International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 3 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT206380

Segregation of Live News Articles Based on Location Using Machine Learning

Heneil Tayade, Chaitanya Shetty, Ratika Jankar, Dr. Amol Pande
Department of Computer Engineering, Datta Meghe College of Engineering, Navi Mumbai, Maharashtra, India

ABSTRACT

As we all know web contains an enormous amount of data which is gigantic and it is changing continuously for each minute and we also know during this hectic lifestyle it's very difficult to stay track of each news and article that's occurring. So, people are mostly focused on the news which goes into their nearby environment. During this paper, we consider displaying the news directing on the nearby cities and also displaying the required news articles supported by a few important cities. Here, we've a web crawler which is used to withdraw the content from the HTML pages of the articles. Random forest, Naïve Bayes and SVM classifiers are used to compute the precision and their accuracy is being calculated. Machine Learning is the well- known technique used for this type of news classification and displaying of the news articles

Keywords : Machine Learning, Random Forest, SVM, Naive Bayes, News classification

I. INTRODUCTION

The large data sets of story articles are spread across the planet at a very constant rate. It's difficult to track stories that are displayed every day for an individual because of different and restless lifestyle. So, people are mainly inquisitive about online news articles and because their interest might change from individual to individual the information displayed should even be changed as the requirement and interest of individual changes. People are generally interested in the news which goes on in their instantaneous surroundings and environment. The data got can be of less importance to an individual. So, the withdrawal of expected and also the applicable information for a specific person could be an important task to be done. This can be drained mainly by concentrating an issue of interest for a specific person per that city or country or mainly to which place that person currently belongs to or is in. During this case, we mainly target the geographical domain per the interest of a specific person or individual. So, if we consider an example where an individual wants to read news of Mumbai but if the news displayed is going to be of other state or city then it'll be an unorganized task to fetch a piece of specific news. So, during this case, we are concentrating on displaying the news per that exact city on which an individual is interested. We have used Machine Learning techniques to extract the news articles that support the various classifiers and their classification [6]. The placement which we would like is of the identical city, state or the country but here we are mainly specializing in extract the news of a specific city. Therefore, the news articles from the various websites of 3 different newspapers are taken to extract the desired information [1]. The three different newspapers are Hindustan Times, Times of India and Indian Express. The fundamental form of the net page is the HTML Language. It contains various elements like comment section, bars, advertisement, news basis, etc. The text classification related here is got down with very little correctness and reliability. So, to form a news article with proper

display of data and also the related data to the user [12]. The matter solving techniques should be added on to get rid of the efficiency and to extend the accuracy of processing the desired news articles and also the required classification of stories supported the town. The processing is done by creating our web crawler to withdraw the desired news and also the required webpage of the most heading of theArticle [2]. The processing requires different operation like tokenizing the text and stemming. After these tokens are finished processing we remove the stop words once the stemming is finished. Eventually, classification is finished and also the trained classifiers would predict the output precision. Random forest, Naïve Bayes and SVM classifiers are used for the classification part.

II. LITERATURE SURVEY

i) Survey report on text classification with different term weighting methods and comparison between classification algorithms.

Author: Anuradha Patra Barkatulallah University Institute of Technology Barkatulaah University Bhopal, MP, India

In this paper supervised and unsupervised algorithms used are K-NN, SVM, neural network, decision tree. These algorithms are compared with each other. For text classification first preprocessing of data is done. Here a list of stop words to be removed is created. After this, a set of words produced by word extraction is then scanned so that every word appearing in the stop list is removed. Stemming is used to reduce a word to its stem or root from. It is used widely in information retrieval tasks to increase the recall rate and give most relevant results such as: happy-> happi. After that indexing is performed [11]. Then term weighting method is performed which classifies the different algorithm like k-nn, naïve Bayes, SVM, neural network, decision tree. This paper survey is on text classification. This survey focuses on the existing literature and explored the documents representation and the analysis of feature selection methods and classification algorithms. Term weighting is one of the most vital parts for construct a text classifier. After studying this paper, we analyzed that no single representation scheme and classifier can be mentioned as a general model for any application. Different algorithms perform differently depending on data collection [13]. We can also use such method which is use in this paper like preprocessing.

ii) Comparison of text classifiers on news articles
Author: Lilima Pradhan, Neha Ayushi Taneja Charu Dixit, Monikasuhag

Text classification is used to classify documents on the basis of their content. The documents are assigned to one or more categories manually or with the help of classifying algorithms. In this paper Naive Bayes, SVM, Decision tree are used. In our paper we can use the random forest algorithm because by researching various papers, we analyzed that decision tree has some problems like optimal decision tree is NP-complete, tree splitting is locally greedy, tree structure prone to sampling and we analyzed that random forest algorithm can overcome these problems of decision tree [3]. Random forest comprises of different single trees each dependent on an irregular example of the preparation information. They are normally more precise than single decision trees. Here the trees are unpruned. While a solitary Decision tree like CART is frequently pruned, a random forest tree is completely developed and unpruned. In this paper k nearest algorithm is used but after studying we analyzed that K-NN classifier is a supervised lazy classifier which has local heuristics. Being a lazy classifier, it is difficult to use this for prediction in real time. Naive Bayes is an easier learning classifier and it is much faster than K-NN. Thus, it could be used for prediction in real time. It

takes a probabilistic estimation route and generates probabilities for each class. It assumes conditional independence between the features and uses a maximum likelihood hypothesis [3]. After observing this paper and studying the result, SVM gives the highest accuracy for classifying the news articles. Thus, we are updating the feature, improving the system and building application based on location.

iii) News classification and its technique: A review

Author: Mazhar Iqbal Rana x, Shehzad Khalid y, Muhammad Usman Akbarz, Department of computer Engineering, Bahria University Islamabad.

In this paper different steps are used for classifying the news. These steps are data collection, preprocessing, feature selection, classification techniques application and evaluating performance measures. In this paper SVM, decision tree, K-NN algorithms are used. After studying these algorithms, we analyzed that SVM is a best algorithm and have a better accuracy and we can apply more significant algorithm in our system [3]. Naïve bayes, random forest gives pretty accurate results too. Review of news classification is bestowed in this paper. We studied this paper and applied the feature which is required and best for text classification.

iv) Multi-Label Topic Classification of News Articles

Author: Zach CHASE Nicolas GENAIN Orren KARNILearning

We consider the problem of learning to classify the topic of news articles for which there are multiple relevant topic labels. We analyze the shortcomings of a number of algorithmic approaches, including Naive Bayes, and develop two alternate approaches to solving the problem [4].We assess the performance of a binary classifier approach in which we learn a set of one-versus-all naive bayes binary classifiers, one for each label class. We also developed and analyzed the performance of two novel algorithms derived from the popular tf-idf weighting scheme, the tf-idf approach captures some interesting aspects of the intuition behind how people may classify news articles, but we were not able to lower the error produced by the tf-idf model sufficiently to make it practically competitive with the binary classification scheme [6]. Future research might look into alternate methods for scoring functions based on tf-idf and the notion of finding high information words to classify multi-label articles.

III. BACKGROUND

Support vector machines is an algorithm that determines the best decision boundary between vectors that belong to a given group (or category) and vectors that do not belong to it [6]. That's it. It can be applied to any kind of vectors which encode any kind of data. This means that in order to leverage the power of svm text classification, texts have to be transformed into vectors. Since a Naive Bayes text classifier is based on the Bayes's Theorem, which helps us compute the conditional probabilities of occurrence of two events based on the probabilities of occurrence of each individual event, encoding those probabilities is extremely useful [4]. Random Forest is used because of its algorithmic simplicity and prominent classification performance for high dimensional data [3]. It has become a promising method for text categorization. Random forest is a popular classification method which is an ensemble of a set of classification trees.

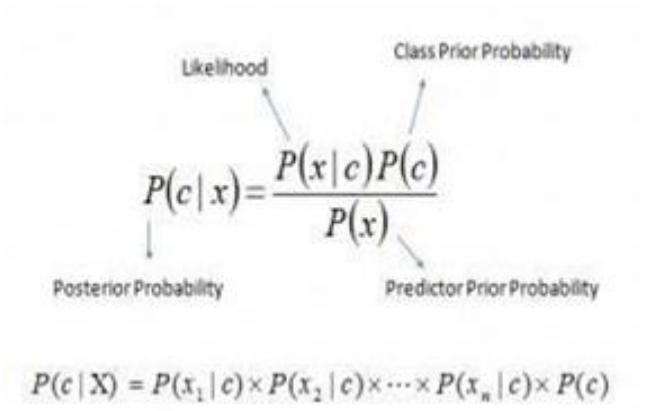
A. Random ForestClassifier

Random Forest is a supervised classification algorithm that was first introduced to us by Leo Breiman. The random forest combines hundreds of number of decision trees, trains each tree on a slightly different set of the observations, splitting nodes in each tree considering a limited number of

the features [3]. The final predictions of the random forest are made by averaging the predictions of each tree. Random forest classifiers use bootstrap aggregating techniques which over and over chooses a random sample of the preparation set with substitution and utilization this example to cause trees to learn since bootstrap methodology reduces the distinction and accordingly prompts better execution. Random forest classifier maintains good accuracy even a large proportion of the data is missing [3].

B. Naive Bayes

Naïve Bayesian classifier calculates the posterior probability. It is based on Bayes theorem. They are the statistical classifiers. Naïve Bayesian classification is referred to as naïve because it assumes that each of its inputs is independent of each other, an assumption that rarely holds true [4].



• P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
• P(c) is the prior probability of class.
• P(x|c)is the likelihoo which is the probability of predictor given class.
• P(x) is the prior probability of predictor.

C. SVM Classifier

A support vector machine is a supervised learning algorithm that sorts data into two categories. SVM also known as support vector network [6]. It is trained with a series of data already classified into two categories, building the model as it is initially trained An SVM yield a guide of the arranged information with the edges between the two as far separated as could be expected under the circumstances. SVM is based on the idea of finding a hyperplane that best divides a dataset into two classes or a group [9]. SVM utilises a lot of mathematical function that is defined as a kernel. The main function of the kernel is to accept the information or data as input and transform it into a required form.

IV. METHODOLOGY

The goal of our approach is to assign an output class to the news articles based on the content. The process starts with the Data Retrieval module which is the collection of our dataset wherein our self-developed web scraping algorithm is employed to extract the actual text from the webpage. Training and Testing is done on the dataset. The Training done is 70% while the Testing done is 30% respectively. Data Preprocessing methods are applied before training the data which is used as an input to the classifier for training. The same data preprocessing methods are used for testing data and this acts as an input to the trained classifier which predicts the output class of the test news articles. This is followed by the evaluation of the accuracy of the trained classifier based on some performance metrics. Following subsections provide further analysis of each of these components.

A. Data Retrieval

This is the primary phase of the method where the news articles are collected from various sources of internet sites. It contains three different steps where the primary step is Parsing the RSS feeds to induce the URL's during this phase the URL's from different websites are been retrieved where the info is collected from these websites. The second step is to gather the URL's in a file where it can be used further for processing so this can be saved and filed for future use. The last is to extract the articles using URL's here

fetching of knowledge is completed from various websites through these URL [6-9]. Each news articles are extracted by visiting every URL's and their websites. Web crawler helps in extracting and displaying the news through HTML page for various newspapers like Hindustan Times, Times of India and Indian Express.

B. Text pre-processing

Text processing contains four different steps. It involves the pre-processing of the extracted data. During this phase first the tokenizing of the fetched data is completed here the string of characters is divided as needed only the desired type of string is saved and also the unwanted string of characters is removed. Stemming is the next step where the words within the articles are reduced to their declension form. This is followed by stop word removal, the stop words are those which are of no importance in fetching and reading the desired data so such articles are discarded from the collected data.

C. Training a classifier

Training a classifier first involves the pre- processing module which extracts relevant a part of the article. This step is important to boost the accuracy of the classifier. Input to the classifier is the training set and therefore the set of labels corresponding to it. The processed news articles are labeled numerically supported town tag which is pre-decided. Since the input to the classifier is two vectors, the set of stories articles and the labels must be vectorized. After vectorization of these two entities, they act as an input to the classifier. Only 70% of the dataset is employed for training the classifier, the rest is utilized for testing. Since training is performed just once, the classifier object which is trained is stored during a pickle file. Pickling is finished to serialize the item and storing it into the disk for the testing phase. A similar process is applied for dumping the Count vectorizer object for further use.

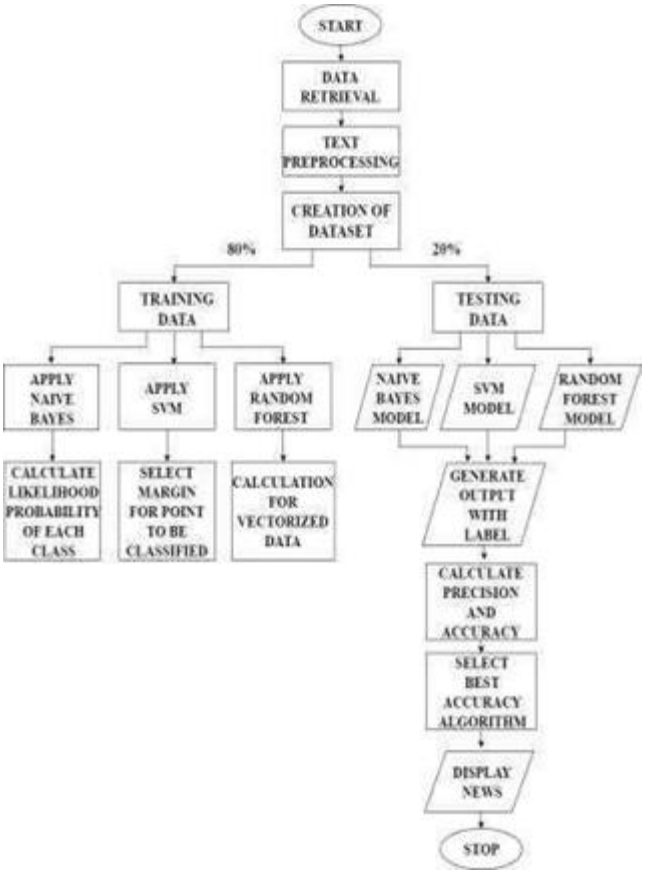


Figure 1. Flowchart of the process

D. Testing a classifier

The pickle file containing the stored classifier and therefore the Count victimizer object is loaded. Since the stored classifier is trained, testing data is fed into it. The classifier predicts the class, during this case-city, of the corresponding article the classification as performed by the trained classifier acts as a basis for determining the accuracy of the model. After testing, the accuracy of the classifier is noted supported various performance metrics like Precision, Recall, and F1 score.

V. RESULT

The final outcome gives the precision, recall and F1 score. The score is considered because the primary performance measure where the higher performance of various algorithms is noted. It's mainly divided into precision and Recall

A. Precision:

It is the ratio of the number of articles that are truly positive that's which are correct to the overall number of articles classified and predicted having a selected category. It's mainly classified as the positive and negative value called as positive predictive value [9]. It may be defined with the following formula:

P=m / (m+n)

here, m stands for True positive and n stands for false positive.

B. Recall:

Recall It is the ratio of the number of articles that are true positive to the number of actual articles with a selected category. Even here it consists of positive and negative values [9].

It can be defined as follows: R=m / (m+t) here, m stands for True positive and t stands for false-positive the end result gives us the processing time and the predicted results with the help of the bar graph as shown in the fig.

Algo	Precision	Recall	F1
Naïve Bayes	0.8717	0.814	0.7924
svm	0.8888	0.851	0.8380
Random Forest	0.8888	0.851	0.8380

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the

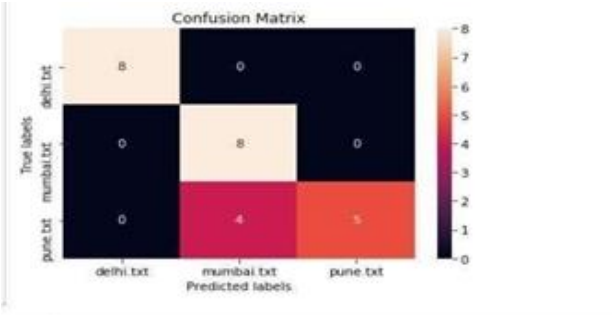


Figure 3. Confusion Matrix for Random Forest

In a confusion matrix we check whether the predicted values given by our algorithms is equal to actual true values which it must be. We consider 8 news articles which are of Delhi region and predict the label using the algorithms. It shows that all the news which are Predicted are in Delhi region itself showing the predicted value is right. Same goes for the Mumbai region news. For the Pune news it is seen that few news is getting classified in Mumbai even if they are in Pune. It is because few locations sharing the same name.

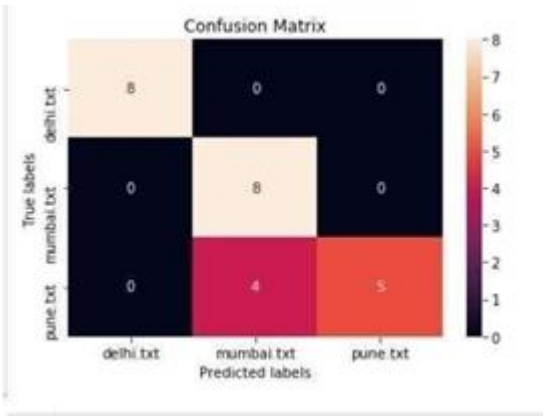


Figure 4. Confusion Matrix for SVM

The 8 news which are in Delhi are predicted in Delhi itself. This 8 news which are happening in Delhi are not characterized in Mumbai and Pune. Therefore, their value in the matrix is 0. Same goes for the Mumbai news. The values of Mumbai news getting predicted as Delhi and Pune is 0. Four of Pune's reports are characterized in Mumbai which isn't right. So, this segment show true values are known [7]. It allows the visualization of the performance of an algorithm. blunder and five of records is characterized in Pune itself.

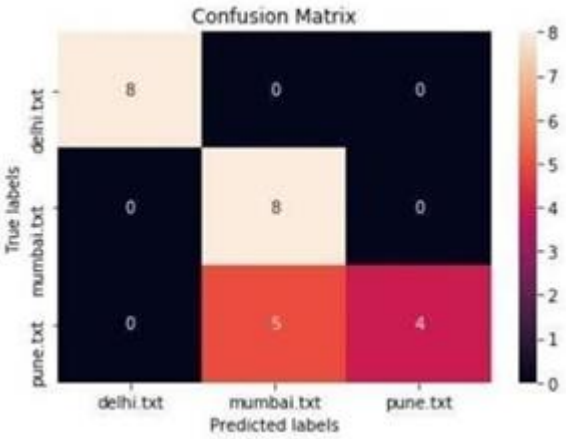


Figure 5. Confusion Matrix for Naïve Bayes.

There are eight news documents which are classified in Delhi itself. In similar manner there these eight documents which are occurring in Delhi are not classified in Mumbai and Pune so that value is shown as zero. There are eight documents which are classified in Mumbai similar manner these documents occurring in Mumbai are not classified in Pune and Delhi so that value is shown by zero. Five of Pune's documents are classified in Mumbai.

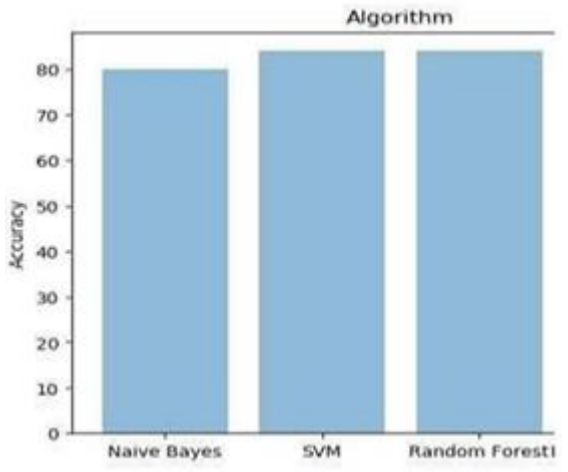


Fig 6. Graph of Accuracy of Algorithms

Thus, with this project is now possible to track live news and segregate it in Realtime into provided locations.

VI. CONCLUSION

In this paper, we've investigated the likelihood to use the machine learning algorithms to classify news articles based on cities. The experiments show that this problem may be successfully solved by using various Classifiers like Naive Bayes, Support Vector Machines and Random Forest. Random Forest has outperformed the opposite classifiers. Naive Bayes has performed well too and the Support Vector machine is at the bottom in terms of the performance metrics utilized in our approach. The proposed system may be used as part of more complex newspaper article classification systems. Our future target is to enhance the accuracy and also try classifiers like Neural Network. We can further increase the number of the input articles to 1000-fold compared to our present dataset for training our model.

VII. REFERENCES

[1]. B. Pendharkar, P. Ambekar, P. Godbole, S. Joshi, and S. Abhyankar." Topic categorization of rss news feeds," Group vol. 4, p. 1, 2007.

[2]. Shen, Z. Chen, Q. Yang, H. Zeng, B. Zhang, Y. Lu, and W. Ma, Web-page classification through summarization, in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004, pp. 242249.

[3]. Leo Breiman, Random forests, Machine Learning. vol. 45, no. 1, pp. 532, 2001

[4]. Lewis," Naive (bayes) at forty: The independence assumption in information retrieval," Machine Learning: ECML-98, pp. 415, 1998.

[5]. J. K. M. Han, Data Mining: Concepts and Techniques, 2nd ed. 2006.

[6]. Yu," SVM tutorial: Classification, regression, and ranking," Handbook of Natural Computing 2009

[7]. Rijsbergen, Information Retrieval,2nded London: Butterworths, 1979.

[8]. Y. Baeza and B. R. Neto, Modern Information Retrieval. Boston, 1999.

[9]. J. Davis and M. Goadrich, The relationship between precision- recall and ROC curves, in Proceedings of the 23rd International Conference on Machine Learning, ser. ICML06. New York, NY, USA: ACM, 2006, pp. 233240 [Online]Available: http://doi.acm.org/10.1145/1143844.1143 874.

[10]. T. Landgrebe, P. Paclk, R. Duin, and A. Bradley, Precision- recall operating characteristic (P-ROC) curves in imprecise environments, in Proceedings of ICPR, 2006.

[11]. Manning, C. and Schutze, H., Foundation of statistical natural language processing Cambridge, Mass: MIT press,1999.

[12]. C. Ee and P. Lim, Automated online news classification with personalization.

[13]. M. Kasthuri, Dr.S.Britto Ramesh Kumar, A Framework for Language In- dependent Stemmer Using Dynamic Programming, International Journal of Applied Engineering Research, ISSN 0973- 4562 Vol.10, pp 39000-39004, Number.18,2015.

Cite this article as :

Heneil Tayade, Chaitanya Shetty, Ratika Jankar, Dr. Amol Pande , "Segregation of Live News Articles Based on Location Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6, Issue 3, pp.505-512, May-June-2020.
Available at doi: https://doi.org/10.32628/CSEIT206380
Journal URL : http://ijsrcseit.com/CSEIT206380