iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** I **Month of publication:** January 2018

DOI: http://doi.org/10.22214/ijraset.2018.1253

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 :6.887
Volume 6 Issue I, January 2018- Available at www.ijraset.com

Development of a Technique for Gender Recognition on Twitter Users

Jatinder Kaur[1], Dr. Kiran Jyoti[2]
[1] *M tech Student, Department of CSE, GNDEC Ludhiana*
[2]*Assistant Professor, Department of IT, GNDEC Ludhiana*

Abstract: Content is as utmost common Internet media sort. Instances of incorporate prominent social networking sites, for example, Twitter, Facebook, Craigslist and so on. Other web applications for example, email, chat rooms, blog and so forth are likewise generally message based. An inquiry we address in this paper manages content based Internet crime scene investigation is the accompanying: given a short content record, would we be able to distinguish the author of text as male or female? The analysis of author classification is influenced by late occasions where individuals try to counterfeit their sexual orientation on the Internet. In this paper, we propose feature selection with hybrid classification technique. Data set is collected and data pre-processing is performed
to mine the data and feature selection is performed to reduce the dimensionality of dataset. Classification is done to improve the performance of individual classifiers. The proposed method is tested along various parameters like accuracy, precision, recall, f Measure.
Keywords: Social Organizing Applications, Feature Selection, Hybrid Classification, Data Pre-processing, Classification

I. INTRODUCTION

With the massive use of social networking sites, enormous routes of data sharing are available. Well-known micro-blogging platform, Twitter provided a very limited information about a user such as Name, Email, Location, URL except age or gender which is a true identity (relevant information) of user profile. All of these information are valuable in variety of fields like advertising, legal-inquiry for cyber-crime, personalization, marketing(online reviews),political challenges. Absence of relevant features such as gender or age have caused great interest in research regarding authorship of users on Twitter. To predict these latent features accurately can be valuable in variety of research and applications. [1].

Various Machine Learning Techniques have been used in existing research for gender recognition like face recognition, speech recognition, gesture or facial expression recognition, text classification[2].In general, we can predict the gender of posted tweets automatically based on the relevant user names but to detect the gender from unknown tweets may produce irrelevant results. The concentration of this exploration to classify gender from unknown tweets. Text Classification one of the predominant approach used in Data Mining to classify text. This paper will discuss various classification techniques used to accomplish research objectives.

A. Support Vector Machines

These are directed training models with related training computations which separate information used for gathering or backslide examination. Given a course of action of preparing cases, each and every set independently
having a point with both orders, After getting ready SVM initiates computation that helps to build a model that designates new instances to single class or the other, framing it a non-probabilistic combined linear classifier [3].

B. Logistic Regression

Logistic Regression is also a classification technique. This is utilized to foresee a binary result (1/0, Yes/No, True/False) given an arrangement of free factors. To speak to binary/absolute result, we utilize imaginary factors. You can likewise consider logistic regression as an exceptional instance of linear regression when the result variable is unmitigated, where we are utilizing log of chances as needy variable. In basic words, it predicts the likelihood of event of an occasion by fitting information to a log it work. Like all regression analysis, the strategic relapse is a prescient investigation. Calculated relapse is used to depict data and to clear up the association between one ward matched variable and no less than one apparent, ordinal, interval or extent level self-ruling components [3].

C. ADA Boost

Boosting comprises of straightly joining an arrangement of powerless classifiers to acquire a solid one. The AdaBoost calculation, presented in 1995 by Freund and Schapire. The calculation takes data as a preparation set (x1; y1);:::; (xm; ym) where every xi has a place with some area or instance space X, and each name yi is in some name set Y. AdaBoost iteratively takes the best basic classifier it could discover at each progression, and adds it to its last arrangement of classifiers while figuring its coefficient.

At every cycle of the calculation, it tries to locate the best classifier for the learning illustrations which have been slightest treated up until now. Clearly, picking the best basic classifier at each progression of AdaBoost is impossible by testing every one of the potential outcomes [3].

II. RELATED WORK

In the previous work, large volume of informal text generated at high speed from twitter have been experimented with traditional batch mining techniques as well as stream mining algorithm[4]. Various individual classification algorithms were not able to produce better

efficiency and accuracy. To overcome this problem ,related attributes of algorithms were tested and integrated to form a classifier which is not yet implemented.

Marco Vicente et al. [5] describes a way to deal with naturally recognize the gender orientation of Twitter clients. A number of features that capture phenomena specific of Twitter users is proposed and evaluated on a dataset of about 242K English language users. Distinctive administered and unsupervised methodologies are utilized to survey the execution of the proposed features involving Logistic Regression, Naive Bayes variations, Support Vector Machines, K-implies and Fuzzy c-Means bunching. An unsupervised

approach in view of Fuzzy c-Means ended up being exceptionally reasonable for this undertaking, restoring the right gender for around 96% of the clients.

Shilpy Singh et al. [6] describes a lot of programming languages & its compatibility problems with databases and few refers to build libraries for re-usability and find reliable Twitter4J libraries for better utilization. In case of data acquisition Twitter APIs can be integrated with any applications and in any format. Twitter API is a cross-platform tool, platform independent and compatible with the latest versions of Java Runtime Environment.Twitter4J is very easy to use, first and foremost to do copy the JAR file to preferred class

path and execute[6]. We explore the techniques for usage of twitter4J libraries for data acquisition and data analytics. This research will assist data scientist, data quality analyst and business users.

It has been acknowledged from existing research that there are certain improvements required to achieve appropriate outcome. Many classifiers were performing well to achieve higher accuracy with smaller training sets but better accuracy yields to larger training sets[5].In Gender Recognition, accuracy is calculated on the basis of classification methods like Naive Bayes, SVM, Logistic Regression,NN,C4.5 Decision tree etc [6].It is not sufficient to classify instances accurately on the basis of training data set and various classification techniques required for better accurate result.

III.PROPOSED TECHNIQUE

In our research, the proposed technique combines the classification methods which would be discussed in detail. The brief description of steps is as per the following:

A. Data Collection

Data collection is done from Twitter by collecting tweets. Input data will be raw content from tweets on social issues in India. This informational collection was utilized to prepare a Crowd Flower AI gender orientation indicator. Patrons were asked to take a look at Twitter account and predict whether the client was a male, a female, or a united (non-individual).The training dataset consists of rows defined by client name, an irregular tweet, profile and picture, area, and connection and sidebar shading[6]. The training data sets are obtained from the following web link https://data.world/crowdflower/gender-classifier-data.

B. Data Pre-Processing

Data pre-processing is a mining system that includes changing crude data into a justifiable configuration. Pre-processing incorporates 3 stages:

1) Tokenization and parsing of words: In this stage, each tweet sentence parts into expressions of any normal handling dialect.

2) Removing stop words: The words that contain less information are called stop words. Thusly, at the period of pre-taking care of, shut this stop word that every word refers to these can erased from dataset and execution will increase.
3) Stemming: It is termed as a approach to lessen the categorical words to one of their kind word stem. For example, "walked", "walking", "walk" as in perspective of the root word "walk". Snowball stemmer is used to reduce the decided word to motivation.

C. *Features Selection*

Dimensionality of dataset can be reduced with the use of Feature selection approach. Feature Selection is the technique which can be applied to expel expressions in the preparation archives that are measurably irrelevant with the class marks. It will diminish arrangement of expressions to be utilized as a part of characterization, along these lines enhancing both productivity and precision[7].

In order to select the best features Symmetric uncertainty is used that will figure the wellness of highlights for include choice by ascertaining amongst highlight and the objective class. The element which has high estimation of SU gets high significance. An element ought to be exceptionally corresponded to the class and very little related to some other component of the class. For this we have utilized data hypothesis in light of entropy which is a measure of vulnerability of an arbitrary variable. This can be characterized by accompanying condition 1 as

$$H(X) = - \Sigma P (xi) \log2 (P (xi) \quad (1)$$

Furthermore, the entropy of X in the wake of watching estimations of another variable Y is characterized in condition 2 as

$$H(X/Y) = - \Sigma P (yj) \Sigma P (xi/yj) \log2 (P (xi/yj)) \quad (2)$$

Here, P (xi) is the earlier probabilities for all estimations of X, and P (xi/yj) is the back probabilities of X when estimations of Y are given. The sum by which the entropy of X diminishes mirrors extra data about X gave by Y is called data increase given the condition 3 as

$$IG(X/Y) = H(X)-H(X/Y) \quad (3)$$

We can conclude that component Y is more associated to include X than to highlight Z, if IG(X/Y) > IG (Z/Y). Here is one more measure balanced vulnerability that demonstrates relationship between's highlights characterized by condition 4 as

$$SU (X, Y) = 2 [IG(X/Y) / H(X) + H(Y)] \quad (4)$$

SU repays data increase's predisposition toward highlights and standardizes its incentive to scope of [0,1] with 1 demonstrating that information of possibly one totally detects the estimation of other and 0 demonstrates that X and Y are autonomous. It reflects combination of highlights equally[8].

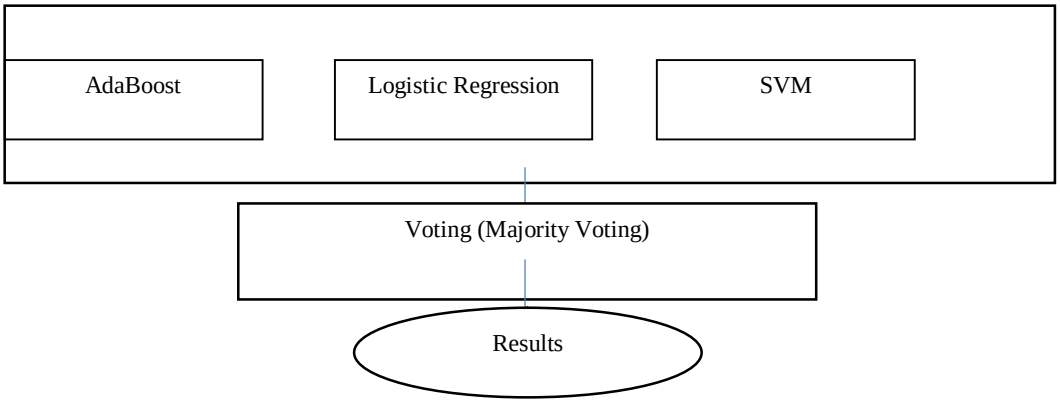
D. Classification

With a specific end goal to enhance the execution of individual classifiers utilized as a part of the paper is to the utilization of meta classifiers. While the benefit of utilizing Hybrid, strategies is the change of the execution, the impediment is about the time it takes to complete the preparation stage. Be that as it may, the principle concern was to construct a model which has a superior execution contrasted with the individual classifiers. This model is novel on the grounds that it utilizes Hybrid technique (Voting), as well as it applies a meta classifier (Bagging) as one of its classifier part. Additionally, parameter advancement approach was utilized on its individual classifier (SVM). Every segment in our model adapts a few sections of the characterization issue and we join these theories to choose the likelihood level [9].

Procedure1. Procedure for hybrid classification algorithm

```
DATA = [antoloji, beyazperde, hepsiburada]
LEARNER = [Naive Bayes, SVM, Bagging, Our Technique]
REPEAT 10 TIMES
FOR EACH data in DATA
TRAIN = random 90% of data
TEST = data – TRAIN
PREDICTOR = Train LEARNER with TRAIN
[accuracy] = PREDICTOR on TEST
END
END
```

FIGURE 1: Hybrid Learning Model

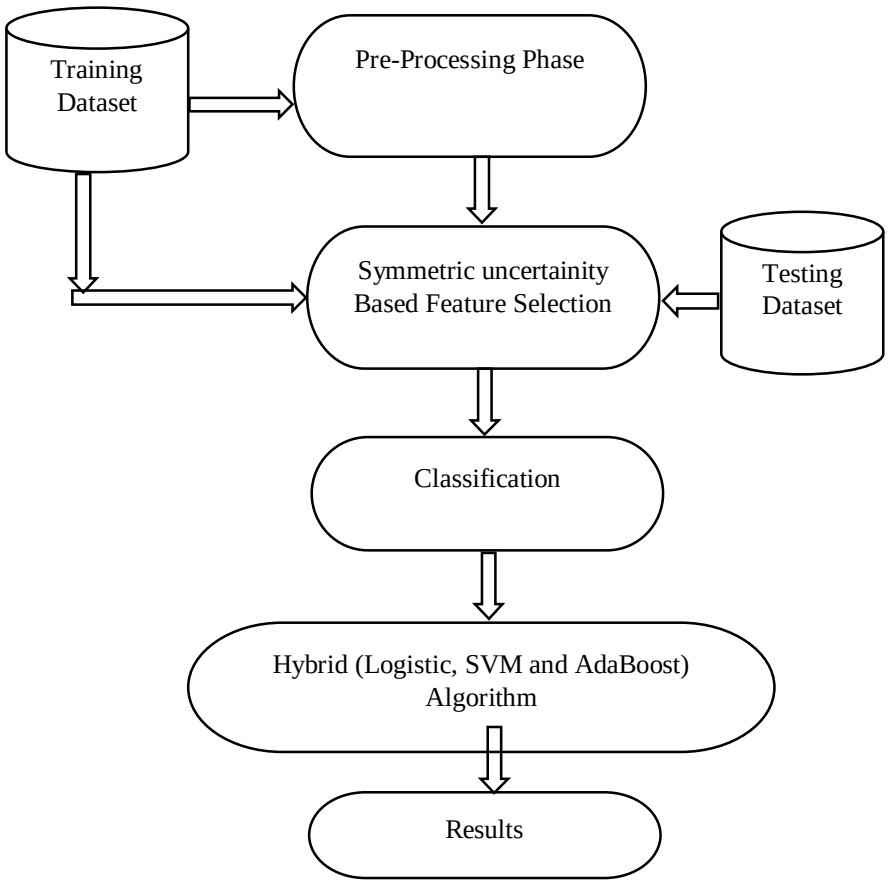


E. Building Proposed Model

The classifier show worked by utilizing proposed half and half strategy is assessed and tried utilizing k-overlap Cross Validation approach (k-overlay CV). In this approach, the preparation set is divided into k little sets. The methodology followed in k-overlap CV approach is as per the following:

A proposed model is prepared utilizing k-1 of the folding as preparing information; the subsequent model is approved on the rest of the piece of the information (i.e., it is utilized as a test set to figure an execution measurement, for example, precision).The execution measurement assessed by k-overlay cross-approval is then the normal of the qualities figured on the up and up.

FIGURE 2: FLOWCHART OF PROPOSED TECHNIQUE



IV. RESULTS AND DISCUSSIONS

The Proposed Enhanced feature selection with hybrid classification algorithm is tested on author gender identification from tweets for predicting the gender. The performance is analyzed based on below mentioned parameters.

The parameters for the assessment of sentiment analysis incorporate different terms like True positives, true negatives, false negatives and false positives which are utilized to contrast the class names allotted with archives. True positive terms are genuinely named positive terms while false positive are not marked by the classifier as positive class but rather ought to have been. True negative terms are effectively named as negative class. Classes that are not named by the classifier as having a place with negative class however ought to have be classified are known to be False negative terms. Confusion Matrix contains these terms that are utilized for assessment.

TABLE 1: CONTINGENCY TABLE

		Correct Labels	
		Positive	Negative
Classified Labels	Positive	True positive	False positive
	Negative	False negative	True negative

TABLE 2: CONFUSION MATRIX OF AUTHOR GENDER IDENTIFICATION FROM TWEETS

	SVM	Logistic	Adaboost	FS with hybrid classification
Correctly Classified ($n_{s\to s}$+ $n_{h\to h}$)	369	367	368	568
Incorrectly Classified ($n_{h\to s}$ + $n_{s\to h}$)	325	327	326	126

Where ($n_{s\to s}$) represents True Positive rate, ($n_{h\to s}$) represents False Positive rate, ($n_{s\to h}$) represents False Negative and ($n_{h\to h}$) represents True Negative

A. Accuracy

Accuracy is the basic measure for arrangement execution. It can be measured as accurately ordered occasions to the aggregate number of examples, while blunder rate utilizes mistakenly grouped cases rather than effectively arranged cases.

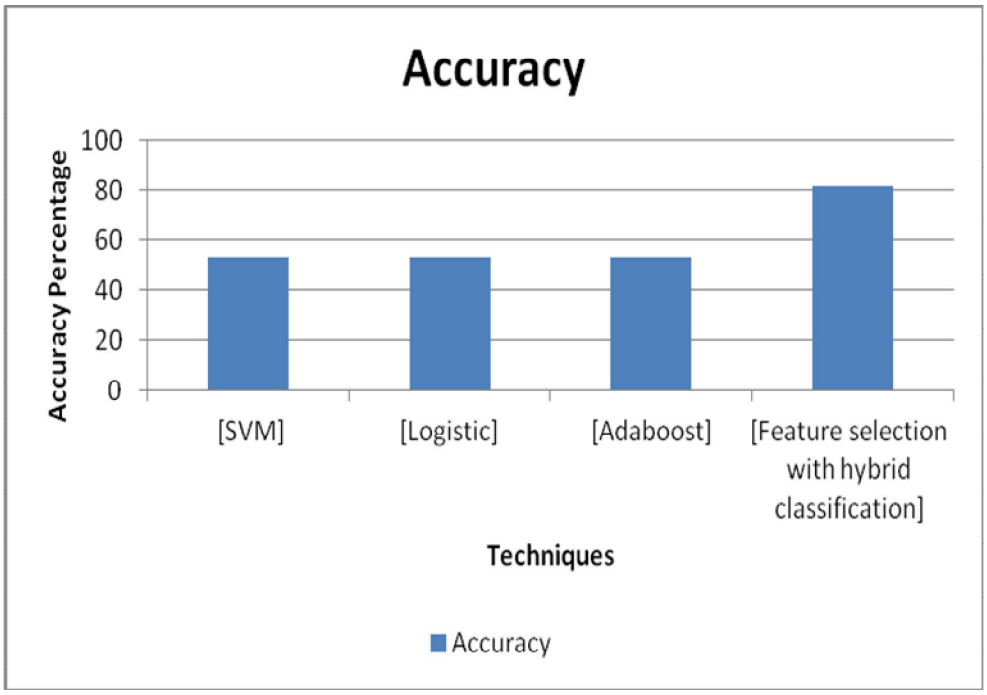
$$\text{Accuracy} = \frac{\textbf{True Positive+True Negative}}{\textbf{True Positive+False Positive+True Negative+False Negative}} \quad (5)$$

TABLE 3: Representing the Accuracy of Proposed Method With Respect To Previous Methods

Algorithms	Accuracy
SVM	53.17
Logistic	52.8818
Adaboost	53.0259
Feature selection with hybrid classification	81.8444

As shown in above table the accuracy of the proposed technique called feature selection with hybrid classification is far better than the previous techniques like support vector machine, logistic, and adaboost.

FIGURE 3: Graph Representing Accuracy Variation



This is the graph representing accuracy of various techniques with respect to the accuracy percentage. It clearly depicts that proposed technique performs better than the previous methods.

B. Precision And Recall

Precision and recall are the two measurements that are broadly utilized to evaluate execution in content mining, and in content investigation field like data recovery. These parameters are utilized for calculating precision and culmination separately.

$$\text{Precision}=\frac{\text{True Positive}}{\text{True Positive}+\text{False Positive}} \quad (6)$$

$$\text{Recall}=\frac{\text{True Positive}}{\text{True Positive}+\text{False Negative}} \quad (7)$$

C. F measure

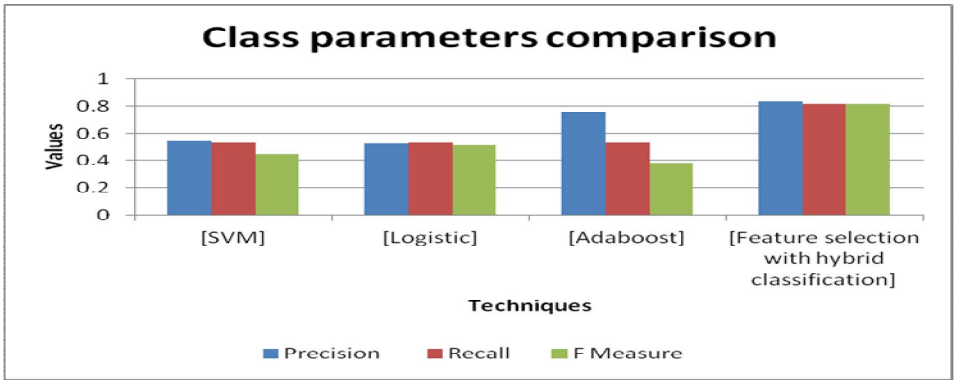
The harmonic mean of precision and recall is called F-Measure. The esteem ascertained utilizing F-measure is a harmony amongst precision and recall.

$$\text{F measure} = \frac{2*\text{recall}*\text{precision}}{\text{precision}+\text{recall}} \quad (8)$$

TABLE 4: Representing the Various Parameters of Proposed Method With Respect To Previous Methods

Algorithms	Precision	Recall	F Measure
SVM	0.544	0.532	0.449
Logistic	0.526	0.529	0.513
Adaboost	0.754	0.53	0.381
Feature selection with hybrid classification	0.834	0.818	0.818

FIGURE 4: Graph Representing Class Parameters Of Various Algorithms



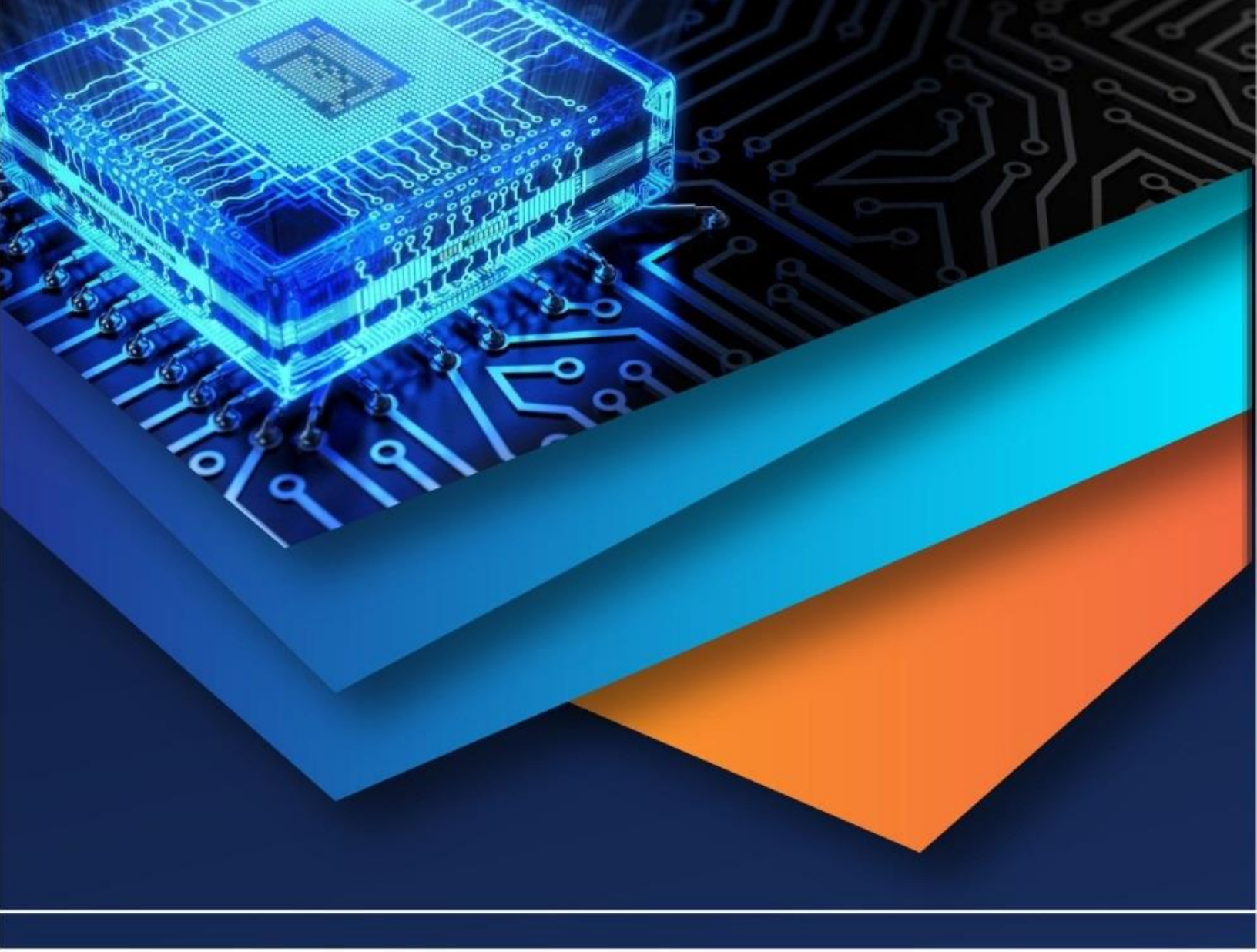
This graph represents various class parameters comparison like precision, recall, F Measure of different techniques along with the proposed technique.

V. CONCLUSION

We perceive that the issue of gender distinguishing proof from content is a transaction between psycho-etymology, bland written work styles of men and ladies [10] and so on. In this paper, many classification techniques are implemented to predict the accurate result. Classification algorithms were used to analyze the relationship between attributes which helped to build an accurate classifier for gender recognition. Purpose of this proposed technique to achieve higher accuracy as compare with earlier techniques. Hybrid Model is implemented on weka environment. Experimental results shows that feature selection with hybrid classification performs better than various methods like SVM, Logistic, Adaboost based on certain parameters like accuracy, precision, recall and fMeasure. So,this proposed technique achieve higher accuracy to predict gender from unknown tweets.

REFERENCES

[1] John D. Burger and John Henderson and George Kim and Guido Zarrella, “Discriminating Gender on Twitter”,Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), July 27 - 31, 2011, pp. 1301-1309.doi:10.1.1.226.279
[2] Sajid Ali Khan, Maqsood Ahmad, Muhammad Nazir and Naveed Riaz,”A Comparative Analysis of Gender Classification Techniques”, Middle-East Journal of Scientific Research 20 (1): 01-13, 2014.
[3] Na Cheng, R. Chandramouli, K.P. Subbalakshmi, “Author gender identification from text”,The International Journal of Digital Forensics & Incident Response,Volume 8 Issue 1 pp. 78-88. doi:10.1016/j.diin.2011.04.002
[4] Zachary Miller,Brian Dickinson”Gender Prediction on Twitter using Stream Algorithms with N- Gram”, International Journal of Intelligence Science pp. 143-148.doi :10.4236/ijis.2012.224019.
[5] Marco Pennacchiotti, Ana-Maria Popescu”A Machine Learning Approach to Twitter User Classification”,Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011, pp. 281-288. doi: 10.1109/FUZZ-IEEE.2015.7338102
[6] Shilpy Singh, Manjunath T N and Ashwini,”A Study on Twitter 4j Libraries for Data Acquisition from Tweets”, International Journal of Computer Applications (0975 – 8887) National Conference on “Recent Trends in Information Technology” (NCRTIT-2016), pp. 29-32.
[7] Zheng R, Li J, Chen H, Huang Z ”A framework for authorship identification of online messages: writing-style features and classification techniques”,Journal for the American Society for Information Science and Technology 57(3):378–393, 2006.doi: 10.1002/asi.20316
[8] Antonio Castro, Brian Lindauer : “Author Identification on Twitter”, (2012)
[9] Marco Paulo Fernandes Vicente “Detecting Portuguese and English Twitter users’ gender”,Department of Information Science and Technology, 2015.
[10] Malcolm Corney, Olivier de Vel,Alison Anderson,George Mohay ”Gender-Preferential Text Mining of E-mail Discourse”,Proceedings of the 18th Annual Computer Security Applications Conference (ACSAC), December 09 - 13, 2002.doi: 10.1109/CSAC.2002.1176299
[11] Valdemir Vapnik. (1979). Retrieved from SVM: ttp://www.svms.org/history.html
[12] Reuters corpora. (2000). Retrieved from http://trec.nist.gov/data/reuters/ reuters.html
[13] Sebastian Raschka : “Naive Bayes and Text Classification“, Article, 04 October 2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY