iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** II **Month of publication:** February 2020

DOI: http://doi.org/10.22214/ijraset.2020.2052

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 8 Issue II Feb 2020- Available at www.ijraset.com

Sentiment Analysis using Ensemble Classifier on Real Time Data Set

Aman Goenka[1], Gayatri Kuyte[2], Prof. Shubhangi Suryavanshi[3]
[1, 2, 3]*Department of Computer Engineering, G.H. Raisoni Institute of Engineering and Technology, Wagholi, Pune 412207*

Abstract: ***Social media has been growing day by day, monitoring and analysis of social media data play an important role in knowing people's behavior. We are performing analysis on Twitter Tweets using Ensemble Classifier which determines the opinion of the people regarding government schemes that are announced by the Central Government.***
This is based on social media twitter datasets of particular schemes and its polarity of sentiments. The popularity of the Internet has been rapidly increasing. Sentiment analysis and opinion mining is the field of study that analyzes people's sentiments, opinions, attitudes, and emotions from written language. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, microblogs, Twitter, and social networks. It is difficult to analyze or summarize the user-generated content. Most of the users write their opinions, thoughts on blogs, social media sites, E-commerce sites, etc. These contents are very important for Organizations, Government and Research and Development department to make decisions. For this Sentiment analysis and opinion mining research is a hot research area which comes under Data Science and Machine learning. We plot and calculate numbers of positive, negative and neutral tweets from each event.
Keywords: Ensemble classifier, Sentiment Analysis, Opinion Mining, Data Science, Machine learning

Article Info

In this article, we show that classifier ensembles formed by diversified components are promising for tweet sentiment analysis and provides more accurate results than traditional classifiers. Detection and handling of concept drift which produces accurate results in less time and reduces model sensitivity to noise.

.

I. INTRODUCTION

Social media has become an emerging phenomenon due to the huge and rapid advances in information technology. People are using social media daily to communicate their opinions with each other about a wide variety of subjects, products, and services, which has made it a rich resource for text mining and sentiment analysis. Social media communications include Facebook, Twitter, and many others. Twitter is one of the most widely used social media sites. This is helpful for evaluating government performance monitoring from People's perspective instead of making People's surveys which are expensive and time-consuming. Sentiment analysis has been first introduced by Liu, B. It is also known as Opinion Mining and subjectivity analysis is the process to determine the attitude or polarity of opinions given by humans to a particular scheme. We consider sentiment analysis a classification problem. Just like in large documents, sentiments of tweets can be expressed in different ways and classified based on the sentiment, i.e., if there are sentiments in the messages, then it is considered to be polar that is either positive or negative otherwise, it is considered neutral. Sentiment analysis can be applied to any textual form of opinions such as blogs, reviews, and microblogs. Microblogs are those small text messages such as tweets, a short message that cannot exceed 160 characters. These microblogs are easier than other forms of opinions for sentiment. Sentiment analysis can be done on a document level or a sentence level. In the first case, the whole document is evaluated to determine the opinion polarity, where, the features describing the product/service should be extracted first. Whereas, in the second one, the document is divided into sentences each one is evaluated separately to determine the opinion polarity. Many researchers have focused on using traditional classifiers, like Naive Bayes, Maximum Entropy, and Support Vector Machines to solve such problems. Here, we show that the use of a combination of multiple base classifiers combined with scores obtained from lexicons can improve the accuracy of tweet sentiment classification. Our experiments on a variety of public tweet sentiment datasets show that classifier ensembles formed by Multinomial Naive Bayes, SVM, Random Forest, and Logistic Regression can improve classification accuracy.

Some challenges can be faced while performing analysis on tweets:

1) Neutral tweets are way more common than positive and negative ones;
2) There are linguistic representational challenges;
3) Tweets are very short and often show limited sentiment cues.

Data streams are highly prone to the phenomena of concept drift, in which the data distribution changes over time. To maintain the performance level of these models, models should adapt to handle the existence of a drift. In this work, we present the Incremental Knowledge Concept Drift (IKCD) algorithm, an adaptive unsupervised learning algorithm for recommendation systems in the news data stream.

Experimental results illustrate an enhanced performance concerning

a) Reducing model sensitivity to noise,
b) Reducing model rebuilding frequency up to 50% in case of re-occurring drift,
c) Increasing accuracy of the model by about 10% with respect the accuracy of confidence distribution batch detection algorithm.

A. Our main Contributions can be Summarized as Follows

1) We show that classifier ensembles formed by diversified components are promising for tweet sentiment analysis and provides more accurate results than traditional classifiers;
2) We compare bag-of-words and feature hashing based on strategies for the representation of tweets and show their advantages and drawbacks;
3) Classifier ensembles obtained from the combination of lexicons, bag-of-words, emoticons, and feature hashing are studied and discussed.
4) Detection and handling of concept drift which produces accurate results in less time and reduces model sensitivity to noise.
5) Incremental learning continuously uses input data to further train the model and consequently extend the existing model's knowledge.

II. RELATED WORK

Several studies on the use of stand-alone classifiers for tweet sentiment analysis are available in the literature. Some of them propose the use of emoticons and hashtags for building the training set, as Go et al. and Davidov, who identified tweet polarity by using emoticons as class labels.

Others use the characteristics of the social network as networked data, like in Hu et al. According to the authors, emotional contagion theories are materialized based on a mathematical optimization formulation for the supervised learning process. Approaches that integrate opinion mining lexicon-based techniques and learning-based techniques have been studied. For example, Agarwal, Zhang, and Saif used lexicons, part-of-speech, and writing style as linguistic resources. In a similar context, Saif introduced an approach to add semantics to the sentiment analysis training set as an additional feature. For each extracted entity (e.g., iPhone), they added its respective semantic concept (like \Apple's product") as an additional feature and measured the correlation of the representative concept as negative/positive sentiments.

Classifier ensembles for tweet sentiment analysis have been underexplored in the literature few used logistic regression classifiers learned from 4-gram hashed byte as features.

They did not attempt any linguistic processing, not even word tokenization. For each of the datasets, they experimented with ensembles of different sizes, composed of different models, and obtained from different training sets, however with the same learning algorithm.

Their results show that the ensembles lead to more accurate classifiers. Rodriguez and Clark proposed the use of classifier ensembles at expression-level, which is related to Contextual Polarity Disambiguation. In this perspective, the sentiment label is applied to a special phrase or word within the tweet and does not necessarily match the sentiment of the entire tweet. Finally, a promising ensemble framework was recently proposed by Hassan, who deals with class imbalance, sparsity, and representational issues.

The authors propose enriching the corpus by using multiple additional datasets also related to sentiment classification. The authors use a combination of unigrams and bigrams of simple words, part-of-speech, and semantic features derived from WordNet and SentiWordNet. Also, they employed summarizing techniques, like Legomena and Named Entity Recognition.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 8 Issue II Feb 2020- Available at www.ijraset.com

III. SENTIMENT ANALYSIS PROCESS

Sentiment Analysis is classified into two main approaches i.e. Supervised Learning Approach and Unsupervised Approach. In Sentiment Analysis Process Following Steps are necessary.

1) *Collection of Peoples tweets:* Tweets are necessary for doing the Sentiment Analysis Task. For the Collection of tweets, there are different techniques which are used in this survey. Schemes of tweets are collected from Twitter websites. The tweets can be a structured, semi-structured and unstructured type. Sentiment Analysis research, There is an open-source framework where the researchers can get their data for the research purpose by installing the required packages and authentication process of social websites, to crawl the reviews from that site is an easy task. Once we have our text data with us then we can use that data for Pre-processing purposes.
2) *Pre-Processing:* Data pre-processing is done to remove incomplete noisy and inconsistent data. Data must be preprocessed before using it in the feature selection task. In the pre-processing following are some tasks: • Removing URLs, Special characters, Numbers, Punctuations, etc. • Removing Stopwords • Removal of Retweets • Stemming • Tokenization
3) *Feature Selection:* Feature selection from pre-processed text is a difficult task in sentiment analysis. The main goal of the feature selection is to decrease the dimensionality of the feature space and thus computational cost. Feature selection will reduce the overfitting of the learning scheme to the training data. Different machine learning algorithms were analyzed on a Scheme dataset with different feature selection techniques features.
4) *Sentiment Word Identification:* Sentiment word identification is a fundamental work in numerous applications of sentiment analysis and opinion mining, such as tweets mining, opinion holder finding, and tweet classification. Sentiment words can be classified into positive, negative and neutral words.
5) *Sentiment Polarity Identification:* The basic task in Sentiment Analysis is classifying the polarity of a given tweets feature. The polarity is in three categories i.e. Positive, Negative and Neutral. Polarity identification is done by using different lexicons e.g. Bing Lui sentiment lexicon, SentiWordNet, etc. which help to calculate sentiment score, sentiment strength, etc.
6) *Sentiment Classification:* Sentiment classification of government schemes tweets dataset and opinion of schemes dataset is done using supervised machine learning approaches like naïve Bayes, SVM, Maximum Entropy, etc. Accuracy depends on which dataset is used for which classification methods. In the case of Supervised machine learning approaches Training dataset is used to train the classification model which then helps to classify the test data.
7) *Analysis of Reviews:* Finally Analysis of the result is important to decide on individual and governmental schemes. In the case of governmental schemes announced by the central government, gets many tweets and if more tweets are positive then people may like that particular schemes. Feedback of particular schemes helps in taking the appropriate discussion to the public and decide for proper implementation of government schemes.

IV. CLASSIFIER ENSEMBLE FOR TWEET SENTIMENT ANALYSIS

Ensemble methods train multiple learners to solve the same problem. In contrast to classic learning approaches, which builds one learner from the training data, ensemble methods build a set of learners and combine them. Dietterich lists three reasons for using an ensemble-based system:

A. Statistical

Assuming that we have several different classifiers, all of them having good accuracy in the training set. If a single classifier is chosen from the ones available, it may not yield the best generalization performance in unseen data. By combining the outputs of a set of classifiers, the risk of selecting an inadequate one is lower.

B. Computational

Many learning algorithms work by carrying out a local search that may get stuck in local optima which may be far from global optima.

For example, decision tree algorithms employ a greedy splitting rule and neural networks algorithms employ gradient descent to minimize an error function over the training set. An ensemble constructed by running the local search from many different starting points may provide a better approximation than any of the individual classifiers;

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 8 Issue II Feb 2020- Available at www.ijraset.com

C. Representational

If the chosen model cannot properly represent the sought decision boundary, classifier ensembles with diversified models can represent the decision boundary. Certain problems are too difficult for a given classifier to solve. Sometimes, the decision boundary that separates data from different classes may be too complex and an appropriate combination of classifiers can make it possible to cope with this issue.

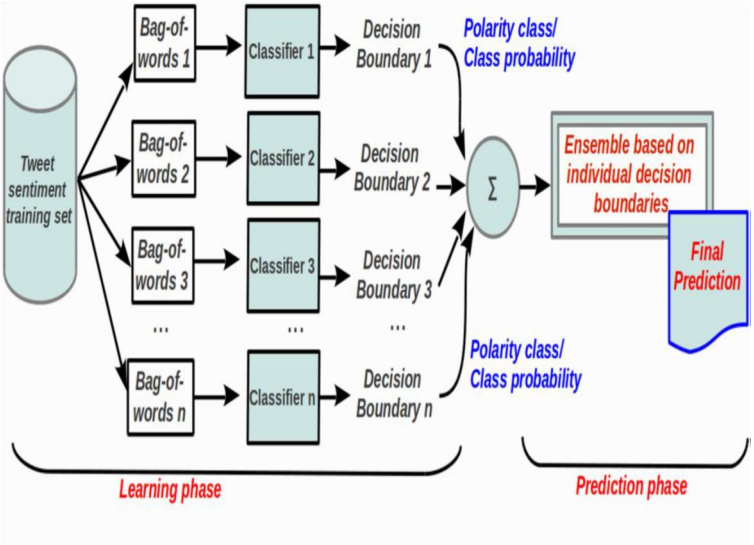


Figure 1: Classifier ensemble for tweet sentiment analysis: refers to the combination rule (e.g., majority vote and average of class probabilities) for the base classifiers.

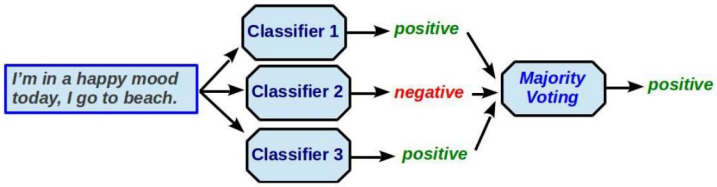


Figure 2: An example of Majority Voting as the combination rule. In this case, the majority of the classifiers agree that the class is positive.

V. DETECTION AND HANDLING OF DRIFT

Recommendation Systems are widely used to help users find items from repositories according to his/her interests. This is driven by the patterns of the users in online shopping, news and video content consumption, search, as well as consumption of content and in many different domains. Fundamentally Recommendation Systems develop a model of the user's interests by observing his/her patterns over a while. Modeling approaches could be widely categorized into two categories, namely (i) Collaborative-based and (ii) Content-based recommendation. Collaborative based Recommendation Systems assume that users with the same "taste" would value items similarly. In turn, the Collaborative-based approaches cluster users with similar tastes into a group and generate recommendations from within this group. On the other hand, Content-based Recommendation Systems assume that a user will react to similar content items in the same manner. In turn, the Content-based Recommendation Systems build user models utilizing features of items that the user favorites in the past. One of the major issues faced by the Recommendation Systems is that the user's interest, and/or the item features themselves may change over time especially for data streams that online users interact with. This problem is referred to in the literature as "concept drift". Concept drift causes the output recommendations to skew from the current user interest over time and become irrelevant. In turn, the existence of concept drift requires Recommendation Systems to adapt the user modeling process to the changing user interest of item features. News, a major application of Recommendation Systems, is one of the domains that are highly prone to concept drift. This is because news is continuously flowing, and usually short-lived. On the other hand, when the user's interest in certain topics of news change over time, models generated from user history becomes obsolete and no longer generates meaningful recommendations.

A. *Types Of Concept Drift*

The literature classifies types of concept drift based on either the time or the predictive views. There are four types of concept drift; namely (i) sudden, (ii) gradual, (iii) incremental, and (iv) reoccurring drift.

1) *Sudden Drift:* This takes place when a concept C1 is abruptly replaced by another concept. For example, a researcher working on big data storage and retrieval algorithms is assigned a task to prepare a survey on security issues in big data. This abrupt change in the researcher's interest from storage and retrieval algorithms to security issues represents a sudden drift.
2) *Gradual Drift:* This takes place when concept C2 starts growing while C1 is still of interest. However, by the time, C2 continues growing until it becomes the dominant concept while C1 starts decaying until it disappears. For example, a researcher who is interested in storage and retrieval algorithms in big data wants to extend his knowledge to security issues in big data. Over time, he/she becomes interested in security issues in big data and stops reading about data storage and retrieval in big data.
3) *Incremental Drift:* This can be identified only over an extended period of time because small changes accumulate over time. For example, if members of the big data community start discussing security issues of big data rather than storage and retrieval algorithms. By the time, the security issues attract their major attention and the storage and retrieval algorithms are out of their interest. It is worth pointing out that incremental drift demonstrates only one active concept at any time, while on the other hand gradual drift occurs when two concepts are concurrently active.
4) *Re-occurring Drift:* Here it refers to the case when a previously concept reappeared after some time. For example, a researcher who is interested in big data storage and retrieval algorithms, changes his/her attention towards security issues in big data.

However, the researcher discovered that security issues are out of his/her interest, and in turn, he/she returns working on storage and retrieval algorithms.

To maintain the performance level of these models, models should be able to handle the existence of a drift. In this, we present the Incremental Knowledge Concept Drift algorithm, an adaptive unsupervised learning algorithm for Sentiment Analysis in the twitter data stream.

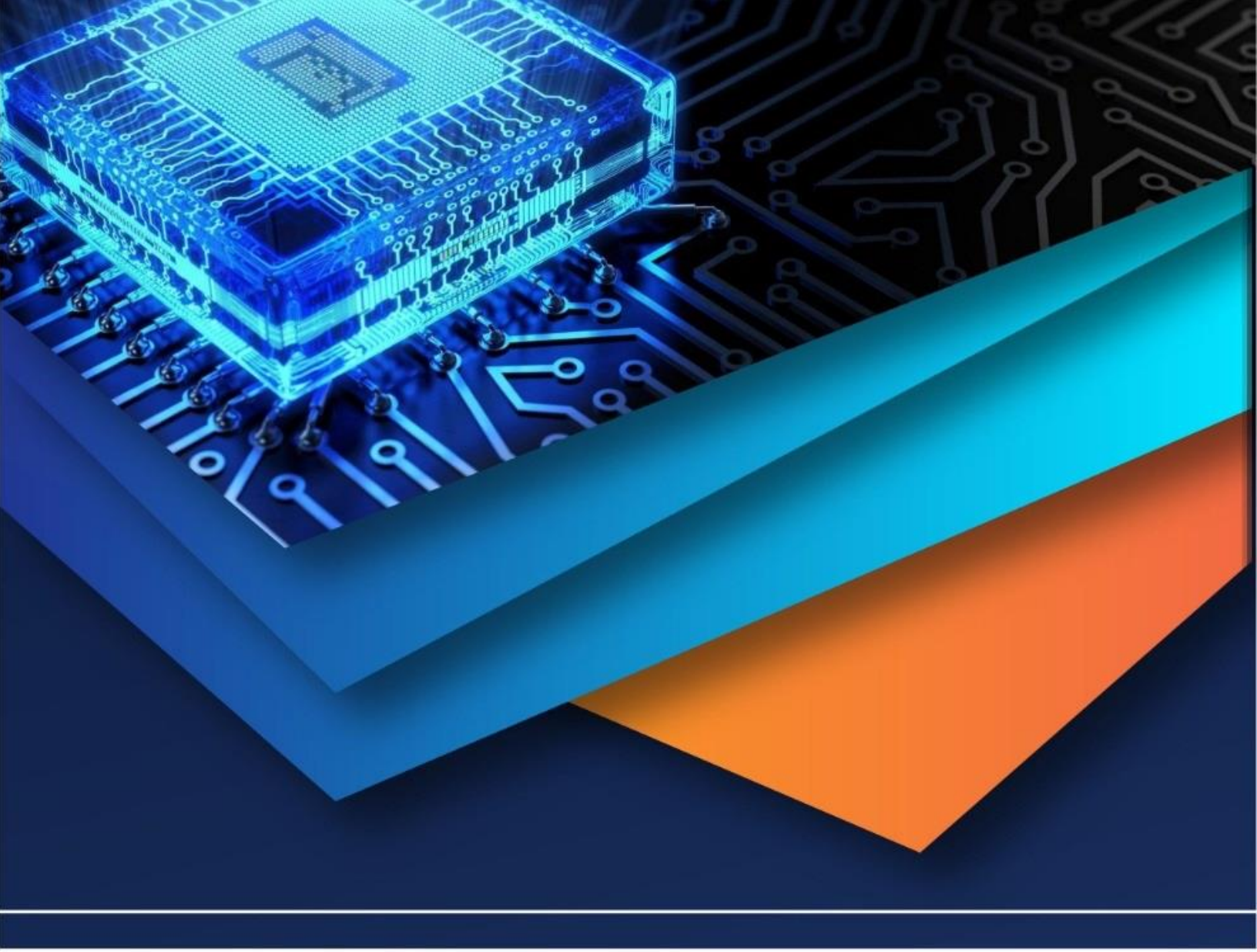
VI. CONCLUSION

In this experimental work, we used a Twitter API which is an open-source API. Tweets from twitter are collected. The use of classifier ensembles for tweet sentiment analysis has been underexplored in the literature. We have demonstrated that classifier ensembles formed by diversified components especially if these come from different information sources, such as textual data, emoticons, and lexicons can provide state-of-the-art results for this particular domain. We also compared promising strategies for the representation of tweets i.e., bag-of-words and feature hashing and showed their advantages and drawbacks.

Feature hashing has shown to be a good choice in the scenario of tweet sentiment analysis where computational effort is of paramount importance. However, when the focus is on the accuracy, the best choice is bag-of-words. Although our results have been obtained for data from Twitter, one of the most popular social media platforms, we believe that our study is also relevant for other social media analysis. Future work about government schemes and sentiment analysis is to find out aspects and their polarity of the schemes which helps for the implantation of government schemes effectively to take the decision to upcoming scheme regarding public satisfaction.

REFERENCES

[1] Improving Twitter Aspect-Based Sentiment Analysis Using Hybrid Approach Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia 2017
[2] Sentiment Analysis of Twitter Data Using Machine Learning Techniques and Scikit-learn .
[3] International Conference On Recent Advances In Computer Science, Engineering And Technology, Sentiment Analysis of Indian Government Schemes using Twitter Datasets, Jan 2018 .
[4] Tweet Sentiment Analysis with Classifier Ensembles Nadia F. F. da Silvaa, Eduardo R. Hruschkaa, Estevam R. Hruschka Jr.b . Institute of Mathematics and Computer Sciences Federal University of Sao Carlos (UFSCAR) S~ao Carlos, SP, Brazil .
[5] International Journal of Computer Science & Information Technology (IJCSIT) Vol 11, No 1, February 2019 DOI: 10.5121/ijcsit.2019.11107 87 Detection and Handling of Different types of Concept Drifts.
[6] Rajkumar S. Jagdale, Vishal S. Shirsat, Sachin N. Deshmukh," Sentiment Analysis of Events from Twitter Using Open Source Tool International Journal of Computer Science and Mobile Computing ISSN 2320–088X IMPACT FACTOR: 5.258 IJCSMC, Vol. 5, Issue. 4, April 2018, pg.475 – 485
[7] Xing Fang and Justin Zhan."Sentiment analysis using product review data". Journal of Big Data 2017. DOI: 10.1186/s40537-015-0015-2
[8] B. J. Jansen, M. Zhang, K. Sobel, A. Chowdury, Twitter power: Tweets electronic word of mouth, J. Am. Soc. Inf. Sci. Technol. 60 (11) (2017)
[9] J.-M. Xu, K.-S. Jun, X. Zhu, A. Bellmore, Learning from bullying traces in social media., in: HLT-NAACL, The Association for Computational Linguistics, 2018



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY