News Classification Using Hybrid Approach Of PSO-KNN

Megha Singla, Brahmaleen K. Sidhu

Abstract: There are different applications which are producing the data in a big way for example various social media platforms. These data need to be analyzed and processed to extract new useful information. This information can also be useful for the decision making process for the organization. In current research there is a news related dataset. This dataset includes various types of news under different categories like technology, entertainment, political etc. various general news are to be categorized into its different categories. These categories correct entry will fine tune the whole system. Later on single sports category news are being categorized into different sports categories like cricket, rugby, football etc. Various classification techniques has been used like SVM, KNN, decision tree etc. over to it new genetic based hybrid approach has been used. This hybrid approach is PSO-KNN. It has been used for classification of the inter news classification and the intra news classification. The results have been compared on different parameters like accuracy, specificity, sensitivity etc. In all the parameters the results have shown improvement over to the SVM, KNN and Decision tree.

Key words: SVM, KNN, PSO, Ngram

I. INTRODUCTION

Text classification is one of the fields where the whole text pertained to the document collected from certain source into multiple classes. This will help in having post features extraction to be comparatively easy with higher success rate. Text classification can be categorization into multiple sub categories.

a. soft classification.
b. hard classification
c. genre classification

In soft classification the whole text is classified into three sub categories out of which one is low, second is medium and third is high category. Low classification includes the text which is having low relationship between the elements. This relationship may also be called as unrelated category to the cause. In medium category the elements are related to the fact with medium level relation. This type of relation is having less strong relationship between the elements. In the third category the relationship is

highly strong. All the elements of the text are related to the requirements analysis usually be drawn on such text. Hard classification is an another type of category in which text of the document will be classified into pre set categories.

The relationship between the elements are sometimes irrelevant because the relationship between the elements are developed based on the requirements and type of the data. Hard classification sometimes leads to the wrong classification.

Genre classification is the one of the most successful type of classification of the text because it classify the text based on the classes of the text like journal, paper, book etc. that helps in successful classification of the documents into its proper category and then further the sub classes are builded based on the soft computing.

1.1 Features selection

There are various techniques and technologies that are required for the features selection.

a. Term Frequency-Inverse Document Frequency (TF-IDF)

For the evaluation of the TF-IDF first TF will be evaluated.

$TF_{ij}=f_{ij}/max_k\ f_{kj}$

F_{ij} is the frequency of word I in document j. $max_k\ f_{kj}$ is the maximum frequency for the k^{th} word in the document.

The calculate the inverse document frequency

as $IDFi=Log_2((N+1)/(N_i+1))+1$

- *Megha Singla, Punjabi University, Patiala, Punjab, India*
- *Brahmaleen K. Sidhu, Punjabi University, Patiala, Punjab, India*

Here N is the total number of training documents, and n_i is the total number of training documents that contain the term i.

$TFIDF_{ij} = TF_{ij} . IDF_i$

Total of the TFIDF is calculated by multiplying the TF and IDF for I term in the document j.

1.2 Classifiers

There are various classifiers which will classify the extracted words into multiple classes.

a. Naïve Bayes classfier: naïve bayes classifier is the way we deal with the number of words. It is assumed by considering that all the words in the document are unique. First step for the classification is first calculate the probability of each class.

$$P(c) = fc/fd$$

Fc is the total number of training documents that contain the label of c. fd is the total number of training documents. So that probability of document in each class is calculated by formula as

$$P(c|d) = P(c) \prod P(xi \in d|c)$$

From this formula the highest probability will be assigned to the class for the document d.

b. Support Vector Machine

It is the multi class classifier. Here the total number of data entities will be distributed on the 3-d plane. Each value from the dataset will be taken and score for each word is marked. Based on the scores the data item will be put into the 3-d plane. This will collect various data entities into multiple classes. Based on the requirements the SVM line will be drawn which will segregate the data items and collect into multiple classes.

c. PSO(Particle swarm optimization)

It is another optimization based technique. It is soft computing based technique where various elements are being put into the multiple layers of the chromosomes. Each chromosome is compared to the threshold value. This process of identification of the optimal element will go on till the final outcome for the optimized element will not be identified. This whole process will classify the data items into two categories. One category is for the elements that are closer to the optimal value. And another category which is far away from the optimized value.

d. KNN(K- Nearest neighbor)

It is another technique where the data items will be having identification based on the identifying the distance. This distance is the Euclidean distance. Sorts these distance in the ascending order of the distances. From the top k number of neighbors will be picked and put into one set. Rest will be put into the other set.

II. LITERATURE SURVEY

Rini Wongso(2017) et. al: author in this paper has worked on the text classification into multiple classes. Each class will be having pre set features. The whole process is done using technique which is the combination of two techniques one is the TF-IDM and SVD. Success rate for the classification using this combined approach is much better compared to the other individual techniques. The whole process of the classification is spanned into various sections or phases. First is input the text dataset, second step includes removing the noise in the text and then after features extraction. In last the classification is performed onto the features. The dataset is taken with Indonesian languages with accuracy of 85%.

Wen Zhang(2011) et. al: in this paper author has worked on the comparison of the TF*IDD and the LSI based techniques. According to study two techniques for the classification for the text is used. The performance of the LSI is better compared to the TF*IDM. The score allocated to the words into the text is not biased using LSI. The scheme of the scores allocations is done by the LSI using merit based process.

Davood Mahmoodi(2011) et. al: author in this paper has proposed SVM based classification technique. It uses the dataset of the Persian based language. They have classified the dataset of the Persian into three categories. Small set is sub divided into the training set and remaining elements are kept as testing set. They have achieved the accuracy of 98.67% for the true classification.

Hao Lin (2014): author in this paper has proposed a classification process for the text being mined using any of the mining technique. For the classification they have used Naïve Bayes based classification. The result generated is much better compared to the SVM. Author in this paper has mainly focused on the efficient way of the classification with the minimum energy lose while classification should be minimized. That technique has to be implemented which is much efficient technique compared to the other technique. so Naïve Bayes is the best technique for the classification of the text.

Krina Vasa (2016): author in this paper has studied the need for the text classification. Text classification is important process as far as current data need is there. There requires various types of classification tools which can classify the text to summarizes the text for various applications like medical diagnosis, sentiment analysis. Researcher has studied various research techniques which are based on machine learning and statistical classification techniques like K-nearest neighbor, Naïve Bayes etc.

III. ALGORITHM

a. Input the text need to be classified.
b. Normalize the text lies in the electronic document based.
c. Extract the features based on TF-IDF based technique.

Term Frequency-Inverse document frequency

It is the most popular scheme. Where each term frequency is calculated in the document.

Like $TF_{ij}=f_{ij}/Max_k\ f_{kj}$

In this the fij is the frequency of the term j in the document i. where the maxk is the frequency for the most common term. It is the kth term.

$IDF_i=\log_2((N+1)/(n_i+1)+1$

In this N is the total number of the text document. ni is the documents count which keep I term.

$TFIDF_i=TG_{ij}.IDF_i$

d. Classify the values for the TFIDF based on optimized criteria. The classification is based on identifying the optimal value from the sub set of values generated against the features set.
e. Performs KNN based classification by calculating the distance of each class from the centroid class.
f. Output the classified classes.

IV. Flowchart

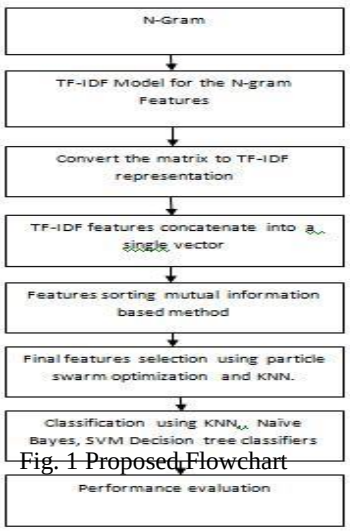

Fig. 1 Proposed Flowchart

V. RESULTS AND DISCUSSIONS

The current research is based on news classification into multiple categories. These categories will be further processed to generate various analysis for the extraction of the results. This whole system follows various steps for the text processing.

5.1 Sample classification

Class	**Number of documents**	**Class**	**Number of documents**
Inter news classification		***Intra news classification***	
Business	510	Athletics	101
Entertainment	386	Cricket	124
Politics	417	Football	256
Sports	511	Rugby	147
Technology	401	Tennis	100

Table 1 Percentage Of Samples Classified Accurately To Class

Current table 1 shows the different samples of the news. Each class will be having different number of news for the classification. Business category has 510, entertainment has 386, politics has 417, sports has 511 and technology has 401. For the intra news there are sports sub category like athletics, Cricket, football, rugby, Tennis etc.

5.2 Accuracy for the Intra news for the PSOKNN and KNN approach

KNN	**PSO_KNN**
89.84238179	100
93.16987741	100
71.80385289	100
91.76882662	100
82.6619965	100

Table 2 PSOKNN and KNN comparison

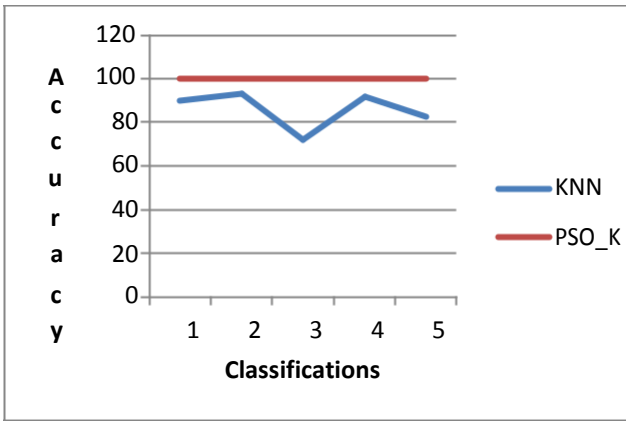

Fig. 2 Accuracy comparison

Fig. 2 shows the accuracy comparison for the KNN and PSOKNN. The hybrid approach has highest accuracy for all the sub categories of sports news. But the KNN has lower accuracy for the few sports sub categories.

5.3 Accuracy for the Intra news for the PSOKNN and SVM approach

SVM	PSO_KNN
100	100
100	100
99.82487	100
100	100
99.82487	100

Table 3 PSOKNN and SVMN comparison

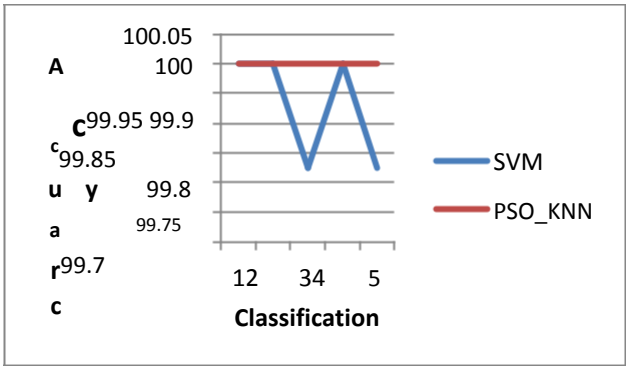

Fig. 3 Accuracy comparison

Fig. 3 shows the accuracy comparison for the SVM and the PSOKNN approach. The hybrid approach has better results compared to the SVM based approach. For the 3rd and 5th category the results for the SVM is lower compared to the hybrid approach.

5.4 Accuracy for the Intra news for the PSOKNN and Decision tree approach

Decision Tree	PSO_KNN
97.54816	100
97.72329	100
95.97198	100
98.42382	100
93.16988	100

Table 4 PSOKNN and Decision tree comparison

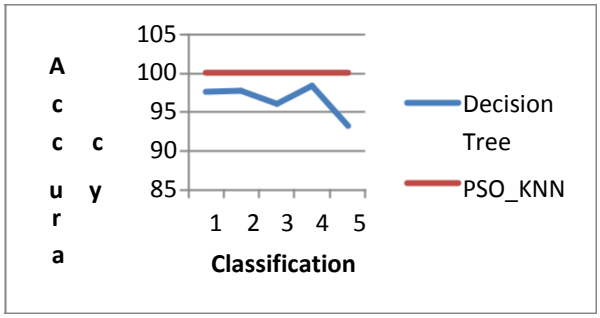

Fig. 4 Comparison for the Decision tree and the PSOKNN

This graph shows the comparison for the decision tree and the PSOKNN. The PSO KNN based approach for the intra news classification is better approach compared to the decision tree. This means the classification has better in all the sub categories.

5.5 Sensitivity comparison

Sensitivity			
SVM	KNN	Decision Tree	PSOKNN
62.85714	0	85.71429	88.14
57.14286	2.380952381	78.57143	85.857
80.55556	100	90.27778	92.779
33.33333	0	88	91
100	0	80.32787	84.4532

Table 5 Comparison for Sensitivity

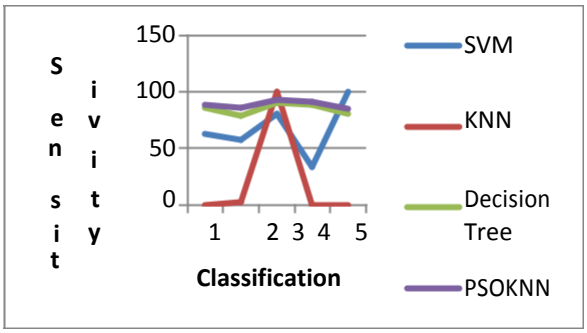

Fig. 5 Sensitivity Comparison

Fig. 5 shows the sensitivity comparison for the different classification technique. These techniques are like SVM, KNN, Decision tree and PSOKNN. The PSOKNN has better results compared to the all the remaining approaches. Hybrid approach has better true positive rate. Lowest and most fluctuating sensitivity is for the KNN.

5.6 Specificity comparison

Specificity			
SVM	KNN	Decision Tree	PSOKNN
100	100	98.4	100
100	100	98.35391	100
99.06103	0.469483568	88.26291	100
100	100	98.57143	100
58.48214	100	97.32143	72.34

Table 6 Specificity Comparison

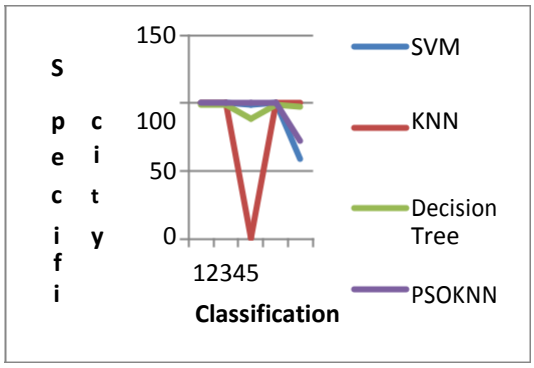

Fig. 6 Specificity Comparison

Graph in the fig. 6 shows the comparison for the specificity for the different approaches. These approaches are like SVM, KNN, Decision Tree, PSOKNN. The PSOKNN has better performance for the true negative identification for the inter news classification.

5.7 Precision Comparison for the inter news

Precision			
SVM	KNN	Decision Tree	PSOKNN
0.882353	0	0.882353	1
0.891892	1	0.891892	1
0.722222	0.253521	0.722222	0.966667
0.956522	0	0.956522	1
0.890909	0	0.890909	0.396104

Table 7 Precision Comparison

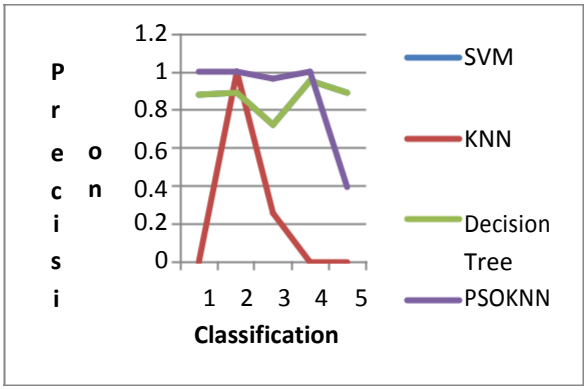

Fig. 7 Precision comparison

Fig. 7 shows the comparison for the different approaches for the inter news classification. The result for the hybrid approach is better compared to the other approaches like decision tree, KNN, and SVM.

5.8 Recall comparison

Recall			
SVM	KNN	Decision Tree	PSOKNN
0.857143	0	0.857143	0.928571
0.785714	0.02381	0.785714	0.971429
0.902778	1	0.902778	0.925556
0.88	0	0.88	0.933333
0.803279	0	0.803279	1

Table 8 Recall Comparison

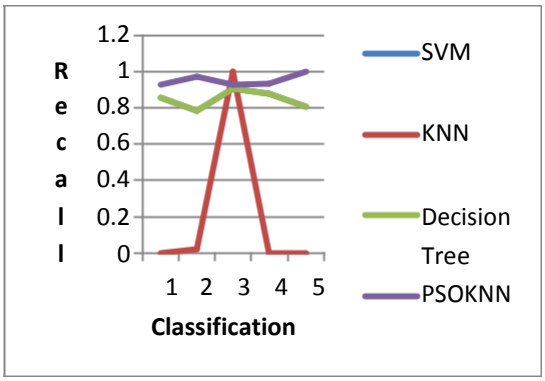

Fig. 8 Recall comparison

Fig. 5.11 shows the comparison for the recall for the different category of the news. These news are for the different categories like sports, entertainment, politics etc. The classes for the different news are done using hybrid approach is better than the existing approaches.

5.9 F-score comparison

Fscore			
SVM	KNN	Decision Tree	PSOKNN
0.869565	0	0.869565	0.97193
0.835443	0.046512	0.835443	0.927273
0.802469	0.404494	0.802469	0.978788
0.916667	0	0.916667	0.9
0.844828	0	0.844828	0.967442

Table 9 F-score Comparison

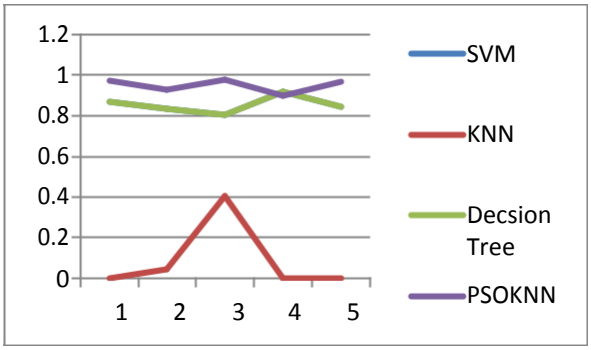

Fig. 9 F-Score comparison

The fig. 9 shows the F-Score comparison for the different approach for the news classification. The Hybrid approach for the PSOKNN has shown the better f-score value for the inter news classification.

VI. CONCLUSION

In current time there are various applications in the electronics world to generates large amount of the data. It will be very difficult to process this large data. These data are so large that even very difficult to store into the single memory. For the better processing there requires processing in terms of the classification of the news. These classes are done in two different scenarios like inter news classification, and the intra news classification. In the inter news classification the news are classified like into five different categories like politics, entertainment, sports etc. In the intra news classification the sports news are classified into different categories like rugby, football, cricket etc. A genetic based hybrid approach has been applied for both inter and intra news classification. The result on different parameters are being compared. These parameters are sensitivity, specificity, accuracy, Recall, precision tec. In all the parameters the results for the hybrid approach is better compared to the various other techniques like SVM, KNN and Decision tree etc.

VII. FUTURE WORK

Currently a hybrid approach for the news classification for the inter and intra news classification has been performed. The results for the hybrid approach is much better compared to the other approaches like SVM,KNN and Decision tree. In future the technique can be applied for the intra news classification for the various other categories other than the sports category.

References

- Rini Wongso, Ferdinand Ariandy Luwinda, Brandon Christian Trisnajaya, Olivia Rusli, Rudy," News Article Text Classification in Indonesian Language", ICCSCI,issue:116, pp:137-143,2017.
- Wen Zhang , Taketoshi Yoshida b , Xijin Tang c," A comparative study of TFIDF, LSI and multi-words for text classification", Expert Systems with Applications,issue 38, pp:2758-2765,2011.
- Davood Mahmoodi1 , Ali Soleimani1 , Hossein Khosravi1 , Mehdi Taghizadeh2," FPGA Simulation of Linear and Nonlinear Support VectorMachine", Journal of Software Engineering and Applications,issue 4,pp:320-328,2011.
- Hao Lin," Research on Energy-Efficient Text Classification", ICITEC, 2014.

- Krina Vasa," Text Classification through Statistical and Machine Learning Methods: A Survey", IJEDR,vol. 4,issue 2, pp:655-658,2016.
- Vangelis Metsis, Ion Androutsopoulos, Georgios Paliouras," Spam Filtering with Naive Bayes – Which Naive Bayes?",issue 27-28, 2006.
- C. C. Aggarwal and C. Zhai, Mining Text Data, 2012.
- M. Kepa, J. Szymanski, "Two stage SVM and kNN text documents classifier," In: Pattern Recognition and Machine Intelligence, Kryszkiewicz M. (Ed.), Lecture Notes in Computer Science, Vol. 9124, pp. 279-289, 2015.
- R. C. Barik and B. Naik, "A Novel Extraction and Classification Technique for Machine Learning using Time Series and Statistical Approach," Computational Intelligence in Data Mining, vol. 3, pp. 217-228, 2015.
- R. Bruni and G. Bianchi, "Effective Classification Using a Small Training Set Based on Discretization and Statistical Analysis," IEEE Trans. Knowl. Data Eng., vol. 27, no. 9, pp. 2349-2361, 2015.
- Chaudhuri, "Modified fuzzy support vector machine for credit approval classification," IOS Press and Authors, vol. 27, no. 2, pp. 189-211, 2014.
- E. Baralis, L. Cagliero, and P. Garza, "EnBay: A novel pattern-based Bayesian classifier," Tkde, vol. 25, no. 12, pp. 2780- 2795, 2013.
- X. Fang, "Inference-Based Naive Bayes: Turning Naive Bayes Cost-Sensitive," vol. 25, no. 10, pp. 2302-2314, 2013.
- H. Wan, L. H. Lee, R. Rajkumar, and D. Isa, "A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine," Expert Syst. Appl., vol. 39, no. 15, pp. 11880-11888, 2012.