Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 4, April 2017, pg.444 – 458

Sentiment Analysis of Tweets using Machine Learning Approach

Ankita Gupta[1], Jyotika Pruthi[2], Neha Sahu[3]

[1] Ankita Gupta, Student, Dept. of Computer Science
THE NORTHCAP UNIVERSITY, Sector 23A Gurgaon, Haryana (122017), INDIA
Email: ank21gupta@gmail.com

[2,3]Jyotika Pruthi, Neha Sahu, Faculties, Dept. of Computer Science
THE NORTHCAP UNIVERSITY, Sector 23A Gurgaon, Haryana (122017), INDIA
Email: jyotikapruthi@ncuindia.edu, nehasahu@ncuindia.edu

ABSTRACT: Sentiment Analysis comes under study within Natural Language processing. It helps in finding the sentiment or opinion hidden within a text. This research focuses on finding sentiments for twitter data as it is more challenging due to its unstructured nature, limited size, use of slangs, misspells, abbreviations etc. Most of the researchers dealt with various machine learning approaches of sentiment analysis and compare their results[1][3][4][5][6][7][13][14][15][19][20][21][22][31] but using various machine learning approaches in combination have been underexplored in the literature. This research has found that various machine learning approaches in a hybrid manner gives better result as compared to using these approaches in isolation. Moreover as the tweets are very raw in nature, this research makes use of various preprocessing steps so that we get useful data for input in machine learning classifiers. This research basically focuses on two machine learning algorithms K-Nearest Neighbours (KNN) and Support Vector Machines (SVM) in a hybrid manner. The analytical observation is obtained in terms of classification accuracy and F-measure for each sentiment class and their average. The evaluation analysis shows that the proposed hybrid approach is better both in terms of accuracy and F-measure as compared to individual classifiers.

Keywords: Sentiment Analysis, Machine Learning, KNN, SVM.

1. INTRODUCTION

Due to the presence of enormous amount of data available on web, various organizations started taking interest in this as mining this information can be very valuable to them. This gives birth to an entirely different and broad field of study known as Sentiment Analysis. Various names have given to this field as opinion mining, opinion extraction etc. However there is slight difference in meaning between these various terms. Before automatic mining of sentiments traditional survey techniques

were highly biased as they were taken individually by users thus a need of an automatic system arose that can directly deal with hundreds of thousands of opinions hidden in users' posts in the form of reviews, blogs etc. Various applications of sentiment analysis are as in product reviews, movie reviews, business, politics, recommender system etc. Based on the opinion about a product or about different aspects of a product, an organization can make changes accordingly. Similarly based on the opinion about a particular political party, government policies' changes can be made accordingly. Two main techniques used for sentiment analysis are machine learning based and lexicon based.

Supervised, Unsupervised and Semi-supervised comes under Machine Learning. Supervised approaches e.g. SVM[2][11][13][14][19][30][32], KNN[20][21], Naive Bayes[3][4][6][7] etc. requires a good quality training set and thus are highly domain dependent but provide better results if trained properly. Unsupervised approaches e.g. K-Means, Self Organizing Maps (SOM) etc. do not make use of training set. Semi supervised approaches require partial labelling of data and are of two types: a) Transductive Learning b) Inductive Learning

Lexicon based approach makes use of dictionary consist of labelled words and with the help of these words, a text is judged whether it is subjective or objective[16][23][25]. This approach is further divided into a) Dictionary based which does not take into account the context of word within a text, and b) Corpus based which expands the dictionary with taking associations between different words into account. A complete survey in this field is provided in [9].
This research analyze sentiment of tweets[1][2][3][4][5][8]. As tweets are very unstructured in nature this research converts them into useful information so that better features can be used for machine learning. Hence in this research we provide a good data preprocessing to tweets followed by hybrid classifier. With the help of processed tweets or data features are generated and fed to the two machine learning algorithms KNN and SVM in a hybrid manner. Different feature have tried by authors for improving the results as in [2][3[4][5][18].

Support Vector Machines(SVM) : SVM invented by Vapnik & Chervonenkis in 1963 is a supervised machine learning algorithm used basically for classification and regression problems. It solves an optimization problem of finding the maximum margin hyperplane between the classes. This is basically required to avoid overfitting. Basically it is a linear classifier separating the classes which can be separated with the help of linear decision surfaces called hyperplanes. Foe classes having binary features SVM draws a line between the classes and for classes having multiple features hyperplanes are drawn. However it can be used for classifying the non linear data also by transforming the feature space into the higher dimensional space so that non linear data in higher dimensional can be separated easily by a hyperplane.
This transformation is made easy with the help of Kernel-trick. With the help of kernels it is not necessary to calculate all the dimensions when transform and calculation of hyperplane can be done in the same lower dimensional feature space. Kernels are not used only for this purpose but also for making the calculation easier in case of many features. Various kernels are used by machine learning approaches e.g. RBF(Radial Basis Function), Linear kernel, Poly kernel etc.

Kernels make the calculation very fast and helps in improving the calculation time remarkably.
K-Nearest Neighbors(KNN) : It is also a supervised learning algorithm based on the classes nearest to the point which is to be classified. Based on the values of the K nearest classes a test set is provided the majority voting class. However to improve this algorithm weights are assigned to each of

the k points according to their distance from the test point. The value of k depends upon the classification problem and the size of dataset.
This research uses the prediction probability of both the algorithms on each test tweet and assign the class based on greater probability. Our approach provides better results as compared to using these algorithms entirely in isolation. This research paper is organised as: 2. provides literature review and 3. provides proposed methodology followed by 4. composed of experimental results and evaluation. 5. provides conclusion and scope.

2. LITERATURE SURVEY

Various researchers have been working on twitter[3][4][5][11][12][16] and from time to time they are publishing their researches . They have used various sentiment analysis techniques for improving the results of classification Their work is also helpful in this research as the sentiment analysis techniques they have used, feature selection techniques, different pre-processing steps they have used is taken care of in this research. This research mainly focuses on supervised approach for sentiment analysis task and has surveyed researches both for twitter and non- twitter data and also for both supervised and lexicon based approaches for better clarification and understanding of the topic chosen.

Many researches defined multiple faces of sentiment analysis as opinion orientation, feature extraction etc. Machine learning classifiers need various features for learning so different researchers from time to time have selected different features for comparing results.

Agarwal et al.[02], Pak and Paroubek[03], Spancer and Uchyigit[04], Koloumpis et al.[05] selected various features as unigrams, bigrams, pos tagging, hash tags, ngrams etc. and found mixed response in classification results. Different features and feature selection methods as semantic features and concepts, information gain, chi-square etc. has been used by Hassan Khan et al.[13], Agarwal et al.[14].

Hassan Khan et al.[13] approach includes rigorous data pre-processing followed by supervised machine learning. They collected labelled datasets of different domains so that machine learning will not be limited to a particular domain. To learn SVM classifier they make use of different training sets each make SVM learn different feature sets -1) Information gain(IG) with feature presence and 2) feature frequency 3) Cosine similarity with feature presence and 4) feature frequency. They found that feature presence is better than feature frequency.

Agarwal et al.[14]], found that for better results using machine learning approaches, finding good features is a challenging task. They gave the concept of "Semantic Parser" and treated concepts as features. They used the minimum Redundancy and Maximum Relevance (mRMR) feature selection mechanism. They used different feature sets for their classification task e.g. unigrams, bigrams, bi-tagged and dependency parse tree along with their proposed scheme so that results can be compared with.

Various approaches and classifiers such as lexicon based approach, Naive Bayes(NB), Support Vector Machines(SVM), Maximum Entropy(MaxEnt) etc. have been used time to time with various parameters for evaluating the results as accuracy, precision, recall, f-measure etc. Narr et al[06] concluded 71.5% accuracy with mixed language NB classifier on unigrams. Saif et al.[07] concluded

that semantic features used by NB classifier increase f1-measure against unigram by 6.47% and pos+unigram by 4.78%.
Asmi[24], Hutto[25], Neviarouskaya[26] proposed rule based approaches for increasing the accuracy. Swati[27], Chikersal[29] and Prabowo and Thelwall[10] proposed hybrid approach consisting of rule based and machine learning classifiers.

Hybrid approaches consisting of machine learning classifiers have been underexplored in the literature with very few researches in this approach as in Revathy[28], F.F. da Silva et al.[11]. F.F. da Silva et al.[11] proposed an ensemble based classification in which various classifiers e.g. SVM, Multinomial Naive Bayes, Random Forest, Logistic Regression are used. They proposed that if we train the different classifiers with different training sets and then by using either average probabilities of different classifiers or maximum voting, we get better results than by using only a single classifier. Moreover they uses two different features for learning the classifiers:-
a) Bag of Words(BOW) b) Feature Hashing
They used four different datasets for training and testing. They found that Feature Hashing is not better than BOW approach in most of the datasets except one.

Our research work mainly focuses on combining the machine learning classifiers and proves that combining gives better results as compared to standalone classifiers. Also this research gives comparative results as against to the feature hashing+lexicon based features used by [11] ,with only a small dataset and few features.

Bhadane et al[15], Apple et al[16] proposed combination of sentiment lexicon with machine learning approaches and found increase in accuracy. Muhammad et al.[17] handled word's polarity in terms of local and global context by giving SmartSA system and found that their system is superior to baseline lexicons and systems like SVM, NB etc. with more F1 score.
Addlight and Supreethi[20] compared two machine learning methods KNN and SVM and found that SVM outperforms KNN. Saif and He[23] gave the concept of SentiCircles for calculating the context of words. They found that it is necessary for better sentiment classification.

Jianqiang et al.[32] discussed the role of rigorous preprocessing in increasing the evaluation measure and gave six different preprocessing methods for the same. Keeping this in mind our approach also uses a good preprocessing to filter the tweets. Khan and Jeong[21] proposed an approach for finding the sentiments about each aspect of a product and this can be a good future work to explore.

3. PROPOSED METHODOLOGY

The proposed hybrid model is defined as the three stage model. In first preprocessing stage of this model, the multi-aspect based filtration and impurities correction is applied. The spell correction, stemming, abbreviation expansion, stopwords removal are defined in this stage to normalize the input tweets. In this stage, the separation of tag tokens, positive aspects and negative aspects from messages is also defined. Negation handling is also done. In second stage, the filtered text is processed to generate the statistical features. In this stage, the transformation of input training and testing set is done to corresponding feature set. These features are processed by the Hybrid classifier for sentiment prediction in final stage of this model. In classification stage, the probabilistic predictive decision is applied for selection of KNN or SVM classifier for individual instance.

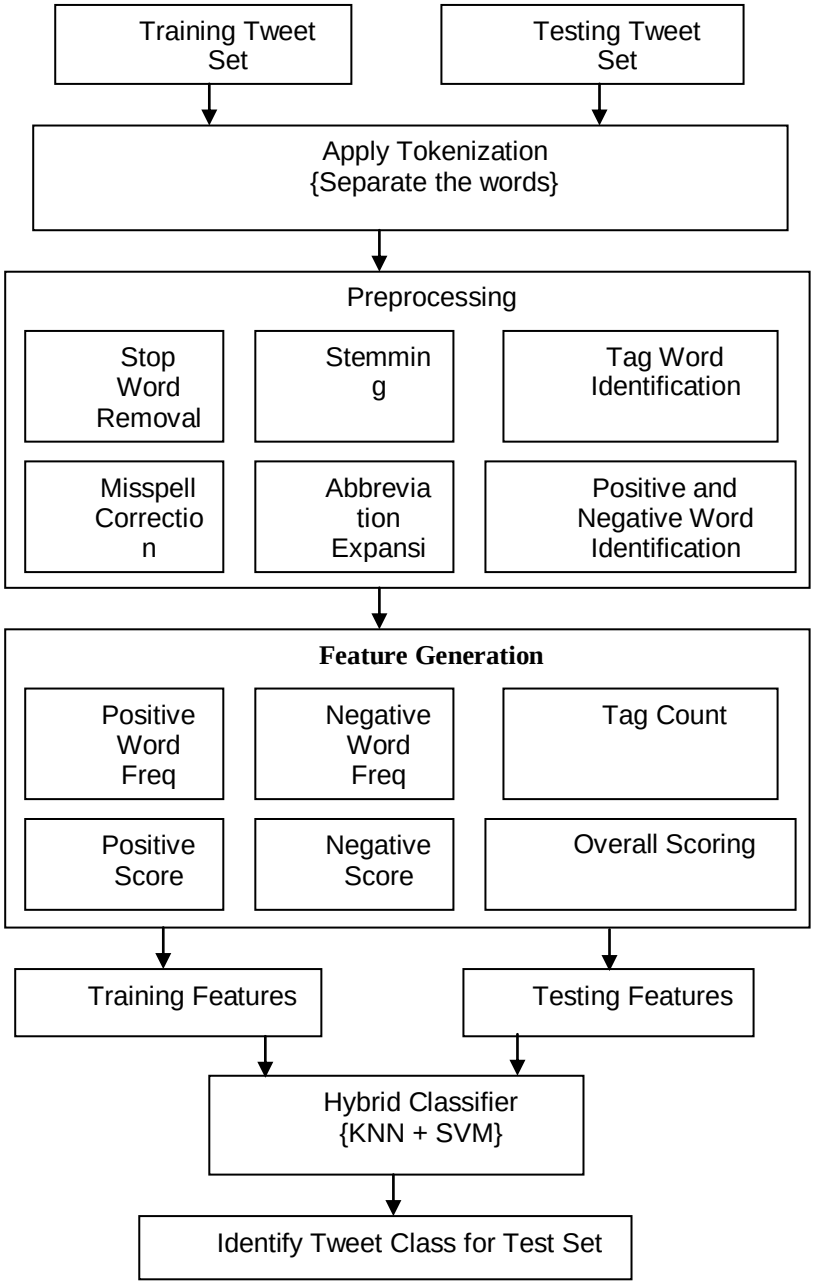


Fig 1: Flowchart of the proposed system

The classification is applied on the tweets acquired from the web. The description of dataset is given below in Table 1:

Features	Values
Dataset Name	Twitter-Sentiment-Analysis-FinalizedFull
Dataset Url	https://github.com/TharinduMunasinge/Twitter-Sentiment-Analysis
Number of Tweets	997
Classes	Positive Tweets , Negative Tweets and Neutral Tweets
File Type	CSV

3.1 Preprocessing:

During preprocessing various steps are taken as stopword removal, handling of negation, abbreviation expansion, misspell correction, stemming, positive word lists of each tweet, negative word lists of each tweet. Porters algorithm is used for stemming.

Data

SrNo	Tweet	Filtered List	Tag Filtere...	Negative L...	Positive List	G	H
1	@united U...	[@unit, ua...	[ua5396, ...	[crap]	[get]		
2	I hate Tim...	[hate, time...	[hate, time...	[hate, porn]	[warner, li...		
3	Tom Shan...	[tom, shan...	[tom, shan...	[]	[]		
4	Found the ...	[found, sel...	[found, sel...	[]	[]		
5	@united a...	[@unit, arri...	[arriv, ye, fli...	[miss, slow]	[]		
6	Driverless ...	[driverless...	[driverless...	[]	[]		
7	how can y...	[not, love, ...	[not, love, ...	[joke]	[love]		
8	Safeway is...	[safewai, r...	[safewai, r...	[]	[rock]		
9	RT @jquer...	[rt, @jqueri...	[rt, ultim, jq...	[]	[]		
10	I saw Nigh...	[night, mu...	[night, mu...	[]	[]		
11	Missed thi...	[miss, , ge...	[miss, gen...	[miss]	[]		
12	is being fu...	[fuck, time,...	[fuck, time,...	[fuck, suck]	[warner]		
13	I hope the ...	[hope, girl,...	[hope, girl,...	[]	[hope]		
14	@aparajul...	[@aparaju...	[good, luck]	[]	[good, luck]		
15	needs so...	[explain, la...	[explain, la...	[]	[]		
16	@united T...	[@unit, tha...	[thank, ma...	[]	[thank, get]		
17	@ontheMA...	[@onthem...	[ditto!, not, ...	[]	[good]		
18	waiting in l...	[wait, line, ...	[wait, line, ...	[]	[]		
19	OMG, I wo...	[oh my go...	[oh my go...	[died, no]	[good]		
20	Theres a g...	[there, goo...	[there, goo...	[]	[]		
21	#MBA Adm...	[mba, adm...	[mba, adm...	[]	[]		
22	am loving ...	[morn, lov...	[morn, lov...	[outlier]	[love]		
23	Goodby, Si...	[goodbi, si...	[goodbi, si...	[]	[enjoi]		
24	12 Gift Ide...	[12, gift, id...	[12, gift, id...	[]	[lover]		
25	So the #C...	[coachella...	[coachella...	[]	[]		
26	New blog ...	[blog, post...	[blog, post...	[]	[]		
27	whoever is...	[whoever, r...	[whoever, r...	[rape, out]	[warner, u...		
28	@Donnie...	[@donnie...	[tell, spoke...	[]	[right, hop...		
29	Three Chi...	[china, aer...	[china, aer...	[]	[invest]		
30	Ok, first as...	[ok, asses...	[ok, asses...	[fuck]	[ok]		
31	heyy loves!...	[heyi, love...	[heyi, love...	[kick]	[loves]		
32	@united w...	[@unit, we...	[well, john,...	[]	[well]		
33	I loved tod...	[love]	[love]	[]	[love]		
34	RT @Wate...	[rt, @water...	[rt, ca, mer...	[]	[profit, well]		

Fig 2: Showing tweets after preprocessing

For stopwords, abbreviation expansion[1] and misspell correction[2] database is created.

Filtered list:-contains tweets after tokenization and applying the above written filters

Tag Filtered list:-contains the filtered list with @ tags removed. The @ tags are used in feature generation as tag count in each tweet.

Negative List:-contains negative adjectives in each tweet.

Positive List:-contains positive adjectives in each tweet.

3.2 Features Generation:

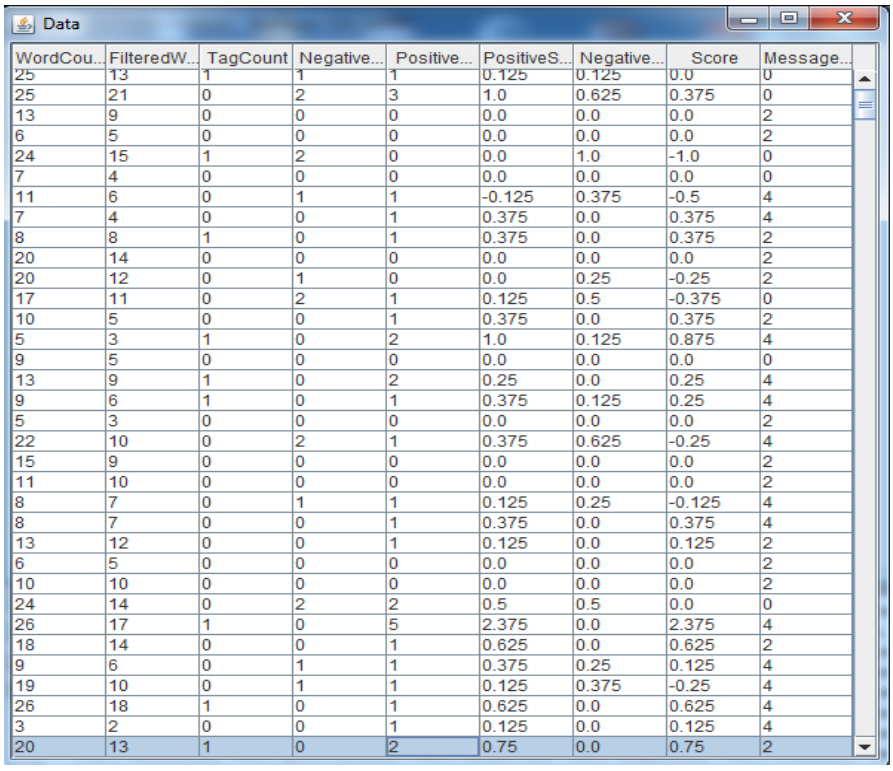
A list of adjectives[3] is used for features generation. This list contains positive score, negative score, overall rating of an adjective among other attributes.

Table 2: Describing various attributes of an adjective

Attributes	Description
Id	Numeric Unique id to all adjectives
Adjective	Stores the textual information to represent the actual adjective
Pscore	Positive score, to represent the positive acceptability of an adjective Lies between 0 & 1
Fscore	Negative score, Lies between 0 & 1
Score	Overall score of adjective lies between -1 & 1 +ve values for +ve adjective -ve value for –ve adjective

1: http://www.illumasolutions.com/omg-plz-lol-idk-idc-btw-brb-jk.htm
2: https://noisy-text.github.io/norm-shared-task.html
3: http://www.sentix.de/index.php/en/item/sentix-website.html

Various features are generated after filtering of tweets for learning the classifiers.



WordCou...	FilteredW...	TagCount	Negative...	Positive...	PositiveS...	Negative...	Score	Message...
25	13	1	1	1	0.125	0.125	0.0	0
25	21	0	2	3	1.0	0.625	0.375	0
13	9	0	0	0	0.0	0.0	0.0	2
6	5	0	0	0	0.0	0.0	0.0	2
24	15	1	2	0	0.0	1.0	-1.0	0
7	4	0	0	0	0.0	0.0	0.0	0
11	6	0	1	1	-0.125	0.375	-0.5	4
7	4	0	0	1	0.375	0.0	0.375	4
8	8	1	0	1	0.375	0.0	0.375	2
20	14	0	0	0	0.0	0.0	0.0	2
20	12	0	1	0	0.0	0.25	-0.25	2
17	11	0	2	1	0.125	0.5	-0.375	0
10	5	0	0	1	0.375	0.0	0.375	2
5	3	1	0	2	1.0	0.125	0.875	4
9	5	0	0	0	0.0	0.0	0.0	0
13	9	1	0	2	0.25	0.0	0.25	4
9	6	1	0	1	0.375	0.125	0.25	4
5	3	0	0	0	0.0	0.0	0.0	2
22	10	0	2	1	0.375	0.625	-0.25	4
15	9	0	0	0	0.0	0.0	0.0	2
11	10	0	0	0	0.0	0.0	0.0	2
8	7	0	1	1	0.125	0.25	-0.125	4
8	7	0	0	1	0.375	0.0	0.375	4
13	12	0	0	1	0.125	0.0	0.125	2
6	5	0	0	0	0.0	0.0	0.0	2
10	10	0	0	0	0.0	0.0	0.0	2
24	14	0	2	2	0.5	0.5	0.0	0
26	17	1	0	5	2.375	0.0	2.375	4
18	14	0	0	1	0.625	0.0	0.625	2
9	6	0	1	1	0.375	0.25	0.125	4
19	10	0	1	1	0.125	0.375	-0.25	4
26	18	1	0	1	0.625	0.0	0.625	4
3	2	0	0	1	0.125	0.0	0.125	4
20	13	1	0	2	0.75	0.0	0.75	2

Fig 3: Showing the results of features generation

Various features used for learning the classifiers are:
Word count:-Total words in each tweet after filtration
Tag count:-Total @ tags used in each tweet
Negative word count:-Total negative words in each tweet
Positive word count:-Total positive words in each tweet
Positive score:-Total positive score obtained by adding the positive scores of each positive adjective.
Negative score:-Total negative score obtained by adding the negative scores of each negative adjective.
Score:-Positive score-Negative score for each tweet
Message class 0: for negative tweets
1: for neutral tweets
2: for positive tweets

3.3 Classification:

After features generation classification is done with our hybrid approach in which prediction probability of both the classifiers is used which is shown in the algorithm below:

Algorithm 1:

Classification(TrainingSet,TestingSet)

/*TrainingSet is the Training Tweet Set and TestingSet is the Testing Tweet Set on which features are generated */

{

1. TrainFeaturesSet=FeatureGeneration(TrainingSet)
   /*Generate Features for Training Set*/
2. TestFeaturesSet=FeatureGeneration(TestSet)
   /*Generate Features for Testing Set*/
3. SWeight=GenerateWeight(TrainFeaturesSet,SVM)
   /*Process the Classifier, train with the Training Feature set and Generate Feature weights for SVM*, KNN is trained directly during testing/
4. For i=1 to TestFeatureSet.Length
   /*Process the Testing Instances*/
   {
5. K1=Predict(TestFeatureSet(i),TrainFeaturesSet)
   /*Apply Prediction on Test Instance respective to KNN Classifier Weight*/
6. S1=Predict(TestFeatureSet(i),SWeight)
   /*Apply Prediction on Test Instance respective to SVM Classifier Weight*/
7. If (K1>Th1 And S1>Th1)
   /*Apply Hybrid Classifier for Test Class Identification, Th1 is the threshold used for prediction probability, Th1=0.5 is used here*/
   {
8. TestFeatureSet(i).Class=IdentifyClass(greater(K1,S1))
   }


```
9.  Else If (K1>Th1)
    /*Apply KNN Classifier for Test Class Identification*/
    {
10. TestFeatureSet(i).Class=IdentifyClass(K1)
    }
11. Else
12. /*Apply SVM Classifier for Test Class Identification*/
    {
13. TestFeatureSet(i).Class=IdentifyClass(S1)
    }
    }
Return TestFeatureSet.Class
}
```

4. EXPERIMENTAL RESULTS AND EVALUATION

In this present work, the SVM and KNN based hybrid classification model is presented to process the tweet features and to identify the hidden sentiments from these tweets.

Implementation is done in Netbeans8.0 with Weka(3.8) integrated into it. Weka is widely used in data mining for preprocessing, clustering, classification etc. and gives results in terms of accuracy, precision ,recall, f-measure etc. MySql is used for storing the various datasets as list of adjectives, abbreviations, misspell corrections, training dataset and testing dataset.

For running classifiers KNN and SVM in isolation, weka is used directly but for their combination weka is integrated into netbeans and confusion matrix and results are manually calculated with the help of results.

The comparative analysis is provided against KNN and SVM based methods separately.

K=15 is used for comparison between KNN and Hybrid approach.

The description of processing training and testing set is shown in Table 3:

Features	Values
Size of Training Set	699(267-positive,264-negative,168-neutral)
Size of Testing Set	298(114-positive,113-negative,71-neutral)
Tweet Classes	Positive, Negative, Neutral
Existing Methods	KNN & SVM
Proposed	Hybrid KNN+SVM

The classification algorithm combining KNN and SVM is given in section 2. Confusion matrix for KNN and SVM is taken from weka by opening the TrainFeaturesSet and TestFeaturesSet there directly and is provided below:

Table 4: Confusion matrix for KNN

		Predicted		
Actual	Class	Negative	Neutral	Positive
	Negative	**77**	19	17
	Neutral	9	**46**	16
	Positive	21	14	**79**

Table 5: Confusion Matrix for SVM

		Predicted		
Actual	Class	Negative	Neutral	Positive
	Negative	**77**	14	22
	Neutral	11	**47**	13
	Positive	16	20	**78**

Confusion matrix for hybrid approach is calculated manually by the analysis results and with the help of confusion matrix precision, recall and f-measure is calculated.

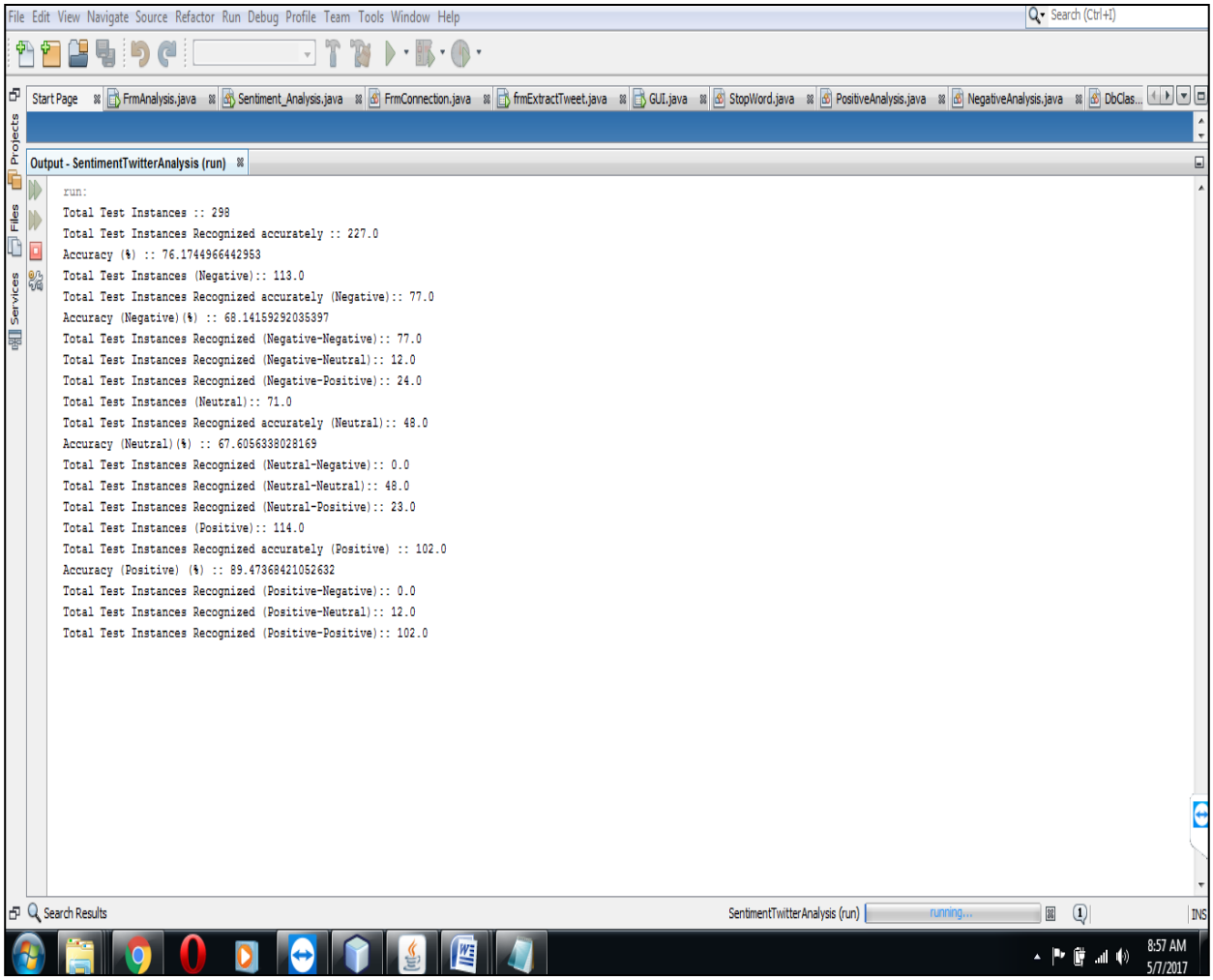


Fig 4: showing analysis results

Analysis results shows True Positives(TP), False Positives(FP), True Negatives(TN) and False Negatives(FN) for each sentiment class. Thus confusion matrix is derived below:

Table 6: Confusion Matrix for KNN+SVM

		Predicted		
	Class	Negative	Neutral	Positive
Actual	Negative	**77**	12	24
	Neutral	0	**48**	23
	Positive	0	12	**102**

With the help of confusion matrices Accuracy, Precision, Recall and F-measure for positive, negative and neutral classes are calculated and is also compared for the 3 approaches used above.

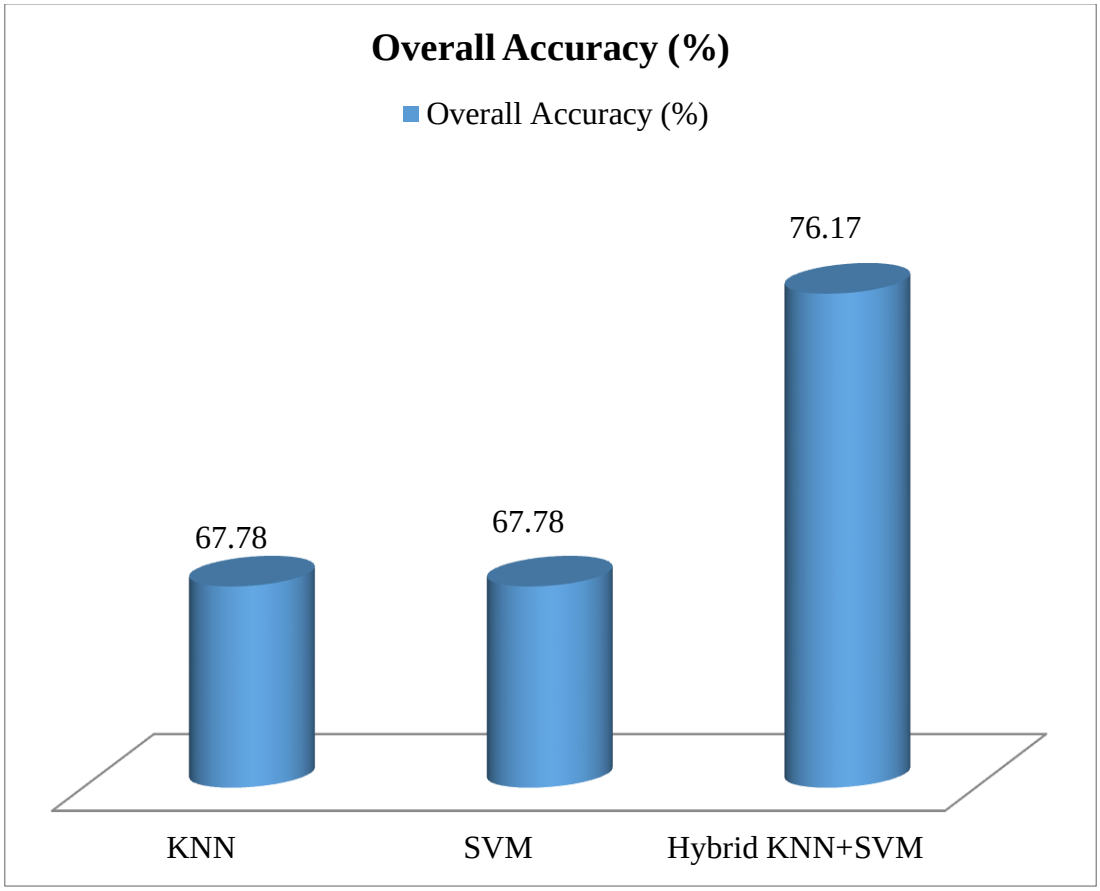


Fig 5: showing overall accuracy for 3 approaches

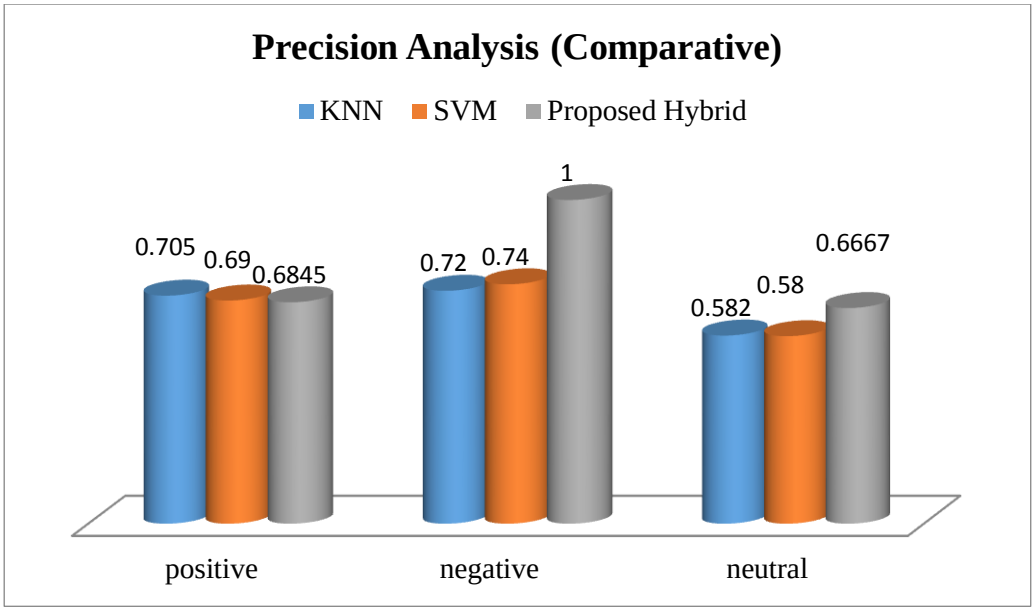


Fig 6: showing precision analysis

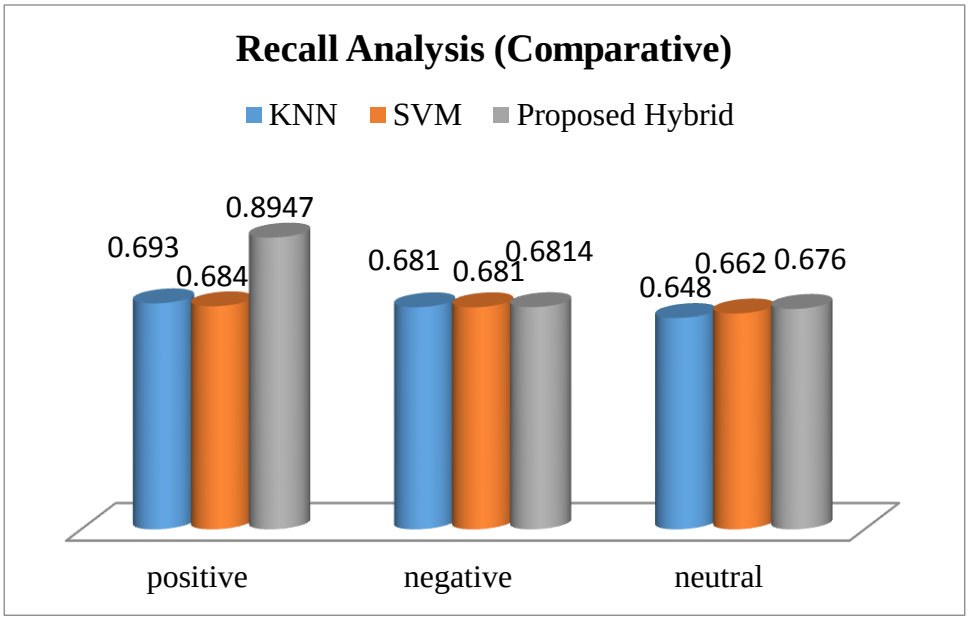


Fig 7: showing recall analysis

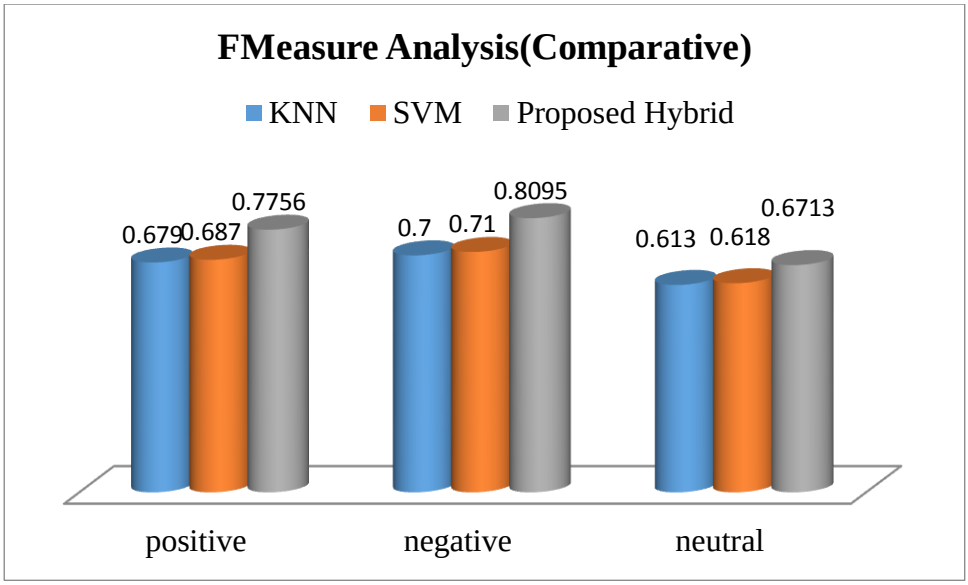


Fig 8: showing f-measure analysis

Thus the above results show that our hybrid approach works better both in terms of accuracy and f-measure.

Comparison of our approach with that of [11]

This research with only a small dataset and few features gives comparative results and even better results as compared to combination of Logistic Regression(LR), Random Forest(RF) And Multinomial Naive Bayes(MNB) along with Feature Hashing and Lexicon features used in [11].

The comparison is shown below in Table 7,Avg is used in place of Average, pos in place of positive, neg in place of negative:

Table 7

Datasets	Avg F-measure(%) (pos,neg)	Avg F-measure(%) (including neutral)	Accuracy(%) (pos,neg)	Accuracy(%) (including,neutral)
OMD[11]	65.35	-	70.62	-
Strict OMD[11]	71.80	-	74.56	-
Sanders[11]	76.25	-	76.63	-
Stanford [11]	78.25	-	79.11	-
HCR[11]	62.20	-	78.35	-
Our dataset (Tharindu Munasinge)	**79.25**	**75.21**	**78.80**	**76.17**

5. Conclusion and Scope

In this paper, a SVM and KNN based hybrid model is presented to improve the classification accuracy. The proposed method classified the tweets in positive, negative and neutral sentiments whereas much of the literature in this field is associated with 2-way classification[10][11]. The work of proposed model has gone through preprocessing stage, features generation stage and classifiers learning stage. The analytical evaluation of proposed model is done in terms of accuracy and f-measure.. The comparative observations

are taken against the SVM and KNN methods. The comparative results shows that the proposed model has improved the accuracy and f-measure of tweet class prediction.

As number of features for learning the classifiers are limited in our approach, we will be using more features and better feature selection methods like Information Gain, Chi-Square etc. in our future work. Our comparison with literature[11] shows that increasing our dataset with more tweets and features can also help in increasing reasonable accuracy and f-measure. Other machine learning methods in combined way can also be explored in the future.

References

[1] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1 (2009): 12.

[2] A Agarwal, B Xie, I Vovsha, O Rambow. "Sentiment analysis of twitter data." Proceedings of the workshop on languages in social media. Association for Computational Linguistics, 2011.

[3] Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREc. Vol. 10. 2010.

[4] Spencer, James, and Gulden Uchyigit. "Sentimentor: Sentiment analysis of twitter data." Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. 2012.

[5] Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. "Twitter sentiment analysis: The good the bad and the omg!." Icwsm 11 (2011): 538-541.

[6] Narr, Sascha, Michael Hulfenhaus, and Sahin Albayrak. "Language-independent twitter sentiment analysis." Knowledge Discovery and Machine Learning (KDML), LWA (2012): 12-14.

[7] Saif, Hassan, Yulan He, and Harith Alani. "Semantic sentiment analysis of twitter." International Semantic Web Conference. Springer Berlin Heidelberg, 2012.

[8] Carpenter, Thomas, and Thomas Way. "Tracking Sentiment Analysis through Twitter." Proceedings of the International Conference on Information and Knowledge Engineering (IKE). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.

[9] Walaa Medhat, Ahmed Hassan, Hoda Korashy, “Sentiment analysis algorithms and applications: A survey”, Ain Shams Engineering Journal (2014) 5, 1093–1113.

[10] Prabowo, Rudy, and Mike Thelwall. "Sentiment analysis: A combined approach." Journal of Informetrics 3.2 (2009): 143-157.

[11] Da Silva, Nadia FF, Eduardo R. Hruschka, and Estevam R. Hruschka. "Tweet sentiment analysis with classifier ensembles." Decision Support Systems 66 (2014): 170-179.

[12] Khan, Farhan Hassan, Saba Bashir, and Usman Qamar. "TOM: Twitter opinion mining framework using hybrid classification scheme." Decision Support Systems 57 (2014): 245-257.

[13] Khan, Farhan Hassan, Usman Qamar, and Saba Bashir. "A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet." Knowledge and Information Systems (2016): 1-22.

[14] Agarwal, Basant,Soujanya Poria,Namita Mittal,Alexander Gelbukhand Amir Hussain. "Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach." Cognitive Computation 7.4 (2015): 487-499.

[15] Bhadane, Chetashri, Hardi Dalal, and Heenal Doshi. "Sentiment analysis: Measuring opinions." Procedia Computer Science 45 (2015): 808-814.

[16] O Appel, F Chiclana, J Carter, H Fujita. "A hybrid approach to the sentiment analysis problem at the sentence level." Knowledge-Based Systems 108 (2016): 110-124.
[17] Muhammad, Aminu, Nirmalie Wiratunga, and Robert Lothian. "Contextual sentiment analysis for social media genres." Knowledge-Based Systems 108 (2016): 92-101.
[18] Zhu, Dengya, and Jitian Xiao. "R-tfidf, a Variety of tf-idf Term Weighting Strategy in Document Categorization." Semantics Knowledge and Grid (SKG), 2011 Seventh International Conference on. IEEE, 2011.
[19] Anjaria, Malhar, and Ram Mohana Reddy Guddeti. "A novel sentiment analysis of social networks using supervised learning." Social Network Analysis and Mining 4.1 (2014): 1-15.
[20] Mukwazvure, Addlight, and K. P. Supreethi. "A hybrid approach to sentiment analysis of news comments." Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), 2015 4th International Conference on. IEEE, 2015.
[21] Khan, Jawad, and Byeong Soo Jeong. "Summarizing customer review based on product feature and opinion." Machine Learning and Cybernetics (ICMLC), 2016 International Conference on. IEEE, 2016.
[22] Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." Journal of Big Data 2.1 (2015): 5.
[23] H Saif, Y He, M Fernandez, H Alani. "Contextual semantics for sentiment analysis of Twitter." Information Processing & Management 52.1 (2016): 5-19.
[24] Asmi, Amna, and Tanko Ishaya. "Negation identification and calculation in sentiment analysis." The Second International Conference on Advances in Information Mining and Management. 2012.
[25] Hutto, Clayton J., and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." Eighth International AAAI Conference on Weblogs and Social Media. 2014.
[26] Neviarouskaya, Alena, Helmut Prendinger, and Mitsuru Ishizuka. "Semantically distinct verb classes involved in sentiment analysis." IADIS AC (1). 2009.
[27] Kawathekar, Swati A., and Manali M. Kshirsagar. "Sentiments analysis using Hybrid Approach involving Rule-Based & Support Vector Machines methods." IOSRJEN 2.1 (2012): 55-58.
[28] Revathy, K., and B. Sathiyabhama. "A hybrid approach for supervised twitter sentiment classification." International Journal of Computer Science and Business Informatics 7 (2013).
[29] Chikersal, Prerna, Soujanya Poria, and Erik Cambria. "SeNTU: sentiment analysis of tweets by combining a rule-based classifier with supervised learning." SemEval-2015 (2015): 647.
[30] 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, 2004, pp.168-177.
[31] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.
[32] Z. Jianqiang and G. Xiaolin, "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis," in IEEE Access, vol. 5, no. , pp. 2870-2879, 2017.

