International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

# A Dissimilarity based Hybrid Approach Along with Novel Score Exploration for Event Detection in Social Streams

**M. Vijaya Maheswari[1], Dr S. Manju Priya[2]**

[1]PhD Research Scholar, Department of Computer Science, Karpagam Academy of Higher Education, Karpagam University Coimbatore, Tamilnadu, India

[2]Associate Professor, Department of Computer Science, Karpagam Academy of Higher Education, Karpagam University Coimbatore, Tamilnadu, India

**Abstract:** ***The Event detection scheme will helps to detect an event from large dataset and it is also needed preserve the privacy of reviewers because of the social and security issues. A novel recommendation system has been proposed in which depth pre-processing is implemented. In depth pre-processing, the short texts or internet slang acronyms or abbreviations are processed. These short texts are considered to be a mandatory factor for deciding the polarity of user reviews. In the further processing of the user reviews, rather than using the bag-of-words for dictionary compilation, an improved reinforced dictionary formation technique is implemented.A novel feature extraction technique Novel score Exploration(NSE) has been proposed based on considering the positivity, negativity and neutrality along with the a novel user reviews exploration method which will be processed by segregating the user reviews as a sentences, from which the sentence score, position score, modality score and other features are extracted. These additional feature are effective in improving the classification accuracy.***

**Keywords:** social network services, event, event detection, opinion mining, extraction, short text

## 1. Introduction

In recent days number of social media are increased and Millions of people shared their opinion through these type of social media [3].

Due to the increase of social media, its data size has become large and complex to mine the data and to find the reviews shared by the users.

The Event detection scheme will helps to detect an event from large dataset and it is also needed to preserve the privacy of reviewers because of the social and security issues [6].

Event detection and personal information protection were the emerging important research topics in social media analysis. Personnel information and customer's comments were two basic availabilities in social networks which was to analyze the worth or quality of the products[5].

Social sites is a common platform where the customers shares their opinions about the product what they bought. These information shared by the customers would help the manufacturers to know about the status of their products[2].

## 2. Literature Survey

- Wen Hua et.al., proposed a method to Understand Short Texts by Harvestingand Analysing Semantic Knowledge.Semantic knowledge based harvesting was done in this work which increases the difficulty of prediction rate[1].
- Cedric De Boom Steven Van Canneyt et.al., proposed a model for the Representation learning for very short texts using weighted word embedding aggregation. Due to low dimensional representations, sufficient information was not obtained to improve accuracy of classifier[4].
- Zheng Yu et.al., proposed a method for Understanding Short Texts through Semantic Enrichment and Hashing.Encode was created by deep neural network which takes a very long time and the network can provide wrong answer under some circumstances[12].
- Jiaming Xu et.al., proposed a system for Self-Taught convolutional neural networks for short text clustering . Flexible self taught convolutional neural network framework for short text clustering was proposed in this work.Determining the window size in convolutional neural network was very difficult task [11].

## 3. Existing System

Even though, recommendation systems are in high need, the techniques to mine the reviews and to analyze the user's mind-set are critical to identify[9]. Due to the changing trends in the online communications, now-a-days there is also a drastic change from traditional way of communication language to usage of internet slang words and short text over the online social media sites. This makes processing of the user's text reviews to be more complex[5].

**Existing technique & drawbacks**

- Even though, recommendation systems are in high need, the techniques to mine the reviews and to analyze the user's mind-set are critical to identify[13].
- Due to the changing trends in the online communications, now-a-days there is also a drastic change from traditional


Licensed Under Creative Commons Attribution CC BY

way of communication language to usage of internet slang words and short text over the online social media sites [8].
- This makes processing of the user's text reviews to be more complex [10].

## 4. Proposed Work

To overcome the above mentioned issues, a novel recommendation system has been proposed in which **Depth Pre-Processing (DPP)** is implemented. Initially in the preprocessing stage, the raw user reviews will undergo stop words removal, stemming and other preprocessing steps for converting the raw textual user reviews to a useful information [18]. Further from the extracted useful information, the necessary words extraction is done from along with the short texts or internet slang acronyms or abbreviations which will be represented as **Short Text Extraction (STE) Method**. These short texts are considered to be an additional mandatory factor for deciding the polarity of user reviews. In the further processing of the user reviews, rather than using the bag-of-words for dictionary compilation, an **Improved Reinforced Dictionary Formation (IRDF) technique** is implemented.

A novel feature extraction technique has been proposed which will be pooled with the existing **Syllable Count with Weight Extraction (SCWE)** algorithm [7]. This novel technique SCWE is proposed based on considering the positivity, negativity and neutrality along with the **a Novel Score Exploration(NSE) Method** which will be processed by segregating the user reviews as a sentences, from which the sentence score, position score, modality score and other features are extracted. These additional feature are effective in improving the classification accuracy. Also a N**ovel Dissimilarity Based Hybrid Classification Algorithm**is proposed to improve the classification effectiveness. As an added flavor to the proposed work, the user's reviews are sorted based on the user's personal information which will help in user's behavior based recommendation.

**Steps**
- Load the user review dataset, which is the input for processing.
- Consider pre-requisites like Positive word Dictionary, Negative Word Dictionary, Word Net Dictionary which consists of the set of adjectives, adverbs, nouns and verbs[23].
- The dataset consist of the reviewers comments about the products which were taken from social sites. These data are initially divided into user's personal data and user's review contents. Since the data are unstructured textual data, these data are preprocessed[25].
- Using the information available from the SentiWordNet dataset, the score is calculated for each word and then the score is computed for the whole comment.
- The features extracted from the NSE phase is trained to the machine, named as training features.
- Now the Testing data is preprocessed and all the feature are extracted. These features are known as testing features[26].
- Both the testing and training feature are loaded into the proposed DbH classifier.

## 5. Proposed Algorithm

**Pre-Requisites**

Inputs: Word Net Library, SentiWordNet Library, Facebook Review Dataset, User Profile Data

**Dictionary Creation**
- Extract Adjective, Adverb, noun, verb from Word Net Library, Positive and Negative words from library.
- Extract short texts, acronyms or internet slang words and replaced with appropriate words from the library [15].
- Preprocess user data and load User Data

**Training model**

Let us train the machine by extracting the adjective and adverb, positive and negative words from the reviews by comparing with the library to create a dictionary[14].

**Novel Score Exploration Technique (NSE)**

The Score for the obtained reviewer's comment was calculated using the SentiWordNet Dataset. SentiWordNet is a lexical resource for opinion mining. SentiWordNet assigns to each set of WordNet three sentiment scores: positivity, negativity, objectivity.Using the information available from the SentiWordNet dataset, the score is calculated for each word and then the score is computed for the whole comment. This score feature will help the classification algorithm to predict the class of the comment with high accuracy.

**Algorithm**

Input: Reviewer's Comments
Output: Score
Step-1: Bag File is loaded
Step-2: Bag File is separated into Dictionary Files and BagOfWords
Step-3: Dictionary Files were used for examining the grammatical model of reviewer's comments
Step-4: BoW is used to replace the short words present in the reviewer's comment
Step-5: $S\_comment$=Sanitation(reviewer comment) ;
// $S_comment$Sanitized Comment
Step-6: Separate the each word present in the comment, $〚WS〛\_comment\ (i)=Separation(S\_comment\ )$;
// $〚WS〛\_comment\ (i)$ Separated word from the reviewer comment; i=1,2,3,..,N
N total number of words present in the comment
Step-14: // $〚SC〛\_(〚WS〛\_comment\ (i)\ )$ Score for ith word
Step-7: if $〚WS〛\_comment\ (i)==ShortText$
Step-8: Replace $〚WS〛\_comment\ (i)$ by Full form of that short Text
Step-9: Then check the spelling of each word using the word application
Step-10: if Spell($〚WS〛\_comment\ (i)$)~=Correct; clear($〚WS〛\_comment\ (i)$); //the misspelled word was not considered
Step-11: if Spell($〚WS〛\_comment\ (i)$)==Correct;


Licensed Under Creative Commons Attribution CC BY

Step-12: then, Grammatical Checking is carried out using the dictionary file whether the $[\![WS]\!]\_comment\ (i)$ $uounorverboradverboretc$
Step-13: $V\_1\ andV\_2$, for every words in the dictionary files

**Dissimilarity based Hybrid Classification (DbHC)**

- In Dissimilarity based Hybrid classification Algorithm, we coupled both the Euclidean distance measuring algorithm and Bhattacharya distance measuring algorithm.
- In conventional Bhattacharya distance calculation, the mean and variance of training and testing features was considered.
- In proposed DbHC work, instead of variance and mean of training and test features, we considered the temp1, temp2, temp3 and temp4 variables which was calculated as mentioned below
- $temp1 = {(V_{train}\ (i))^2}/{(V_{test})^2}$ ;

  $temp2 = {(V_{test})^2}/{(V_{train}\ (i))^2}$; $temp3 = (M_{train}\ (i) - Mtest)2$; $temp4=(Vtraini+Vtest)2$
- Then, the Euclidean distance is also calculated
- Using the result of Bhattacharya distance and the Euclidean distance, the DbHC classifier provides the result

This module executes in the testing Phase.The features extracted from the previous phase is trained to the machine, named as training features. Now the Testing data is preprocessed and all the feature are extracted based on the previous algorithmic steps[17]. These features are known as testing features. Both the testing and training feature are loaded into the proposed DbHclassifier. It classifies the test samples which class it belongs i.e., Positive or negative or ordinary and provides the accurate result while comparing the existing SVM and PLNN algorithms[16].

**Algorithm**

Input: train, test and label
Output: result
Step-1: for i=1 to $R\_train$
Step-2: $M\_train$= mean(train);
Step-3: $V\_train$= variance(train); end loop
Setp-4: $M\_test$= mean(test);
Step-5: $V\_test$= variance(test);
Step-6: for i=1 to $R\_(V\_train\ )$
Step-7: $temp1$= $[\![(V\_train\ (i))]\!]$ ^2/ $[\![(V\_test)]\!]$^2
Step-8: $temp2$= $[\![(V\_test)]\!]$ ^2/ $[\![(V\_train\ (i))]\!]$^2 ; end for
Step-9: for i=1 to $R\_(M\_train\ )$
Step-10: $temp3$= $[\![\ [\![(M]\!]\ \_train\ (i)-M\_test)]\!]$^2; end for loop
Step-12: $temp4$= $[\![\ [\![(V]\!]\ \_train\ (i)+V\_test)]\!]$^2; end for
Step-13: for i=1 to length($temp1$)
Step-14:
$temp5$=(1/4)∗(log⁡(1/4)∗log⁡($temp1(i)$+$temp2(i)$+2)
)+(1/4)∗(($temp3(i)$)/($temp4(i)$)); end for
Step-15: for i=1 to $R\_train$
Step-16: for j=1 to $C\_train$
Step-17: temp6=[[(train(i,j)~test(1,j))]]^2
Step-18: $temp7$=$temp6$+$temp5$; end for loop
Step-19: [Vtemp6 Itemp6]=min(temp6);
Step-20: [Vtemp5 Itemp5]=min(temp5);
Step-21: if Itemp7== Itemp5
Step-22: result=label(Itemp7);
Step-23: else, if Vtemp7< Vtemp5
Step-24: result=label(Itemp7);
Step-25: else, result= label(Itemp5);
Step-26: end if; end ifStep-11: for i=1 to ($V\_tr$ )

## 6. Experimental Analysis

- The Experimental analysis is carried out on Matlab8.1 in the Windows 7 Platform. The dataset taken for analysis is extracted from a social media dataset, here a Facebook[19]. It is the users comments about the kindle product.The user's varied comments are analyzed by mining the comments which will help to capture sentiments of the comments[20].
- The proposed novel Dissimilarity based Hybrid Classification (DbHC) algorithm are used to extract the sharp features. It has hugely reduced the computational complexity and it has reduced the overall computation time of the system.
- The features extracted from the NSE algorithm has paved way for the effective classification of the DbH classifier algorithm.
- While comparing the existing PLNN model, there is a 35.8% of improvement in the sensitivity measure / true positive rate.

## 7. Results and Discussions

- The comparative results are shown in table 1, that the proposed DBH model is effective in performance and accuracy.
- While comparing the existing PLNN model, there is a 35.8% of improvement in the sensitivity measure / true positive rate.
- There is a approximately 6.1215% increase in the positive predictive rate comparison

**Table 1:** Performance Analysis comparison between PLNN SVM and DbHC

| Comparison Parameters | SVM classifier | PLNN classifier | DbHC Classifier |
|---|---|---|---|
| True Positive | 5 | 52 | 66 |
| True Negative | 920 | 905 | 922 |
| False Positive | 2 | 17 | 15 |
| False Negative | 76 | 29 | 0 |
| Sensitivity | 6.1728 | 64.1975 | 100 |
| Specificity | 99.7831 | 98.1562 | 98.3991 |
| Precision | 71.4286 | 75.3623 | 81.4815 |
| Recall | 6.1728 | 64.1975 | 100 |
| Jaccard Coefficient | 92.2233 | 95.4138 | 98.5045 |
| Dice Coefficient | 95.9544 | 97.6531 | 99.2466 |
| Kappa Coefficient | 0.1481 | 0.9136 | 0.89 |
| Accuracy | 92.2233 | 95.4138 | 98.5045 |


Licensed Under Creative Commons Attribution CC BY

**Accuracy**

Accuracy refers to the closeness of a measured value to a standard or known value. Accuracy is also referred to a weighted arithmetic mean of Precision and Inverse Precision (weighted by Bias) as well as a weighted arithmetic mean of Recall and Inverse Recall (weighted by Prevalence)

$Acc=(TP+TN)/(P+N)$ $or$ $(TP+TN)/(TP+TN+FP+FN)$

The comparison chart of Accuracy between PLNN, SVM and DbHCare shown in Fig1.

Also an approximate of 3.09% improvement in the accuracy of the result retrieval on comparing with the PLNN.

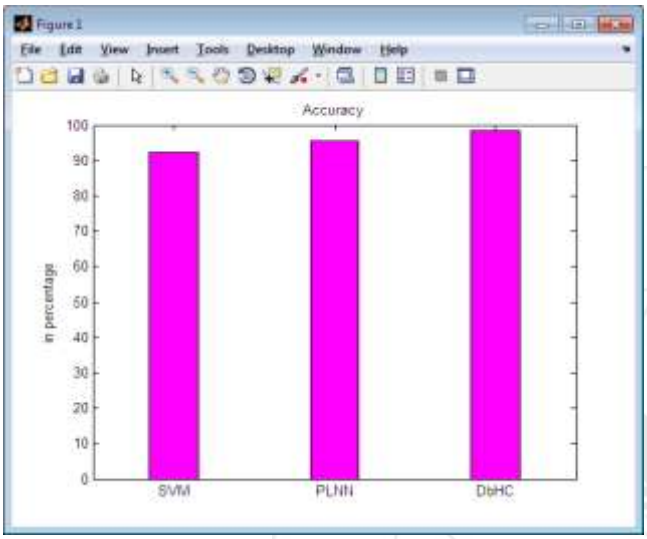


**Figure 1:** Accuracy comparison between PLNN, SVM and DbHC

**Fault Acceptance Rate (FAR)**

False acceptance, also called as a type II error, is a mistake occasionally made by the security systems. In an instance of false acceptance, an unauthorized person is identified as an authorized person[21]. The FAR is defined as the percentage of identification instances in which false acceptance occurs. This can be expressed as a probability.

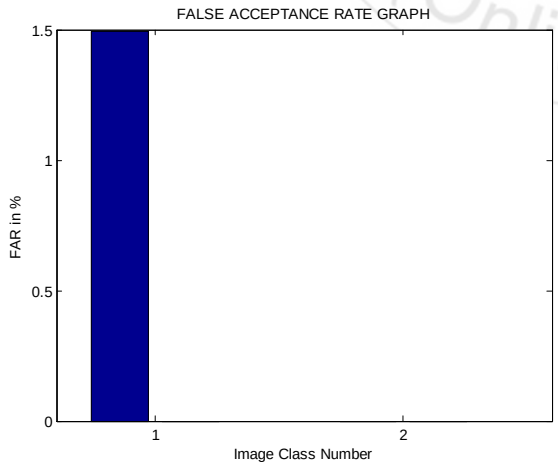


**Figure 2:** False Acceptance Rate Graph

In x axis, Image Class Number
1-No of Positive samples
2- No of Negative samples
Y axis is False Acceptance Rate

**Fault Rejection Rate (FRR)**

This is defined as a percentage of genuine users rejected by the security system.In the verification security system the user will make claims to their identity and hence the system must not reject an enrolled user and number of False Rejections must be kept as small as possible.Thus False Rejection must be minimized in comparison to False Acceptance[22].

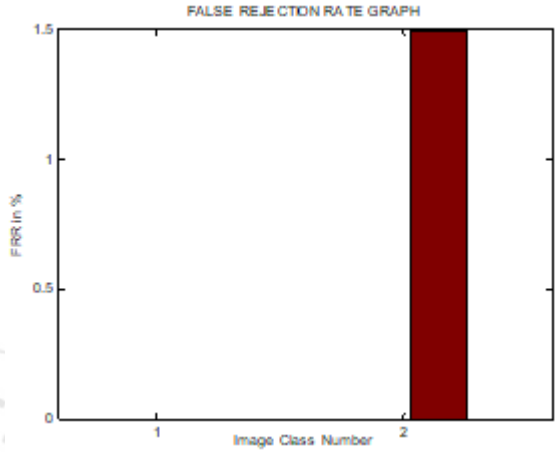


**Figure 3:** False Rejection Rate Graph

In x axis, Image Class Number
1-No of Positive samples
2- No of Negative samples
Y axis is False Rejection Rate

**Genuine/Global Acceptance Rate (GAR)**

The Genuine Accept Rate (GAR) or True Accept Rate (TAR) can be used as an alternate to FRR while reporting the performance of a security verification system[24]. This is defined as a percentage of the genuine users which is accepted by the system. It is given by GAR=100-FRR.

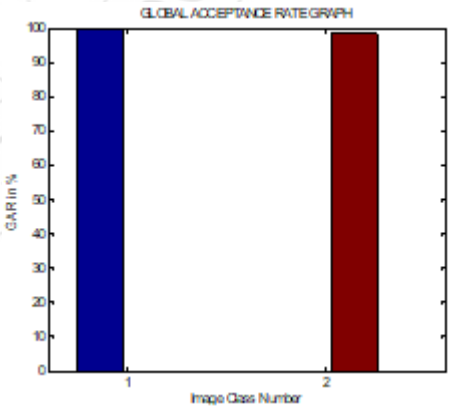


**Figure 4:** Global Acceptance Rate Graph

In x axis, Image Class Number
1-No of Positive samples
2- No of Negative samples
Y axis is Global Acceptance Rate

## 8. Conclusion

A novel Text mining system has been proposed with privacy preservation model to analyze the product information. Along with the privacy Preservation model, the event detection model has undergone a Depth Pattern Component


Licensed Under Creative Commons Attribution CC BY

Analysis Technique which provides the effective feature for machine classification method to provide the results with better accuracy. The proposed work DBH has been fine-tuned with the implication of Word Syllable Count and Weight Extraction Technique along with a Novel Score Exploration approach to extract feature attributes. A novel Dissimilarity based Hybrid Classification technique has been proposed as a novel classifier.Our proposed DBH work has been compared with the existing SVM techniques and proved to be outperforming than all the other techniques. The proposed DBH work has been compared with the existing SVM & PLNN work on the scale TP, TN, FP, FN, FRR, FAR, GAR, CM, Kappa coefficient, Sensitivity, Specificity and Accuracy.

## 9. Future Enhancement

The proposed novel Dissimilarity based Hybrid (DbH) classifier algorithm uses the extracted feature from the syllable count and weight extraction algorithm along with a novel score exploration, as a future work, it can extract additional intensive features from the words to train the classifier model.

Further the proposed model can be implemented to extract the online reviews from the other social sites(Reviews about the Products purchased in Online Shipping Sites and etc.) and tested for accuracy.

A recommender system can be designed based on the classification of the buyer's comments and can be implemented in online shopping sites like (snap deal and etc.)

## References

[1] Hua, Wen, et al. "Understand Short Texts by Harvesting and Analyzing Semantic Knowledge." IEEE transactions on Knowledge and data Engineering 29.3 (2017): 499-512.

[2] Shirakawa, Masumi, et al. "Wikipedia-based semantic similarity measurements for noisy short texts using extended naive bayes." IEEE Transactions on Emerging Topics in Computing 3.2 (2015): 205-219

[3] Xu, Bei, and Hai Zhuge. "An angle-based interest model for text recommendation." Future Generation Computer Systems 64 (2016): 211-226.

[4] De Boom, Cedric, et al. "Representation learning for very short texts using weighted word embedding aggregation." Pattern Recognition Letters 80 (2016): 150-156.

[5] Farman Ali, Kyung-Sup Kwak, Yong-Gi Kim, "Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification", Applied Soft Computing 47 (2016).

[6] Nikolaos D. Doulamis,, Anastasios D. Doulamis, Panagiotis Kokkinos, and Emmanouel (Manos) Varvarigos, "Event Detection in Twitter Microblogging", IEEE TRANSACTIONS ON CYBERNETICS 2015.

[7] VijayaMaheswari M , Dr.S.ManjuPriya "A Textual Pattern Encoding Method for Privacy Preservation in Social Streams" in Journal of Advanced Research in Dynamical & Control Systems (JARDCS), Issue 5, July 2017,Pages 129 – 134.

[8] Vijay B. Raut,D.D. Londhe,"Opinion Mining and Summarization of Hotel Reviews", 2014 International Conference on Computational Intelligence and Communication Networks (CICN).

[9] K Indhuja, Raj P C Reghu, "Fuzzy logic based sentiment analysis of product review documents" , 2014 First International Conference on Computational Systems and Communications (ICCSC).

[10] He, Jiangning, Hongyan Liu, and Hui Xiong. "SocoTraveler: Travel-package recommendations leveraging social influence of different relationship types." Information & Management 53.8 (2016): 934-950.

[11] Xu, Jiaming, et al. "Self-Taught convolutional neural networks for short text clustering." Neural Networks 88 (2017): 22-31.

[12] Yu, Zheng, et al. "Understanding short texts through semantic enrichment and hashing." IEEE Transactions on Knowledge and Data Engineering 28.2 (2016): 566-579.

[13] VijayaMaheswari M , Dr.T.Christopher "A Comparative study on various approaches for Event Detection in Social Streams" in International Journal of Engineering Research & Technology (IJERT), Volume 4, Issue 4, April 2015, Pages 1106-1109.

[14] Rao, Yanghui, et al. "Social emotion classification of short text via topic-level maximum entropy model." Information & Management 53.8 (2016): 978-986.

[15] John Ranelluccia, Eric G. Poitrasb, François Bouchetc, d, Susanne P. Lajoiee, Nathan Halle, "Emotions, Technology and Social Media", A volume in Emotions and Technology 2016.

[16] F.A. Pozzia, E. Fersinib, E. Messinab, B. Liuc,"Sentiment Analysis in Social Networks", 2017.

[17] Andreas Weiler, Michael Grossniklaus, Marc H. Scholl, "An evaluation of the run-time and task-based performance of event detection techniques for Twitter" , Information Systems (2016).

[18] Nikos Tsirakis, Vasilis Poulopoulos, Panagiotis Tsantilas, Iraklis Varlamis,"Large scale opinion mining for social, news and blog data", The Journal of Systems and Software 000 (2016).

[19] Jemal Abawajy, Mohd Izuan Hafez Ninggal and Tutut Herawan, "Privacy Preserving Social Network Data Publication", IEEE TRANSACTIONS 08 March 2016.

[20] Prashant Jawade,Poonam Joshi, "Securing Anonymous and Confidential Database through Privacy Preserving Updates", International Journal of Applied Information Systems (IJAIS) 2016.

[21] Xiang Sun, Yan Wu, Lu Liu, John Panneerselvam, "Efficient Event Detection in Social Media Data Streams ",IEEE International Conference on Computer and Information Technology(2015).

[22] VijayaMaheswari M , Dr.T.Christopher "A Review on Cluster Based Approach in Data Mining" in


Licensed Under Creative Commons Attribution CC BY

International Journal of Engineering Research & Technology (IJERT), Volume 4, Issue 3, March 2015, Pages 595-598.

[23] Kumar Ravi, Vadlamani Ravi,"A Survey on Opinion mining and Sentiment Analysis : Tasks,Approaches and Applications",Elsevier, Knowledge- Based Systems. (2015).

[24] Wei Wei ,Gao Cong, Chunyan Miao, Feida Zhu and Guohui Li, "Learning to Find Topic Experts in Twitter via Different Relations", IEE Transactions on Knowledge & Data Engineering (2016).

[25] Farzindan Atefeh, Wael Kherich, "A Survey of Techniques for Event Detection in Twitter",Computational Intalligence (2013).

[26] Suvarna D.Tembhurikar, Nitin N.Patil, "Topic Detection Using BNgram method and Sentiment Analysis on Twitter DataSet", IEEE Transactions on Cybernetics (2015).


Licensed Under Creative Commons Attribution CC BY