International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 9 Issue: 5
DOI: https://doi.org/10.17762/ijritcc.v9i5.5464
***Article Received**: 22 March 2021 **Revised**: 05 April 2021 **Accepted**: 17 April 2021 **Publication**: 31 May 2021*

News Classification Using Machine Learning

Shweta D. Mahajan
Computer Engineering
Bharati Vidyapeeth College of EngineeringNavi Mumbai
Shwetam613@gmail.com

Abstract: There are plenty of social media webpages and platforms producing the textual data. These different kind of a data needs to be analysed and processed to extract meaningful information from raw data. Classification of text plays a vital role in extraction of useful information along with summarization, text retrieval. In our work we have considered the problem of news classification using machine learning approach. Currently we have a news related dataset which having various types of data like entertainment, education, sports, politics, etc. On this data we have applyingclassification algorithm with some word vectorizing techniques in order to get best result. The results which we got that have been compared on different parameters like Precision, Recall, F1 Score, accuracy for performance improvement.
Keywords *- Machine learning, News classification, Naive Bayes classifier, Natural language processing*

I. INTRODUCTION

There is a huge amount of information that we have to deal withdaily in the form of news articles. These techniques are essential used in textual data management techniques because the textual data is easily rising with the passage of time. Text mining tools are required to perform indexing and retrieval of text data. Text mining is used to extracting hidden information which is finding of some information from large amount of dataset is a very strong tool that is used for classification purpose and this text in the form of unstructured text. Unstructured text rising mostly the information is available in digital form for example world wide web, e-mail, publication, etc. A word and a sentence ambiguity may include in an actual form or sequence occurred in that text any type of data this term is called as unstructured text. This text in sequence of extraction on useful information are using processing and pre-processing techniques are required. The processing method cannot be used in a computer. The computer using the text as order of string and not provide any kind of information. For better classification in every field large number of text material are available. These include medical, finance, image-processing, and many other fields, where the major objective of text mining is to extract useful information from semi-structured or unstructured text by making best use of techniques i.e., supervised or unsupervised classification or Natural Language Processing. All the traditional natural language processing algorithms have been known to majorly operate on words to decide predefined classes for particular text or text-documents. The data set used in this study has been taken by the News Classification Dataset(.json) separated into train data and test data. The news classification performed major work based on text mining. News articles are in the form of large informationbut nowadays long length news is not worked so our main focus is on specifically limited news headlines. News classification depend on statistics and deep learning. There are many classifiers to classify the news classification like naïve bayes, support vector machine (SVM) [4], decision tree, k-nearest neighbour (KNN) and many more. In this paper Naive Bayes technique are used with word vectorization.

II. LITERATURE SURVEY

[1] Have proposed Mykhailo and Volodymyr, which is a naïve bayes technique used for fake news detection. They are implemented one software and test dataset on Facebook post. They are targeted on accuracy and to check true and fake article. They are verified similarities on spam messages and fake messages using naïve bayes approaches. Explained with formula and test dataset on Facebook post and calculate test accuracy evaluation in precision and recall. Then they classifywhich post is fake and correct.

[2] This paper focus on Indonesian news classification with categories. They are used Nazief-adriani stemmer method for each word reduced into basic word and for classification technique is used for naïve bayes. They are uses dictionary from Katlego. In that remove Prefix and suffix.
They used TF-IDF concept also check in one document how many terms are there and how much weight are reduced[2]. They explain the stages wise which is used in naïve bayes one is stage of training and second is phase classification explain with examples. After the result analysis they receive accurate 94% values.

III. OBJECTIVES

- To implement for classification of news into accurate category

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 9 Issue: 5
DOI: https://doi.org/10.17762/ijritcc.v9i5.5464
***Article Received**: 22 March 2021 **Revised**: 05 April 2021 **Accepted**: 17 April 2021 **Publication**: 31 May 2021*

- News classification to determine the precision, recall, F1score for accuracy
- To implement graphical representation of news in the form of Bar chart, Tree map, word cloud
- To implement and compare naïve Bayes techniques andwith TF-IDF and check the accuracy

III. EXISTING SYSTEM

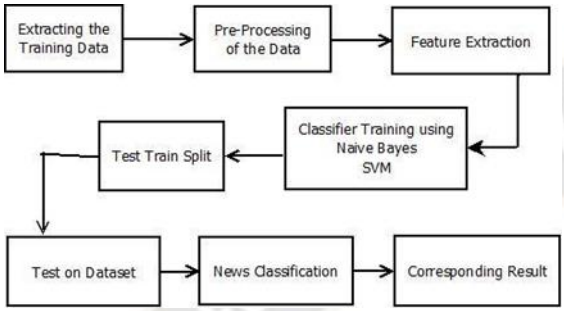

Figure 1 Flow chart – Classifier Training

Figure 1 show the process of classification and given step toretrieve result by step by step in that pre-processing step, especially in feature extraction. Apart from, classification technique they use Naïve Bayes and SVM classifier [10].

IV. PROPOSED SYSTEM

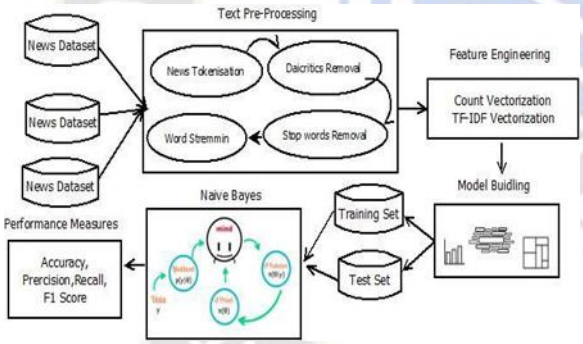

Figure 2 Flow chart – Proposed Model

Nowadays in this pandemics people want to read a news but what was happed to they search a news in google or any of thewebsite all news are shown one by one [7]. But they want a one particular category of news like political, sports like that. So, we are categorised news. We are taking a news dataset from google and to check that news is in which categories.

V. METHODOLOGY

Our approach is to take output to the news headlines based on shot description. Fig 2 shows the whole process of the flow. The beginning with news dataset is call data retrieval module means the collection of datasets in our implementation we are collected from web sites and extract actual text form. The dataset is divided into two-part train dataset and test dataset. Train data is collected 80% of the dataset. Text Pre-processing are applied to the train data is used to classifier for training and for that various methods we are follows like news tokenisation, diacritics removal, stop words removal, word stemming. This input to trained classifier and build model to naïve Bayes classifier [6] and check the predicts the result of test news headlines. This is accompanying by the evaluation of the accuracy of news classifier based on performance measures.

A. Text Pre-processing:

We are collecting news in many of the sources is to be seen in newspapers, magazines, etc. Dataset are available in many formats that is it may in.csv, json,.pdf, .doc, or in .html format. After the collection of news is done [7]. Dataset we retrieve from various sources so it has to be required for cleaning that it should be free from noisy and useless information. We need delete unrelated words from data means full stop, brackets, semicolon, etc [6]. so, data is short listed from these words those are appear in text are called stop words. In this paper used some libraries for example NumPy, pandas, sklearn.

News Tokenization: Divide text into small words or segments is called to be text tokenization [9]. News headlines every word. Every word in the headline and content is evaluated as a string, which will then be broken down into tiny pieces. The end outcome would be used as information for text mining processing. All of the headlines are merged to form a set of words.
Diacritics Removal: The meaning of diacritics varies depending on the language [9]. Commas, semi-colons, quotations, double quotes, full stops, underscores, special characters, and brackets, among other things, are eliminated from all words.
Stop words Removal: Stop words are all the words in a text that join words or connector lines [9]. They are simply considered once and then eliminated. Those words, which appear frequently in news headlines, are deemed ineffective in terms of frequency.
Word Stemming: The process of deleting a portion of a word or reducing a word to its stem or root is known as stemming. It's possible that we're not reducing a word to its dictionary root.

B. Feature Engineering:

The Process of transforming data to improve the predictive performance of the trained models is called as feature engineering [5].

Count vectorization: It is basically count like how many wordsare there in the feature set. In machine learning feature set has to be extracted from text document. In the feature set has many dimensions as the many unique number in full dataset and this approach every unique word as separate feature and directed shown as each set of features as a document [6]. Count of word assign in a document in to its related to that feature is called ascount vectorization method.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 9 Issue: 5
DOI: https://doi.org/10.17762/ijritcc.v9i5.5464
***Article Received**: 22 March 2021 **Revised**: 05 April 2021 **Accepted**: 17 April 2021 **Publication**: 31 May 2021*

It is not able to specific word combinations.

TF-IDF vectorization: It is common words based on rescale the frequency [2]. How frequently used same words are scores in all present document like “that”, “is”, “then”. It takes advantage of the concept of Term frequency- Inverse document frequency. Team frequency is number of occurrences in a document. Inverse Document frequency is downscaling words that appear across document [2]. This common word is allowed to reduce the weight. for example, one certain term such as “of”, “is”, “that” may appear many times but it has low importance then we need to weight down the frequent terms.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Thus, most of the document some word is frequent and the denominator numerator get close to each other and IDF score get zero, thus words which is not discriminative enough get close to zero weight that means infinity. We have applied one type of feature extraction methods and trained one classification technique that is naïve Bayes. Comparative analysis of naïve Bayes methods and naïve Bayes with TF-IDFvectorizer [2]. While working with features generated by TF- IDF vectorization method rather that count vectorization and classifier check accuracy with the both methods.

C. Data Visualization:

We're visualising three different approaches to categorise data[3][4]. Figure 3. We present categorising distribution in a bar chart with two dimensions: categories and counts, and we present how many counts are in each category [8]. Like that Figure 4. we show treemap for news category where as we show one more visualisation is use that is wordcloud Figure 5.For all dataset word [3].

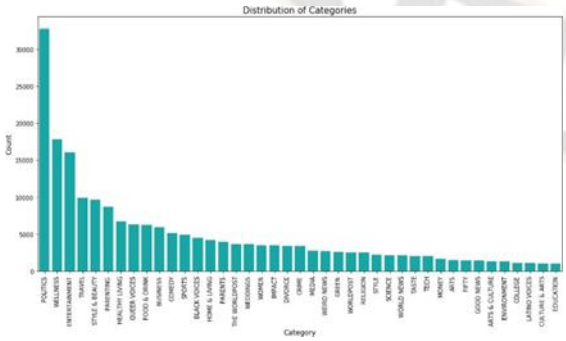

Figure 3. Bar Chart for Distribution of categories

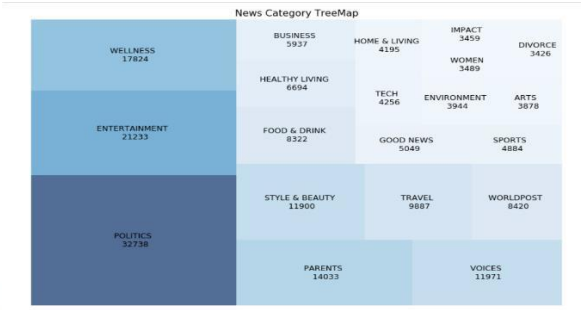

Figure 4. TreeMap for Distribution of categories

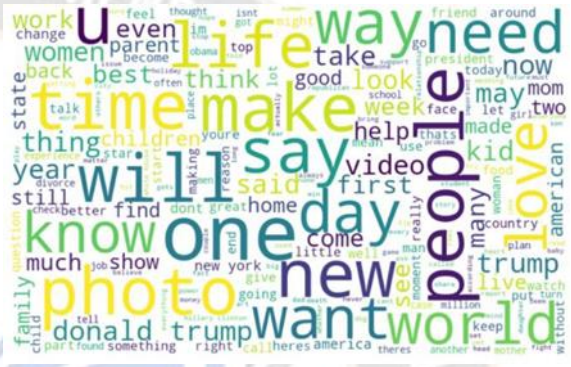

Figure 5. Wordcloud for all words

D. Naïve Bayes:

In machine learning, naïve Bayes classifiers is a simple probabilistic classifiers family. It is based on Bayes theorem and text features [1]. Features are presumed to be independent of one another. It individual calculates text probability within class label and classes. We trained the naïve Bayes classifier with count vectorization and with TF-IDF vectorization [2]. In this text classification techniques naïve Bayes classifier gives accuracy results shows but for instance, not every predication is accurate. It shows 80% to 85% predication are accuracy.

This accuracy is achieved by applying count vectorization as feature extraction observed and we got highest accuracy and the least accuracy observed by using TF-IDF vectorization approach with naïve Bayes. Naïve Bayes is best classifier with textual as well as numeric data formats and easy to implement and to compute, relatively robust, fast, accurate and to compute but it shows poor performance when features are the short text classification. Naïve Bayes is classifiers the particular feature assuming that value and this value is independent of any other feature.

The key concept is to regard each piece of news as a separate entity. First, we will use some data set using from internet in that dataset and their entities for example author, category, date, headline, link, short description. The Bayes theorem is the foundation of Nave Bayes, which indicates that features in a dataset are mutually independent. The probability of occurrence of a trait is unaffected by its occurrence.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 9 Issue: 5
DOI: https://doi.org/10.17762/ijritcc.v9i5.5464
***Article Received**: 22 March 2021 **Revised**: 05 April 2021 **Accepted**: 17 April 2021 **Publication**: 31 May 2021*

E: Performance measures:

We use precision, recall and F1 score, support as the performance measure [5] and compare with naïve Bayes classifier Figure 6. and TF-IDF Figure 7. Precision is the ratio of the number of correct results to number of total results [3]. Recall is ratio of number of correct results to number of correctresults that should have been returned [3]. F1 score is functionof precision as well as recall [3]. Support is how many times iscoming in a dataset [9].

	precision	recall	f1-score	support
0	0.64	0.42	0.51	833
1	0.49	0.51	0.50	1233
2	0.48	0.66	0.56	641
3	0.77	0.61	0.68	671
4	0.51	0.31	0.38	421
5	0.66	0.71	0.68	4313

Figure 6. Performance measure using Naïve Bayes

	precision	recall	f1-score	support
0	0.02	0.02	0.02	740
1	0.03	0.04	0.03	1185
2	0.02	0.03	0.02	688
3	0.02	0.01	0.01	693
4	0.00	0.00	0.00	423
5	0.10	0.11	0.11	4270

Figure 7. Performance measure using TF-IDF

VI. IMPLEMENTATION AND RESULTS

The four existing techniques are considered for implementation purposes. The results of the four models presented are consistent with the suggested model, and the category of news is correctly identified [8]. The demonstration is done with certain machine learning algorithms and python programming on Jupiter Notebook. Following example of the news classification method using Naïve Bayes [2].

No	News	Category
1	India's largest ever 'eye sky' takeneighbours	TRAVEL
2	US Vice President Mike Pence Did Not Fake Getting COVID-19 Vaccine	POLITICS
3	Rapper Skinnyfromthe9 Handcuffed Weed Bust	CRIME
4	How reach Kedarnath Temple: quickguide	TRAVEL

Table1: Result of Naïve Bayes classification

CONCLUSION

After conducting research and analysis, the findings of this study show that Naive Bayes can successfully categorise news, and TF-IDF is at the bottom of the performance measures employed in our methodology. Our next goal is to enhance accuracy while also experimenting with classifiers such asSVM and decision trees.

ACKNOWLEDGMENT

This progress report on the creation of "News Classification" brings me tremendous delight. We are appreciative to Bharati Vidyapeeth College of Engineering's Institute of Engineering for providing us with this fantastic opportunity to work on this big project. Dr. Dyanand Ingle, our project supervisor, provided us with invaluable guidance and suggestions. We'd also want to thank Prof. Sheetal Thakare, Project Coordinator, for providing us with all of the materials we needed to complete our project. We would like to express our gratitude to the Department of Computer Engineering's teaching and non-teaching personnel for their important assistance and support throughout the project hours. We shall not forget to thank everyone who has helped us achieve this goal.

REFERENCES

[1] Mykhailo Granik, Volodymyr Mesyura "Fake News Detection Using Naive Bayes Classifier" 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering(UKRCON)

[2] Garin Septian, Ajib Susanto, Guruh Fajar Shidik Facultyof Computer Science "Indonesian News Classification based on NaBaNA" 2017 International Seminar on Application for Technology of Information and Communication (iSemantic)

[3] Vignesh Rao,Jayant Sachdev "A Machine Learning Approach to classify News Articles based on Location" Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2017) IEEE Xplore Compliant - Part Number:CFP17M19-ART, ISBN:978-1-5386-1959-9

[4] David Martens, Bart Baesens, and Tony Van Gestel, "Decompositional Rule Extraction from Support Vector Machines by Active Learning" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO.2, FEBRUARY 2019

[5] Z. Liu, X. Lv, K. Liu, and S. Shi, "Study on SVM compared with the other text classification methods," 2nd Int.Work. Educ.Technol. Comput. Sci. ETCS 2010, vol. 1, pp. 219-222, 2010.

[6] M. IKONOMAKIS, S. KOTSIANTIS, V. TAMPAKAS, "Text Classification Using Machine Learning Techniques" WSEAS TRANSACTIONS on COMPUTERS, Issue 8, Volume 4, August 2005, pp.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 9 Issue: 5
DOI: https://doi.org/10.17762/ijritcc.v9i5.5464
Article Received: 22 March 2021 Revised: 05 April 2021 Accepted: 17 April 2021 Publication: 31 May 2021

966-974.

[7] Kiran Shriniwas Doddi , Dr. Mrs. Y. V. Haribhakta , Dr. Parag Kulkarni, "Sentiment Classification of News Articles", Kiran Shriniwas Doddi et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 4621-4623

[8] J SreeDevi, M Rama Bai, M Chandrashekar Reddy,
"Newspaper Article Classification using Machine Learning Techniques" International Journal of Innovative Technology and Exploring Engineering (IJITEE)ISSN: 2278-3075,
Volume-9 Issue-5, March 2020

[9] Mazhar Iqbal Rana, Shehzad Khalid, Muhammad Usman Akbar "News Classification Based on Their Headlines: A Review" ISBN: 978-1-4799-5754-5/14/$26.00 ©2014 IEEE

[10] Anjali Jain, Avinash Shakya, Harsh Khatter, Amit Kumar Gupta "A smart system for fake new detection using machine learning" 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT) ISSN: 2321-9939