INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

(Volume2, Issue4)

Available online at: www.Ijariit.com

Non-Probabilistic K-Nearest Neighbor for Automatic News Classification Model with K-Means Clustering

AKANKSHA GUPTA

Computer Science and Engineering Department, Shaheed Udham Singh College of Engineering & Technology Tangori

gupta.akanksha881@gmail.com

ABSTRACT—The news classification is the branch of text classification or text mining. The researchers have already done a lot of work on the text classification models with different approaches. The news works has to be classified in the form of various categories such as sports, political, technology, business, science, health, regional and many other similar categories. The researchers have already worked with many supervised and unsupervised methods for the purpose of news classification. The supervised models have been found more efficient for the purpose of news classification. The k-means algorithm has been used for the classification of the keywords into the multiple groups. The k-nearest neighbor (kNN) classification algorithm has been utilized to estimate the category of the news in the processing. The proposed model has been recorded with the average accuracy of the 93.28% obtained after averaging the accuracy of all test cases, which higher than the previous best performer naïve bayes and SVM based news classifier, which has posted nearly 83.5% of accuracy for classifying the news data. The proposed model has been tested with the 91%, 95%, 90% and 97% of the accuracy over the input test cases of S1, S2, S3 and S4 respectively, which higher than all of the existing models. Hence the proposed model can be declared as the better solution than the previous classification models.

KEYWORDS—News classification, k-nearest neighbor, k-means classification, support vector machine, N-gram analysis.

I. INTRODUCTION

Data mining is method of discovering the knowledge based data like patterns, associations, changes, anomalies and important structures, from the in-depth knowledge to keep in information, knowledge warehouse or different data repositories. knowledge to the wide accessibility of big quantity of knowledge data in electronic records and at hand would like for turning such data into helpful information and knowledge for broad application as well as marketing research, business management and call support, data processing has attracted a good deal of attention in data trade in recent year.

New classification is method of mechanically classifying the news knowledge into the varied classes on the idea of knowledge patterns, associations, changes, anomalies and important structures, from the knowledge data to classify and store the news information, where the data has been collected from the different information repositories containing the news data. knowledge to the wide accessibility of big quantity of knowledge data in online news records and would like for turning such data into helpful information and knowledge for broad application as well as marketing research, business management and call support, data processing has attracted a good deal of attention in data trade in recent year.

News classification is that the method of assignment text documents to 1 or a lot of predefined classes. this permits users to seek out desired data quicker by looking solely the relevant classes and not the complete data house. The importance of text classification is even a lot of apparent once the knowledge house is big like the globe Wide net. samples of net classification systems embrace Yahoo! directory and Google net directory. However, such classification services square measure administered by human specialists, and that they don't proportion well with the expansion rate of web content on the net. To modify the classification method, machine learning ways are introduced. in a very text classification technique supported machine learning, classifiers square measure designed (trained) with a collection of coaching documents. The trained classifiers will so assign documents to their appropriate classes.

Online news articles represent a kind of net data that square measure often documented. Currently, on-line news square measure provided by several dedicated newswires like one Reuters and PR Newswires. it'll be helpful to collect news from these sources and classify them consequently for ease reference. In this paper, we tend to describe a operating news arrangement, named classifier, that performs automatic on-line news classification. The classification model adopts SVM classification technique to classify news articles into classes. These classes are often either a collection of predefined classes, i.e., general classes, or special classes outlined by users themselves. The latter are referred to as the customized classes. With customized classes, The classification model permits users to quickly find the specified news articles with minimum effort.

II. LITERATURE REVIEW

Li, Jinyan et. al. [1] have proposed the hierarchical classification in text mining for sentiment analysis of online news. In this paper, the authors have evaluated several popular classification algorithms, along with three filtering schemes. The filtering schemes progressively shrink the original dataset with respect to the contextual polarity and frequent terms of a document. The proposed approach is called "hierarchical classification". The effects of the approach in different combination of classification algorithms and filtering schemes are discussed over three sets of controversial online news articles where binary and multi-class classifications are applied.

Prollochs, Nicolas et. al. [2] has worked on the sweetening of sentiment analysis of economic news by detective work negation scopes. To predict the corresponding negation scope, connected literature usually utilizes 2 approaches, namely, rule-based algorithms and machine learning. However, a radical comparison is missing, particularly for the sentiment analysis of economic news. To fill this gap, this paper uses German accidental announcements as a standard example of economic news so as to pursue a two-sided analysis. First, we have a tendency to compare the prognostic performance employing a manually-labeled dataset. Second, we have a tendency to examine however detective work negation scopes will improve the accuracy of sentiment analysis.

Cui, Limeng et. al. [3] has developed a hierarchy methodology supported LDA and svm for news classification. In this paper the authors have centered on news text classification that is purposeful for data supplier to arrange and show the news however conjointly for the users to succeed in the dear data simply. A hierarchy methodology supported LDA and SVM is projected to accomplish this task and a number of other experiments are conducted to gauge the projected methodology. The results show that the projected methodology is promising in text classification issues.

Ouyang, Yuanxin et. al. [4] has projected the news title classification with support from auxiliary long texts. In this paper, the authors have aimed at the reports title classification that is a vital associate for the evaluation of typical member in extracted data and propose an approach that employs external data from long text to handle the text categorization. Afterwards, the Restricted Boltzman Machine are utilized to pick out options and so finally perform classification algorithm known as Support Vector Machine.

III. EXPERIMENTAL DESIGN

In this paper, the proposed model has been designed for the automatic news classification over the online news portals. The news sources, which collects the data from the variety of the APIs, cannot classify all of the data manually, hence requires the automatic module for the news classification based upon the knowledge based method. In the proposed model, the news ranking based method has been proposed to classify and rank the news by calculating the similarity between the keyword data collected from the news body and the knowledge data. The news classification decision is taken on the basis of the structural similarity and the density of the matching keywords. Due to numbers of calculation taken between the test sample and all the training samples, the traditional method of Ranking has less computational complexity. To overcome the complexity, this paper introduced combination Ranking algorithm with a clustering method.

Algorithm 1: Ranking Index algorithm

1. Input news data
2. Fetch keyword array from news data
3. Match keyword array with all keyword lists of the given category
4. Build Ranking index with the pre-defined rank values in the category specific ranking arrays
5. Repeat the step 3 and 4 for all categories

In the second step of the news classification engine, the weights are calculated on the basis of the keyword terms data extraction from the news text by using the keyword extraction method based upon the N-gram text analysis. In the further step, the every category is clustered by using the K-means algorithm, which extracts the clusters on the basis of inter-similarity between the term entities. The algorithm steps have been elaborated in the following steps:

Algorithm 2: Weighted k-means clustering algorithm

a. Initialize the value of K as the number of clusters of document to be created.
b. Get the centroid data from pre-defined centroid set
c. Assign each object to the group that has be closest centroid
d. Update the centroid by calculating the average value of the existing data on the cluster; $Ci = 1\ n\ dj\ n\ j$ =1 (5) Ci : centroid to-i from the cluster n : number of documents in a cluster dj : document vector to-j
e. If it is not the last iteration
    i. Repeat the iteration
    ii. GOTO step a
f. Otherwise
    i. Return the indexed matrix

This produces a separation of the objects into groups from which the metric to be minimized can be calculated. After clustering for each category, the cluster centers were chosen to represent the category and they become the new training sets for KNN algorithm. By this method, the number of samples for training is reduced efficiently, so the time for calculating similarities in KNN algorithm also reduced.

Algorithm 3: k-NN (k-Nearest Neighbor) Clustering

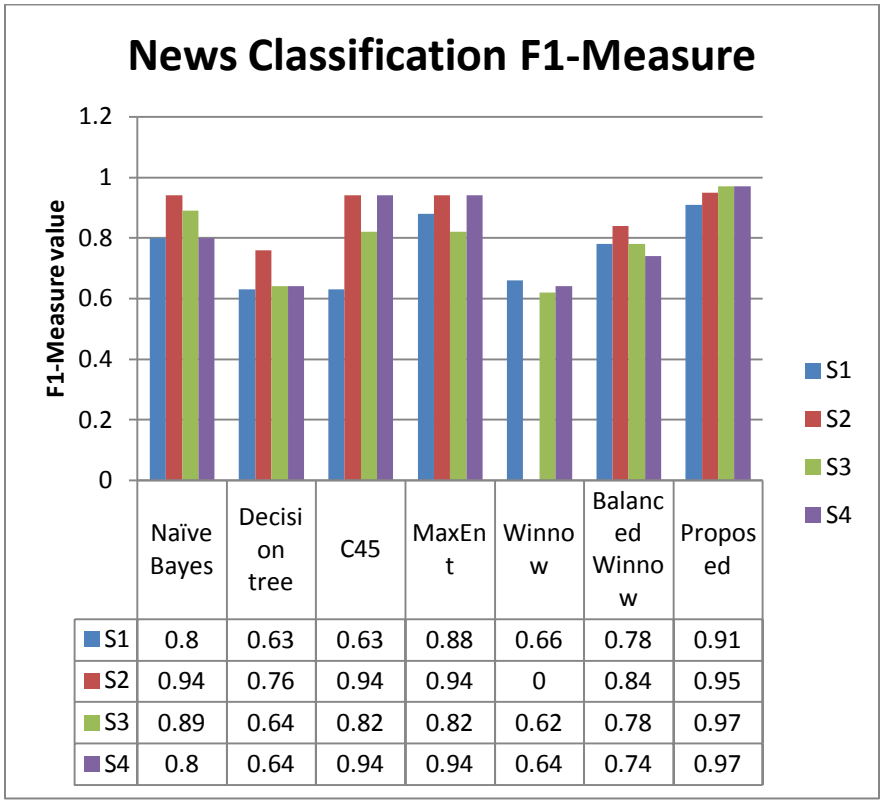
1. Input dataset with news rank values
2. K is a pre-defined number of clusters
3. Algorithm determines the pre-defined data points equal to the cluster number
4. The algorithm evaluates the distance of each data point from all of the pre-defined initial data points.
5. The point is added to the cluster with the lowest distance
6. Evaluate all points with method from 4 to 5 until last point.
7. Return the clustered (classified) data

RESULT ANALYSIS

The results of the proposed model have been collected under the various scenarios. The multiple test cases have been utilized for the in-depth analysis of the proposed model in classifying the news data. The statistical performance parameters of F1-measure and average prediction accuracy have been used for the purpose of news classification.

4.1. Result Evaluation based on F1-Measure

There are twenty news entries, which are permanently stored in the MySQL database, for the retrieval and classification of the given news data. The 20 test cases has been used for the evaluation of the proposed model for the measurement of the errors and the accuracy in the classification.



	Naïve Bayes	Decision tree	C45	MaxEnt	Winnow	Balanced Winnow	Proposed
S1	0.8	0.63	0.63	0.88	0.66	0.78	0.91
S2	0.94	0.76	0.94	0.94	0	0.84	0.95
S3	0.89	0.64	0.82	0.82	0.62	0.78	0.97
S4	0.8	0.64	0.94	0.94	0.64	0.74	0.97

Figure 4.1: Evaluation of the proposed model based upon the F1-Measure

The above figure (4.1) defines the performance evaluation on the basis of the F1-measure over the 20 test cases. The proposed news classification model has been recorded with the improved performance in comparison with the existing model. The above figure 4.1 clearly justifies the robustness of the proposed model.

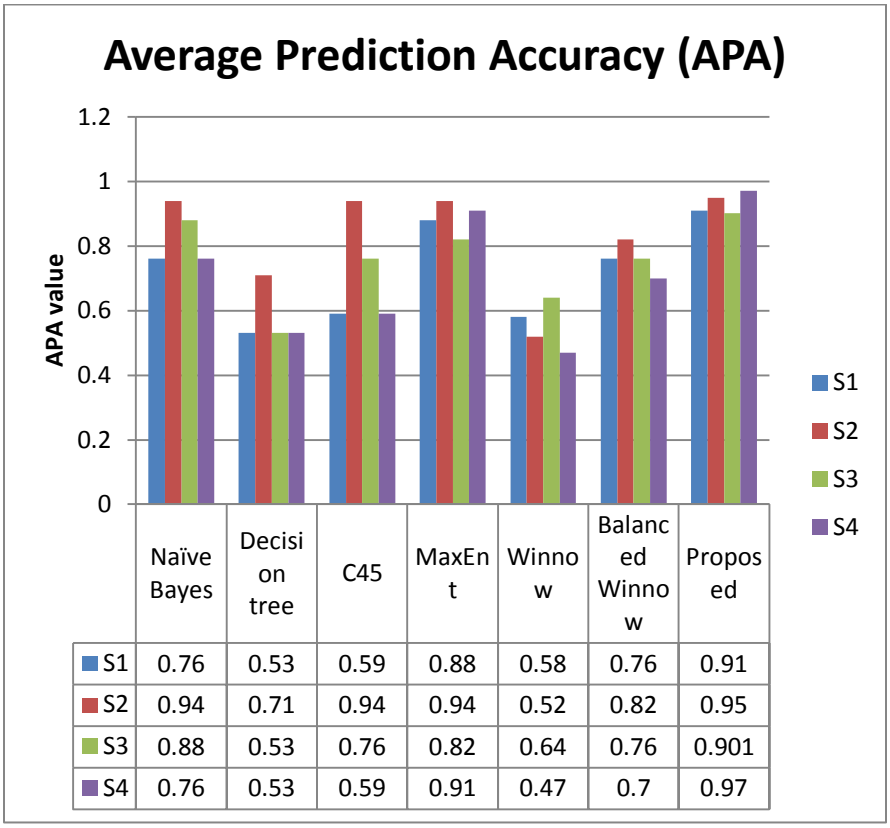
Test Cases	Naïve Bayes	Decision tree	C45	MaxEnt	Winnow	Balanced Winnow	Proposed
S1	0.81	0.64	0.64	0.89	0.65	0.78	0.93
S2	0.93	0.75	0.95	0.93	0	0.84	0.95
S3	0.86	0.64	0.84	0.83	0.64	0.77	0.97
S4	0.80	0.65	0.94	0.95	0.63	0.73	0.971

Table 4.1: The table containing the obtained values of F1-measure from the experiment

The above table shows the performance of the proposed model based upon the parameter of F1-measure. The proposed model has been evaluated over the various kinds of the pre-processing models. The different kinds of pre-processing models has been used, which utilizes the different levels of the keyword filtering such as stop words, high frequency keywords, unique keywords based upon the high frequency, etc. The proposed model results have been obtained in the form of F1-measure from the proposed model. The proposed model has clearly outperformed all of the existing combinations as per shown in the table 4.1.

4.2. Performance evaluation based upon APA

The average prediction accuracy (APA) has been calculated from the experiments conducted over the proposed model. The twenty news entries have been used for the experiments, from where the statistical type 1 and type 2 parameters are calculated, which leads towards the calculation of the statistical measures. The APA is one of the primary parameters for the evaluation of the proposed model for the evaluation of its comparison with the existing models of automatic news classification.



	Naïve Bayes	Decision tree	C45	MaxEnt	Winnow	Balanced Winnow	Proposed
S1	0.76	0.53	0.59	0.88	0.58	0.76	0.91
S2	0.94	0.71	0.94	0.94	0.52	0.82	0.95
S3	0.88	0.53	0.76	0.82	0.64	0.76	0.901
S4	0.76	0.53	0.59	0.91	0.47	0.7	0.97

Figure 4.2: The proposed model evaluation based upon the average prediction accuracy (APA)

The above figure 4.2 shows the results obtained from the news classification system developed under the proposed model. The proposed model results have obtained in the form of multiple classes of the news data processed by using the variety of pre-processing methods as per defined in the latter section. The proposed model results have clearly outperformed the existing model by approximately 5% or higher than the existing news classification models.

Test Cases	Naïve Bayes	Decision tree	C45	MaxEnt	Winnow	Balanced Winnow	Proposed
S1	0.75	0.53	0.59	0.88	0.58	0.76	0.913
S2	0.95	0.71	0.94	0.94	0.52	0.82	0.94
S3	0.88	0.53	0.76	0.82	0.64	0.76	0.90
S4	0.75	0.53	0.59	0.91	0.47	0.7	0.98
Average	0.84	0.58	0.72	0.8875	0.5525	0.76	0.94

Table 4.2: The comparative analysis based upon the APA

The proposed model results have been listed in the table 4.2 with all of the existing models presented in the proposed model. The results have clearly indicated that the proposed model has outperformed the existing models of automatic news classification.

CONCLUSION

The proposed model has been tested with the locally stored news database of first type because of its stability and pretype information. The pretype information contains the manually classified news data type stored in the database under the column pretype, which stays helpful in the classification accuracy measures. A number of experiments have been conducted over the proposed model by using the various forms of the input data generated after various levels of pre-processing. The proposed model has been tested for the various

performance measures which includes the precision, recall, average prediction accuracy and F1-measures. All of the above performance measures has been obtained after the estimation of the statistical type 1 and type 2 errors over the input data. The proposed model has been found accurate higher than 90% in all of the rounds if the true negative cases are also being analyzed. The proposed model has been recorded with the average accuracy over all of the test cases nearly at 93% which is better all of the other models used under the existing model. The proposed model has outperformed all of the existing models designed with the different filters over the differently processed datasets.

REFERENCES

[1] Byun, Hyeran, and Seong-Whan Lee. "Applications of support vector machines for pattern recognition: A survey." In *Pattern recognition with support vector machines*, pp. 213-236. Springer Berlin Heidelberg, 2002.

[2] Cui, Limeng, Fan Meng, Yong Shi, Minqiang Li, and An Liu. "A Hierarchy Method Based on LDA and SVM for News Classification." In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*, pp. 60-64. IEEE, 2014.

[3] Korde, Vandana, and C. Namrata Mahender. "Text classification and classifiers: A survey." *International Journal of Artificial Intelligence & Applications* 3, no. 2 (2012): 85.

[4] Krishnalal, G., S. Babu Rengarajan, and K. G. Srinivasagan. "A new text mining approach based on HMM-SVM for web news classification."*International Journal of Computer Applications* 1, no. 19 (2010): 98-104.

[5] Li, Jinyan, Simon Fong, Yan Zhuang, and Richard Khoury. "Hierarchical classification in text mining for sentiment analysis of online news." *Soft Computing* (2015): 1-10.

[6] Lu, Lie, Hong-Jiang Zhang, and Stan Z. Li. "Content-based audio classification and segmentation by using support vector machines."*Multimedia systems* 8, no. 6 (2003): 482-492.

[7] Morariu, Daniel, L. Vintan, and Volker Tresp. "Feature Selection Methods for an Improved SVM Classifier." In *Proceedings of 14th International Conference on of Intelligent Systems (ICIS06), ISSN*, pp. 1305-5313. 2006.

[8] Ouyang, Yuanxin, Yao Huangfu, Hao Sheng, and Zhang Xiong. "News Title Classification with Support from Auxiliary Long Texts." In *Neural Information Processing*, pp. 581-588. Springer International Publishing, 2014.

[9] Pröllochs, Nicolas, Stefan Feuerriegel, and Dirk Neumann. "Generating Domain-Specific Dictionaries Using Bayesian Learning." *Available at SSRN 2522884* (2014).

