Machine learning is used to classify and mine tweets on multiple levels

Ravi kumar kuchipudi ,research scholar, ANU, Computerscience and Engineering,ravikumary4@gmail.com

Dr V.V. Jaya Rama Krishnaiah, Associate professor, ASN Engineering college ,Tenali

jkvemula@gmail.com

ABSTRACT

Research on sentiment analysis can be found under the discipline of Natural Language Processing (NLP). It is helpful in determining the feelings and meanings that lie beneath the surface of a piece of literature. When attempting to extract sentiments from Twitter data, you face a particularly unique combination of obstacles as a result of the platform's unstructured format, its relatively small size, and the presence of slang, misspellings, and acronyms. The vast majority of researchers analysed the results of multiple sentiment analysis machine learning algorithms and compared their findings. However, merging these methodologies is a topic that has received insufficient attention in the relevant academic literature. According to the findings of this research, mixing many different machine learning algorithms together produces superior outcomes compared to doing it individually. Due to the unprocessed nature of the tweets, this study employs a wide range of preprocessing techniques in order to generate data that can then be used in machine learning classification models. An inquiry into the potential benefits of combining the machine learning algorithms known as KNN and SVM is the primary focus of this work. When making an analytical observation, the classification accuracy and F-measures for each emotion class, as well as the average of those values, are used as inputs. The results of the evaluation show that the proposed hybrid approach is more accurate and has a better F-measure than individual classifiers.

INTRODUCTION

The vast amounts of data that are accessible on the internet have piqued the curiosity of a number of companies, which have launched initiatives to mine this information for useful insights. As a direct consequence of this, a brand-new academic subfield known as sentiment analysis came into being. Opinion mining, opinion extraction, opinion mining, and a number of other names have been used to refer to this area of study. To make things crystal clear, there is a nuanced difference in meaning between each of these several formulations. Before automated sentiment mining was developed, more traditional surveying methods were used.

Because the tweets of individual users contained a substantial amount of bias, it was necessary to implement an automated system in order to deal with the hundreds of thousands of opinions that were buried in reviews, blogs, and other forms of user-generated content. Analysis of people's feelings can be put to a wide variety of purposes; only a few of them are product reviews, movie reviews, business, politics, and recommendation systems. An organisation has the ability to make adjustments in response to the feedback it receives from customers regarding a product or various aspects of a product. In a similar vein, adjustments to the policies of the government might be made in accordance with the attitudes of the people toward a certain political party. The use of machine learning and sentiment analysis that is lexicon-based are two of the most popular approaches that are taken.

Everything falls under the general category of machine learning. Many of these techniques, such as SVM, KNN, Naive Bayes, and K-means clustering, have been supervised, such as [2] [3] [4] [6] [7], etc., require a high-quality training set and are therefore very domain-dependent; however, they deliver better results when properly trained. [2] [3] [4] [6] [7], for example, necessitate a high-quality training set and are thus domain-dependent;[[Unsupervised algorithms like K-Means and Self-Organizing Maps do not make use of the training set in their computations (SOM). There are two different kinds of semi-supervised methods: those that need some labelling of the data and those that don't need it at all. Both transductive learning and inductive learning are forms of learning (b).

It is feasible to assess whether or not a piece of writing is subjective or objective by consulting the dictionary and looking up the words marked [16], [23], and [25]. a) dictionary-based, in which the context of the term within a text is ignored; and b) corpus-based, in which the dictionary is expanded by considering the links between distinct terms.You may find an in-depth analysis of this subject in the following reference [9].

This study investigates the range of feelings that can be found in tweets. This work converts tweets into data that could be used to improve machine learning. This is necessary due to the unstructured nature of tweets. As a result, the purpose of this research is to demonstrate an effective preprocessing of tweets followed by a hybrid classifier. Both the KNN and SVM machine learning algorithms are fed in a hybrid approach using processed tweets or data attributes as the input. In an effort to produce better results, the authors have played with a variety of different aspects of the system.

In 1963, Vapnik and Chervonenkis created a method for supervised machine learning that is used for classification and regression. This system was given the name Support Vector Machine (SVM). The solution to an optimization problem can lead to the discovery of class margin hyperplanes. This is absolutely necessary in order to prevent the garment from being over-sized. To put it in its most basic terms, it is a linear classifier that divides classes by the use of linear decision surfaces called hyperplanes. The Support Vector Machine (SVM) draws a line between classes that have binary features, whereas the hyperplane method draws lines between classes that have multiple features. By giving the feature space more dimensions, which makes it easy for a hyperplane to separate data that doesn't follow a straight line, this can be used to classify data that doesn't follow a straight line.

This transition is made easier by the use of the kernel-trick. It is not necessary to calculate all dimensions if the transform and hyperplane computations may be conducted in the same lower-dimensional feature space thanks to kernels. In this case, the total number of dimensions does not need to be calculated. Not only are kernels useful in situations such as this one, but they can also be put to use to expedite calculations in situations involving characteristics that are particularly complex. Many different types of kernels are utilised by machine learning systems. Some of these kernels are RBF (Radial Basis Function), a linear kernel, a polygonal kernel, and many others.

The use of kernels, which considerably accelerate the process, enables the performance of calculations at a much quicker rate. K-Nearest Neighbors (KNN) is yet another technique for supervised learning that is predicated on the classes that are located nearest to the point of categorization. A test set that is determined by the values of the K classes that are nearby is sent to the voting class that receives the majority of the votes. The locations in the algorithm are given weights according to how far away they are from the test point. The algorithm has k locations total. The quantity of the dataset is another factor that influences the value of k and can produce classification problems of varying degrees of difficulty.

In this investigation, the prediction probabilities of the algorithms are compared for each of the test tweets, and the class is determined according to the method that has a higher probability of accurate prediction. Our strategy produces superior results because it integrates a number of different approaches rather than just one. The following outline constitutes the framework of this thesis: The second part of the report, which is titled "presents a review of the literature and a proposal for a methodology," is followed by the fourth section, which is titled "experimental data and evaluation," and then the final section, which is titled "conclusion and recommendations."

Twitter has been the subject of research by a great number of academics and scientists, and these individuals have periodically disseminated their findings. The classification results of their work have been improved by using a range of sentiment analysis tactics, and this research addresses the approaches for sentiment analysis, feature selection techniques, and other preprocessing stages that they used. This study has looked at studies on Twitter data as well as data from other sources, as well as supervised and vocabulary-based approaches. This was done for the purpose of better clarifying and understanding the topic that was selected.

Several scholars have used terms such as opinion orientation and feature extraction, amongst others, to refer to what is now known as sentiment analysis. When it comes to machine learning-based classifiers, different researchers have chosen different features to compare the results of their work.

The classification results that were obtained by Agarwal et al., Pak and Paroubek, Spancer and Uchyigit, and Koloumpis et al. [05] when they utilised a variety of features, such as unigrams, bigrams, and pos-tagging, were met with a variety of responses. Hassan Khan et al. [13] and Agarwal et al. [14] are two researchers who used information gain and chi-square to choose concepts and semantic characteristics.

Hassan Khan et al. [13] utilise supervised machine learning, but only after subjecting the data to stringent preprocessing. In order to widen the usefulness of machine learning, they gathered labelled datasets from a range of domains and used those datasets. In order to train the SVM classifier, they use a variety of training sets. Each training set teaches the SVM a different collection of characteristics, and the SVM learns these features from the individual training sets. 3) The cosine measure's correlation with feature presence, and 4) the frequency of feature presenceIn this study, it was found that the most important factor was the presence of features, not how often they appeared.

It is a challenging task, as stated by Agarwal et al. [14], to obtain good features for machine learning algorithms. [Citation needed] The term "Semantic Parser" was presented as a concept, and its properties were considered to be those of a concept. The mRMR feature selection method was utilised by them in the process of picking features. Along with the system that they proposed, they made use of a number of other feature sets, including unigrams, bigrams, bi-tagged, and dependency parse trees, when attempting to solve their classification problem.

It is fairly unusual to use lexicon-based techniques such as Naive Bayes (NB), Support Vector Machines (SVM), and Maximum Entropy (MaxEnt) with a large number of parameters to evaluate the results. Some of these parameters include accuracy, precision, recall, and f-measure. Narr et al. [06] discovered that a mixed-language NB classifier has an accuracy rate of 71.5 percent on unigrams.When semantic attributes were included, it was demonstrated that the NB classifier's f1-measure against unigram increased by 6.47 percent, and that f1-measure against posunigram increased by 4.78 percent [07]. Accuracy can be improved using the rule-based methods that Hutto and Neviarouskaya have given. Swati, Chikersal, Prabowo, and Thelwall [27, 29, 10, and 11] introduced a hybrid methodology by utilising a combination of rule-based and machine learning classifiers.

There are a few studies on hybrid strategies incorporating machine learning classifiers that can be found in the research literature. Some examples of these studies include Revathy [28] and F.F. da Silva et al. [11]. [11] An ensemble-based classification method was created by Da Silva and his coworkers. This method makes use of several different classifiers, including SVM, Multinomial Naive Bayes, and Random Forest. [12] They claim that we can achieve superior outcomes by first training the various classifiers on distinct sets of training data and then employing either the weighted average of the probabilities predicted by the various classifiers or the method that yields the most votes, rather than relying solely on a single classifier. In addition, the classifiers are trained with the use of two separate features:

a) Feature Hashing and b) Word Bag, in that order

They practised and evaluated on four individual datasets simultaneously. With one exception, it was shown that the BOW approach performed better than the feature hash algorithm for the majority of the datasets.

One of the key objectives of our research is to demonstrate that employing multiple machine learning classifiers together produces greater results compared to using each of them alone. Even though this study only uses a small dataset and a few features, it also looks at how [11] used feature hashing and lexicon-based features.

Both Bhadane et al. [15] and Apple et al. [16] have found that combining a sentiment lexicon with machine learning methods leads to an increase in the accuracy of their results. When it comes to dealing with the polarity of words, Muhammad et al. [17] tested the SmartSA system and found that their system is superior to baseline lexicons and systems like SVM, NB, etc. with greater F1 scores. This was observed after testing the SmartSA system.

As stated by Addlight and Supreethi [20], KNN is considered to be a superior method for machine learning when compared to SVM [21]. SentiCircles, which were proposed by Saif and He [23], can be utilised to ascertain the context of individual words. They came to the conclusion that more work needs to be done on classifying how people feel.

According to Jianqiang et al. [32], who presented six different preprocessing procedures, it has been demonstrated that rigid preprocessing results in an improvement to the assessment measure. When this is taken into consideration, the preprocessing that we use to filter the tweets turns out to be very effective. Khan and Jeong [21] developed an approach that can be used to determine how customers feel about each component of a product. [Citation needed]

2 PROPOSED WORK

The hybrid model that is being suggested has been given the designation of the three-stage model. In the first stage of this model's preprocessing, multi-aspect based filtration and impurity correction are utilised. At this point, the rules for correct spelling, stemming, expanding abbreviations, and removing stopwords have all been defined. At this level, the messages, the positive and negative aspects, and the tag tokens are all kept distinct from one another. Additionally, handling of negatives is done. The second stage involves the generation of the statistical features that the filtered text possesses. During this stage, the input training and testing sets are turned into the necessary feature set. The final stage of this model applies the hybrid classifier to perform the processing of these attributes in order to predict the user's emotion. For each instance of the thing being studied, a KNN or SVM classifier is chosen based on a probabilistic prediction.

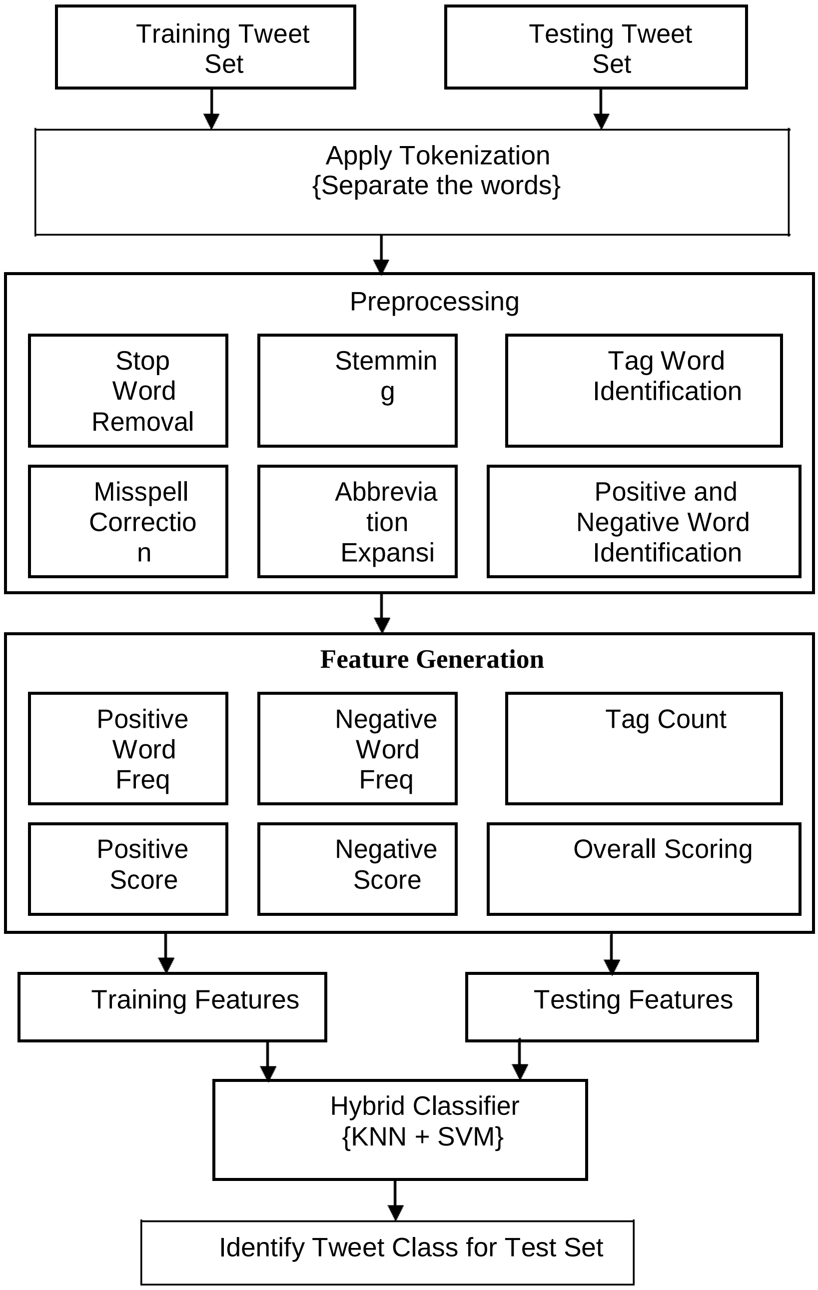

Fig 5.1: Flowchart of the proposed system

The classification is applied on the tweets acquired from the web. The description of dataset is given below in Table 5.1:

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

5.2 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 5.2: 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.

5.2.1 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 5.2: 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.

Data

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 5.3: Showing the results of features generation

There are a number of ways to learn the classifiers, including:

After filtering, the total number of words in each tweet is

The total number of @tags used in a tweet is counted.

This is the total number of negative words in each of your posts.

Each tweet's positive word count:

The sum of the positive values assigned to each of the positive adjectives is known as the "positive score."

Negative score:-The sum of the negative scores of all negative adjectives multiplied together.

Each tweet receives a positive or negative score based on its content.

Class 0 tweets are reserved for those that express displeasure.

1: for unbiased tweets

2: in support of tweets that express gratitude.

5.3 Classification:

Using a hybrid technique that incorporates the prediction probability of both classifiers, we then classify the data using the following algorithm:

Algorithm one:

Classification (TrainingSet, TestingSet)

A training set and a testing set are two separate sets of tweets that are used to generate features.

(

In order to generate features for the training set, you must first create a train features set.

"Generate Features for the Testing Set" is the command that creates the TestFeaturesSet.

3. SWeight = GenerateWeight (TrainFeaturesSet, SVM)

When testing, the KNN is trained using the training feature set, while the SVM is trained using the feature weights generated during processing.

Testing instances are processed in order from i = 1 to TestFeatureSet.Length.

5. K1=Predict (TestFeatureSet (i), TrainFeaturesSet)

KNN Classifier Weight should be taken into account when making predictions on test cases.

6. S1 = Predict (TestFeatureSet (i), SWeight)

SVM Classifier Weight should be taken into account while making predictions for a test case.

(Th1 > K1; S1 > Th1)If this is the case,

In this case, Th1 = 0.5 is selected as the threshold for the prediction probability of the hybrid classifier.

(

8. TestFeatureSet (i).

Class=IdentifyClass (greater (K1, S1))

(

If (K1 > Th1), then

Use a KNN Classifier to Identify Test Classes

(

10. TestFeatureSet (i).

Class = IdentifyClass (K1)

(

11. Else

SVM Classifier is used to identify test classes.

(

13. TestFeatureSet (i).

Class=IdentifyClass (S1)

(

(

TestFeatureSet.Class must be returned.

(

5.4. EXPERIMENTAL RESULTS AND EVALUATION

With the SVM and KNN based hybrid classification model, Twitter characteristics and hidden sentiments may be processed and identified in this study. Netbeans 8.0 and Weka (3.8) are used to implement the project. There are several uses for Weka in data mining, such as preprocessing (such as clustering), classification (such as clustering), and f-measures (such as accuracy and precision). The training and testing datasets, as well as the list of adjectives, abbreviations, and misspelt words, are all stored in MySql databases using the mysqli command line interface.

While weka can be used directly to run classifiers like KNN and SVM on their own, it must be integrated into netbeans and the confusion matrix before the results can be manually calculated. The KNN and SVM-based algorithms are compared individually in the comparative study. The KNN vs. hybrid method comparison is based on a K=15 model.

The description of processing training and testing set is shown in Table5.3:

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 5.4: Confusion matrix for KNN

		Predicted		
	Class	Negative	Neutral	Positive
	Negative	**78**	18	18
Actual	Neutral	8	**47**	16
	Positive	22	15	**79**

Table 5.5: Confusion Matrix for SVM

		Predicted		
	Class	Negative	Neutral	Positive
	Negative	**77**	14	22
Actual	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.

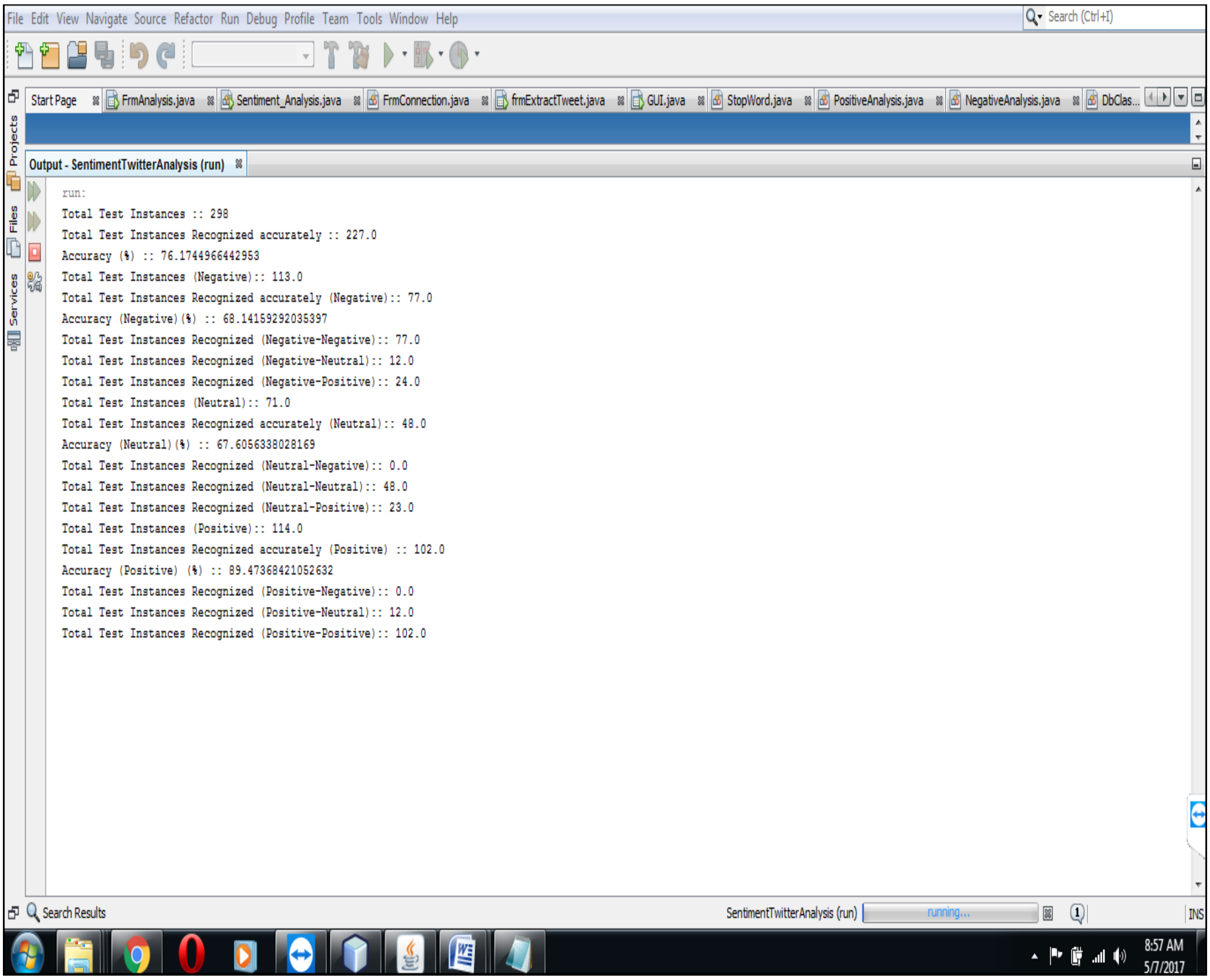

Figure 5.4. Showing the analysis results

Every sentiment category has its own set of true positive (TP), false positive (FP), and true negative (TN) and false negative (FN) outcomes, which can be viewed in the graphs below: As a result, we have the following confusion matrix:

Table 5.6: Confusion Matrix for KNN+SVM

		Predicted		
	Class	Negative	Neutral	Positive
Actual	Negative	**79**	11	25
	Neutral	0	**49**	24
	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.

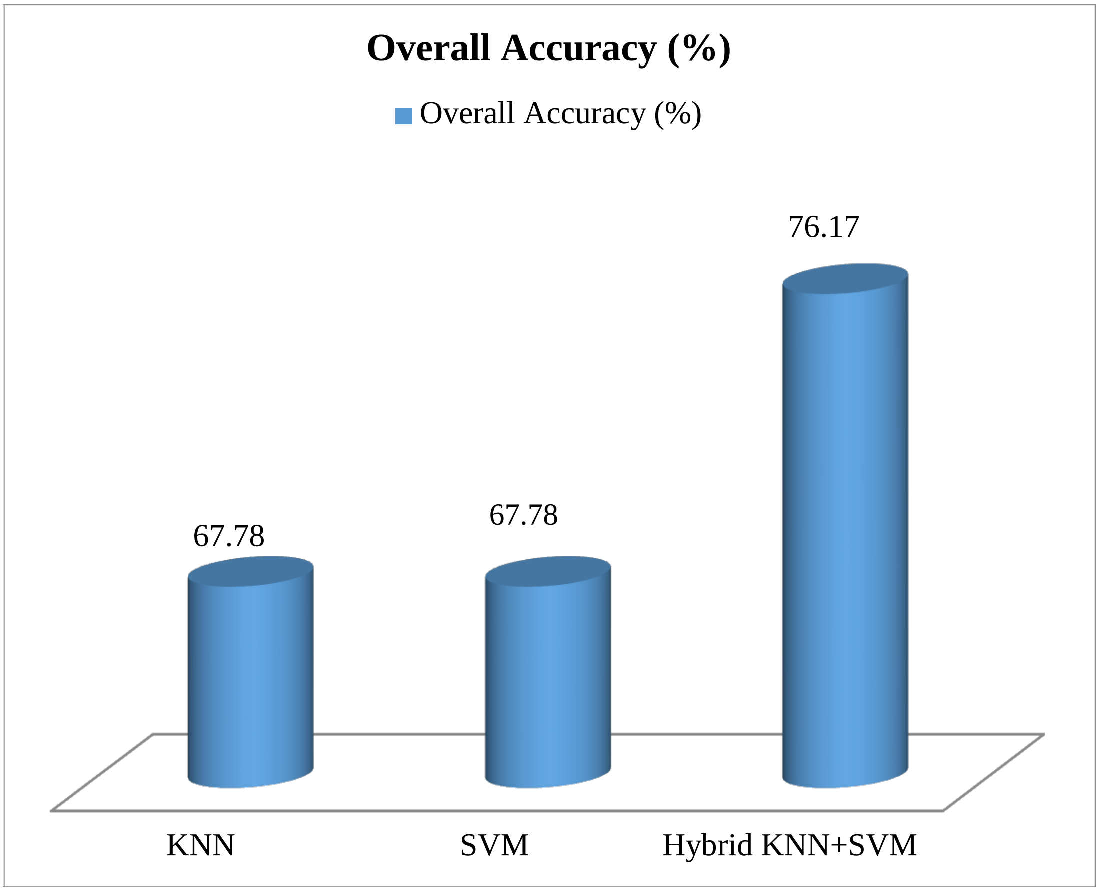

Fig 5.5: showing overall accuracy for 3 approaches

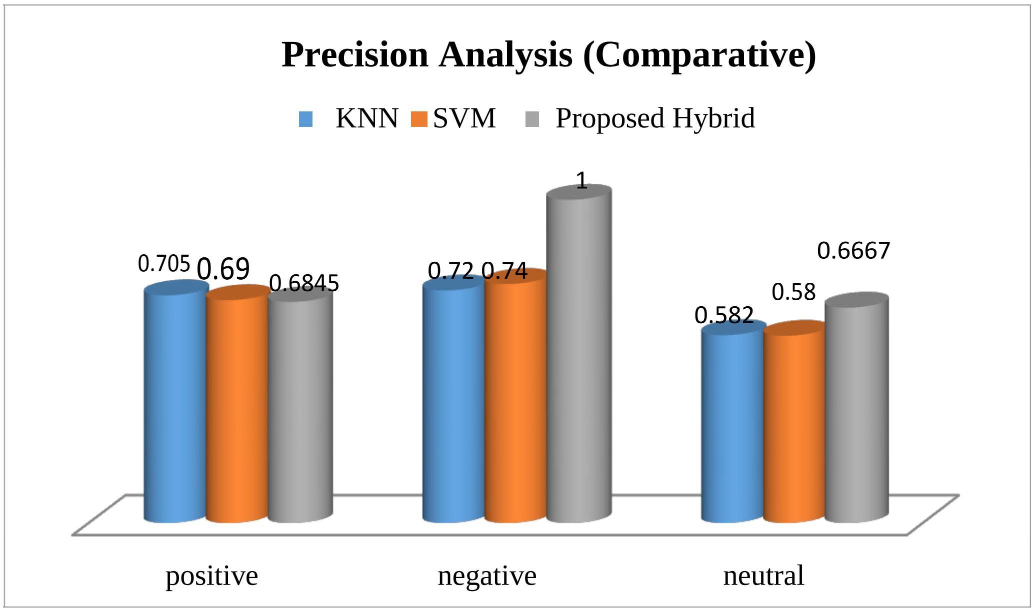

Fig 5.6: showing precision analysis

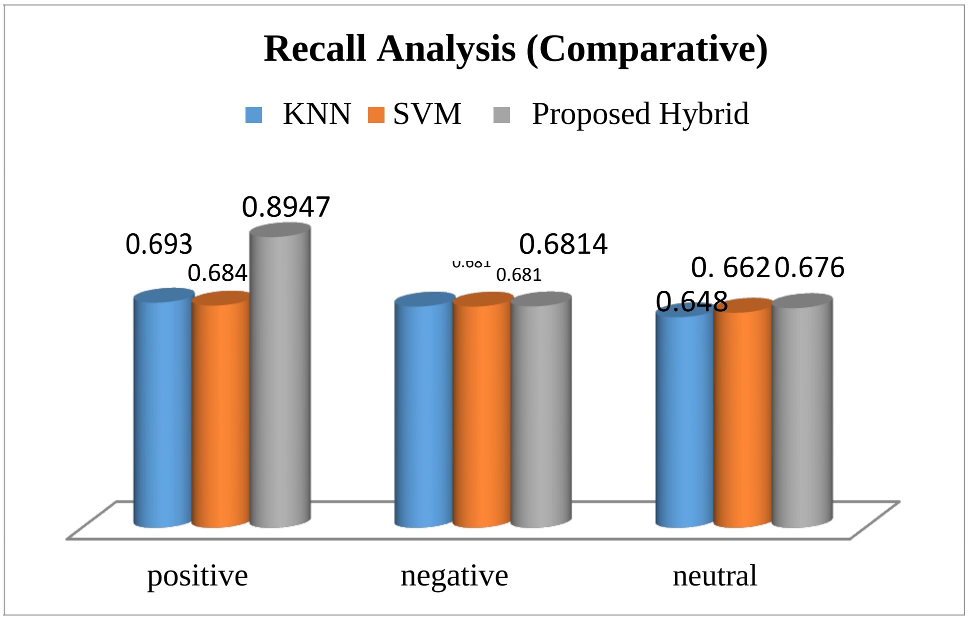

Fig 5.7: showing recall analysis

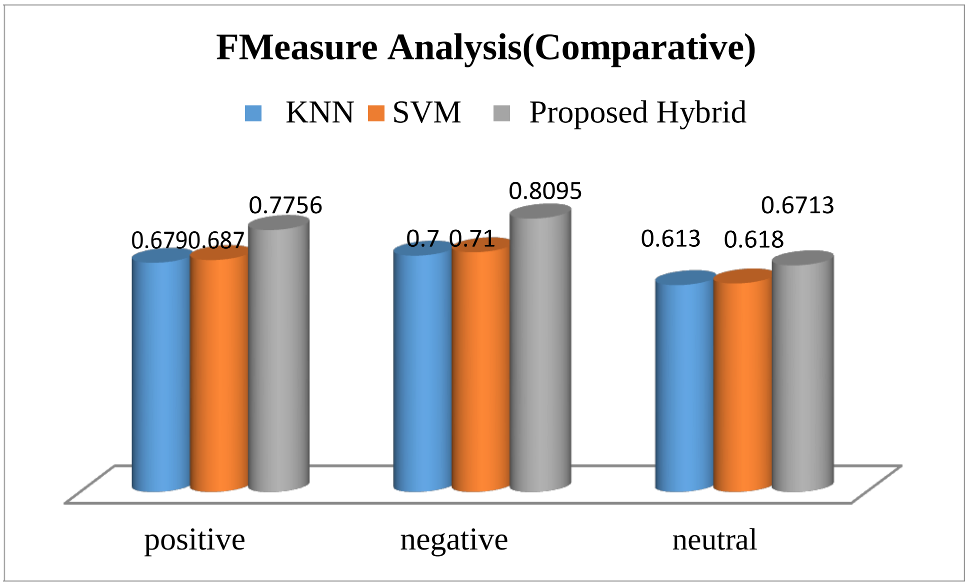

Fig 5.8: showing f-measure analysis

From the data above, you can see that our hybrid strategy beats the competition in terms of both accuracy and f measure.

When compared to [11], our strategy appears to be superior.

Compared to [11], which used a combination of Logistic Regression (LR), Random Forest (RF), and Multinomial Naive Bayes (MNB) along with Feature Hashing and Lexicon features, this research gives similar or even better results.

In Table 5.7, Avg is used in place of Average, Pos is used in place of Positive, and Negative is used in place of Negative.

Table 5.7

Datasets	Avg F-measure(%) (pos,neg)	Avg F-measure(%) (including neutral)	Accuracy(%) (pos,neg)	Accuracy(%) (including,neutral)
OMD[11]	65.36	-	7163	-
Strict OMD[11]	71.81	-	84.56	-
Sanders[11]	76.25	-	86.63	-
Stanford [11]	72.23	-	79.11	-
HCR[11]	61.21	-	78.35	-
Our dataset (Tharindu Munasinge)	**78.28**	**74.22**	**79.81**	**77.27**

5. Conclusion and Scope

As part of this research, a hybrid model that is SVM and KNN based has been presented as a means of improving classification accuracy. The majority of the research that has been done on this subject focuses on a two-way classification of tweets; however, the approach that has been offered classifies tweets according to whether they have a good, negative, or neutral attitude. In relation to the suggested model, preprocessing, the generation of features, and the learning of classifiers have all been finished. Analytical consideration is given to the proposed model in regard to its degree of precision and f-measure. In this investigation, we looked at how the SVM and KNN algorithms compare to one another. The findings indicate that the model that was proposed is superior in terms of its ability to predict tweet class and possesses a lower f-measure.Because the number of features that can be used for training classifiers is restricted by the strategy that is currently being used, we will expand the number of features that we utilise in our future work and employ improved approaches for feature selection such as Information Gain and Chi-Square. Previous studies may lead us to the conclusion that increasing the number of tweets and features contained within our dataset will result in an increase in both accuracy and f-measure. In the not too distant future, additional machine learning approaches could be combined in a similar fashion.

REFERENCES

[1] J. Clement, “Twitter: number of active users 2010-2019,” Statista, Aug. 2019. [Online], Available: https://www.statista.com/statistics/282087/number-of-monthly-activetwitter-users.
[2] E. Kouloumpis, et al., “Twitter Sentiment Analysis: The Good the Bad and the OMG!” in Proceedings of Fifth International AAAI Conference on Weblogs and Social Media, pp. 538-541, 2011.
[3] K. Arun, et al., “Twitter Sentiment Analysis on Demonetization tweets in India Using R language,” International Journal of Computer Engineering in Research Trends, vol. 4, no. 6, pp. 252-258, 2017.
[4] S. A. El Rahman, et al., “Sentiment Analysis of Twitter Data,” International Conference on Computer and Information Sciences (ICCIS), Saudi Arabia, pp. 1-4, 2019.
[5] S. Kumar, et al., “Exploring Impact of Age and Gender on Sentiment Analysis Using Machine Learning,” Electronics, vol. 9, no. 2, pp. 1-14, 2020.
[6] H. Saif, et al., “Semantic sentiment analysis of twitter,” International Semantic Web Conference, pp. 508-524, 2012.
[7] S. Shayaa, et al., “Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges,” in IEEE Access, vol. 6, pp. 37807-37827, 2018.
[8] T. M. Nisar and M. Yeung, “Twitter as a Tool for Forecasting Stock Market Movements: A Short-window Event Study,” The Journal of Finance and Data Science, vol. 4, no. 2, pp. 101-119, 2018.
[9] A. K. Soni, “Multi-Lingual Sentiment Analysis of Twitter data by using classification algorithms,” 2007 Second International Conference on Electrical, Computer and Communication Technologies, pp. 1-5, 2017.
[10] Z. Jianqiang and G. Xiaolin, “Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis,” in IEEE Access, vol. 5, pp. 2870-2879, 2017.
[11] A. Deshwal and S. K. Sharma, “Twitter Sentiment Analysis using Various Classification Algorithms,” 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, pp. 251-257, 2016.
[12] A. Balahur and M. Turchi, “Improving sentiment analysis in twitter using multilingual machine translated data,” Proceedings of the International Conference Recent Advances in Natural Language Processing, pp. 49-55, 2013.
[13] S. K. Mahata, et al., “Analyzing Code-Switching Rules for English–Hindi Code-Mixed Text,” Emerging Technology in Modelling and Graphics, pp. 137-145, 2020.
[14] K. Dashtipour, et al., “Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques,” Cognitive Computation, vol. 8, pp. 757-771, 2016.
[15] I. Mozetič, et al., “Multilingual Twitter Sentiment Classification: The Role of Human Annotators,” PLOSONE, vol. 11, no. 5, 2016.
[16] M. E. Basiri and A. Kabiri, “Uninorm operators for sentence-level score aggregation in sentiment analysis,”4th International Conference on Web Research (ICWR), Tehran, pp. 97-102, 2018.
[17] S. Sahu, S. K. Rout and D. Mohanty, “Twitter Sentiment Analysis - A More Enhanced Way of Classification and Scoring,” 2015 IEEE International Symposium on Nanoelectronic and Information Systems, pp. 67-72, 2015.
[18] L. Brönnimann, “Multilanguage sentiment-analysis of Twitter data on the example of Swiss politicians,” M.Sc. Thesis, University of Applied Sciences Northwestern Switzerland, 2014.