

H M Keerthi Kumar has been financially supported by UGC under Rajiv Gandhi National Fellowship (RGNF) Letter no: F1-17.1/2016-17/RGNF-2015-17-SC-KAR-6370/(SA-III Website), JSSRF (University of Mysore), Karnataka, India.

References

[1] C. D. Santos and M. Gatti. “Deep convolutional neural networks for sentiment analysis of short texts.” In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 69-78.

[2] A. Ortigosa, J. M. Martín, and R. M. Carro. “Sentiment analysis in Facebook and its application to e-learning.” Computers in Human Behavior Vol. 31, pp.527-541. 2014.

[3] R. Ahmad, A. Pervaiz, P. Mannan, and F. Zaffar. “Aspect Based Sentiment Analysis for Large Documents with Applications to US Presidential Elections 2016.” Social Technical and Social Inclusion Issues (SIGSI), 2017, pp. 13.

[4] K. Xu, S. S. Liao, J. Li, and Y. Song. “Mining comparative opinions from customer reviews for Competitive Intelligence.” Decision support systems, Vol. 50, no. 4, pp.743-754. 2011.

[5] A. Tripathy, A. Agrawal, and S.K. Rath. “Classification of sentiment reviews using n-gram machine learning approach.” Expert Systems with Applications, Vol. 57, pp. 117-126. 2016.

[6] M. E. Moussa, E. H. Mohamed, and M. H. Haggag. “A survey on Opinion Summarization Techniques for Social Media.” Future Computing and Informatics Journal (2018). In press.

[7] I. Hemalatha, G. P. S. Varma, and A. Govardhan. “Preprocessing the informal text for efficient sentiment analysis.” International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) 1, no. 2: pp.58-61. 2012.

[8] A. S. Manek, P. D. Shenoy, M. C. Mohan, and K. R. Venugopal. “Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier.” World Wide Web Vol. 20, no. 2, pp.135-154. 2017.

[9] A. Kennedy and D. Inkpen. “Sentiment classification of movie reviews using contextual valence shifters.” Computational intelligence, Vol. 22, no. 2, .pp.110-125. 2006.

[10] M. Z. Asghar, A. Khan, S. Ahmad, and F. M. Kundi. “A review of feature extraction in sentiment analysis.” Journal of Basic and Applied Scientific Research, Vol. 4, no. 3, pp.181-186. 2012.

[11] A. Sharma and S. Dey. “A comparative study of feature selection and machine learning techniques for sentiment analysis.” In Proceedings of the 2012 ACM research in applied computation symposium, pp. 1-7. ACM, 2012.

[12] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. “Lexicon-

based methods for sentiment analysis." Computational linguistics, Vol. 37, no. 2, pp.267-307. 2011.

[13] A. Mudinas, D. Zhang, and M. Levene. "Combining lexicon and learning based approaches for concept-level sentiment analysis." In Proceedings of the first international workshop on issues of sentiment discovery and opinion mining, pp. 5. ACM, 2012.

[14] L. Zheng, H. Wang, and S. Gao. "Sentimental feature selection for sentiment analysis of Chinese online reviews." International journal of machine learning and cybernetics, Vol. 9, no. 1, pp.75-84. 2018.

[15] D. Cai, X. He, W. V. Zhang, and J. Han. "Regularized locality preserving indexing via spectral regression." In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 741-750, ACM, 2007.

[16] B. Pang, L. Lee, and S. Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, pp. 79-86. 2002.

[17] M. S. Mubarok, Adiwijaya, and M. D. Aldhi. "Aspect-based sentiment analysis to review products using Naïve Bayes." In AIP Conference Proceedings, vol. 1867, AIP Publishing, no. 1, pp 1-8.2017.

[18] G. Gautam, and D. Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." In Contemporary computing (IC3), 2014 seventh international conference on, pp. 437-442. IEEE, 2014.

[19] A. G. Prasad, S. Sanjana, S. M. Bhat, and B. S. Harish. "Sentiment analysis for sarcasm detection on streaming short text data." In Knowledge Engineering and Applications (ICKEA), 2017, 2nd International Conference on, pp. 1-5. IEEE, 2017.

[20] P. Melville, W. Gryc, and R. D. Lawrence. "Sentiment analysis of blogs by combining lexical knowledge with text classification." In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 1275-1284. 2009.

[21] O. Kolchyna, T. T. P. Souza, P. Treleaven, and T. Aste. "Twitter sentiment analysis: Lexicon method, machine learning method and their combination." arXiv preprint arXiv:1507.00955. 2015.

[22] Y. Bao, C. Quan, L. Wang, and F. Ren. "The role of pre-processing in twitter sentiment analysis." In International Conference on Intelligent Computing, pp. 615-624. Springer, 2014. Cham.

[23] J. Brooke, M. Tofiloski, and M. Taboada. "Cross-linguistic sentiment analysis: From English to Spanish." In Proceedings of the international conference RANLP-2009, pp. 50-54. 2009.

[24] L. Deng, Y. Hu, J. P. Y. Cheung, and K. D. K. Luk. "A Data-Driven Decision Support System for Scoliosis Prognosis." IEEE Access 5, pp. 7874-7884. 2017.

[25] F. K. Ahmad. "Comparative Analysis of Feature Extraction Techniques for Event Detection from News Channels' Facebook Page." Journal of Telecommunication, Electronic and Computer Engineering (JTEC) Vol. 9, no. 1-2 , pp.13-17. 2017.

[26] H. M. Kumar, B. S. Harish, S. V. Kumar, and V. N. Aradhya. "Classification of sentiments in short-text: an approach using mSMTP measure". In Proceedings of the 2nd International Conference on Machine Learning and Soft Computing. pp. 145-150. ACM. 2018.

[27] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. "Learning word vectors for sentiment analysis." In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1, Association for Computational Linguistics, pp. 142-150. 2011.

[28] M Hu, and B. Liu. "Mining and summarizing customer reviews." In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 168-177. 2004.

[29] S. Feng, J. S. Kang, P. Kuznetsova, and Y. Choi. "Connotation lexicon: A dash of sentiment beneath the surface meaning." In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 1774-1784. 2013.

[30] K. S. Srujan, S. S. Nikhil, H. Raghav Rao, K. Karthik, B. S. Harish, and H. M. Kumar. "Classification of Amazon Book Reviews Based on Sentiment Analysis." In Information Systems Design and Intelligent Applications, pp. 401-411. Springer, Singapore, 2018.

[31] M. B. Revanasiddappa, B. S. Harish. A New Feature Selection Method based on Intuitionistic Fuzzy Entropy to Categorize Text Documents, International Journal of Interactive Multimedia and Artificial Intelligence, (2018), http://dx.doi.org/10.9781/ijimai.2018.04.002

[32] R. Dehkharghani. Building Phrase Polarity Lexicons for Sentiment Analysis, International Journal of Interactive Multimedia and Artificial Intelligence, (2018), http://dx.doi.org/10.9781/ijimai.2018.10.004

[33] H. M. Kumar and B. S. Harish. "Classification of Short Text Using Various Preprocessing Techniques: An Empirical Evaluation." In Recent Findings in Intelligent Computing Techniques pp. 19-30. Springer, Singapore, 2018.

H. M. Keerthi Kumar

He received B.E in Information Science and Engineering and M.Tech in Software Engineering from Visvesvaraya Technological University, India. He is currently pursuing Ph.D degree in Computer Science from University of Mysore, India. His area of research includes Data Mining, Pattern Recognition and Machine Learning.

B. S. Harish

He obtained his B.E in Electronics and Communication (2002), M.Tech in Networking and Internet Engineering (2004) from Visvesvaraya Technological University, Belagavi, Karnataka, India. He completed his Ph.D. in Computer Science (2011); thesis entitled "Classification of Large Text Data" from University of Mysore. He is presently working as an Associate Professor in the Department of Information Science & Engineering, JSS Science & Technology University, Mysuru. He was invited as a Visiting Researcher to DIBRIS - Department of Informatics, Bio Engineering, Robotics and System Engineering, University of Genova, Italy from May-July 2016. He delivered various technical talks in National and International Conferences. He has invited as a resource person to deliver various technical talks on Data Mining, Image Processing, Pattern Recognition, and Soft Computing. He is also serving and served as a reviewer for National, International Conferences and Journals. He has published more than 50 International reputed peer reviewed journals and conferences proceedings. He successfully executed AICTE-RPS project which was sanctioned by AICTE, Government of India. He also served as a secretary, CSI Mysore chapter. He is a Member of IEEE (93068688), Life Member of CSI (09872), Life Member of Institute of Engineers and Life Member of ISTE. His area of interest includes Machine Learning, Text Mining and Computational Intelligence.

H. K. Darshan

He is currently pursuing B.E in Information Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysuru, Karnataka, INDIA, His area of interest includes Pattern Recognition, Machine Learning, and Text Mining.