Journal of Education for Pure Science- University of Thi-Qar
Vol.12, No2 (December, 2022)
Website: jceps.utq.edu.iq Email: jceps@eps.utq.edu.iq

DOI: http://doi.org/10.32792/utq.jceps.12.02.32

News Classification by N-Gram and Machine Learning Algorithms

Melad Farawn Ogaib[1] Prof. Dr. Kadhim Mehdi Hashim[2]

[1,2] Department of Computer Science, College of Education for pure Sciences, University of Th-Qar[1], Iraq.
Imam Ja'afar Al-Sadiq University[2], Iraq.

Received 11/8/2022 Accepted 6/9/2022 Published 01/12/2022

Abstract:

News is information obtained from different sources such as television, internet, newspapers and magazines. Online news is published in very large numbers, and because there are so many news, it will be challenging for users to find the pertinent information that matches their preferences. In this paper, the news is categorized so that a specific category can be obtained quickly and easily. The BBC's newsgroup was used in its five categories: sports, politics, business, technology and entertainment. The classification algorithms Multinomial Naive Bayes (MNB) and decision tree (DT) were applied to the news data set after extracting the features from it using n-gram method . Multinomial naïve Bayes algorithm has proven superiority over decision tree with an accuracy 98.2%.

Keywords: Multinomial Naïve Bayes, Decision Tree, N-gram .

Introduction:

Natural Language Processing (NLP) is examination of language data most commonly which is in form of transcripts such as documents or articles using computational modeling. The purpose of NLP is to use linguistic insights to construct a representation of the text that gives structure to unstructured natural language[1]. The creation of automatic translation techniques and software, sentiment analysis, text summarization, authorship identification are all issues that are currently being researched in the NLP domain. Speech recognition is closely related to (NLP) and could be regarded a subtopic of it[2].

1.Machine Learning (ML): is an area of artificial intelligence and computer science that focuses on the use of data and algorithms to mimic the way humans learn, with the goal of constantly increase accuracy. It also refers to the development of fully automated devices that are controlled by creating predefined algorithms and rules. Machine learning algorithms implement and generate the best results using data and a set of pre-defined rule [3], [4].

Designing algorithms that can learn from data is the aim of ML. A range of real-world issues have been addressed using ML techniques, including speech recognition, fraud detection, customer relationship management, and others. Take the classification of email messages as spam or not, where the effectiveness of a machine learning system is determined by the proportion of emails that are correctly classified [5].

2 News Classification

The amount of news published on various websites in general, and news websites in particular, has increased in recent years. Text and visual descriptions (photos and video) are included in these articles, as well as multimedia resources. Journalists and media watchdogs alike are now faced with the task of sifting through massive amounts of content in order to uncover important topics and events everywhere[6] . It is difficult to find news relevant to a person's interests because many news articles are useful but may be irrelevant. News articles are categorized by news type/category for easy selection of people according to their likes. News and the type of news, may affect the attractiveness of an individual [7].

3. RELATED WORKS

News are published on the Internet in very large numbers, searches take a long time, therefore, many previous studies focused on the classification of news, including:

Z. Jawad et al. (2020) [8], to accomplish the necessary reliability and efficiency, a combination system is used to discover best prediction among three classifiers (Naive Bayes, Support Vector Machine, and K-Nearest Neighbors). This method minimizes the variation in classifiers' performance. The following four data sources were used: 20-newsgroup, Reuters-21578, Watan-2004, and Khalaf2018. Results from the proposed system were superior to those from the individual classifiers. The respective F1- scores for the suggested system for 20-newsgroup, Reuters-21578, Watan-2004, and Khalaf-2018 were 95%, 93%, 94%, and 92%.

M. Rahman, M. Khan et al. (2021) [9], creating an automated system for classifying Bengali news articles and creating advanced solutions for a set of textual data. Despite limited data set, TextGCN outperformed BiLSTM, GRU-LSTM, LSTM, Char-CNN, and BERT in the online Bangla news ranking. According to the results comparison, Text-GCN achieved a maximum accuracy of 96.25 %, Text-GCN has the highest accuracy, recall, and F1 score, with 96.25% in each segment.

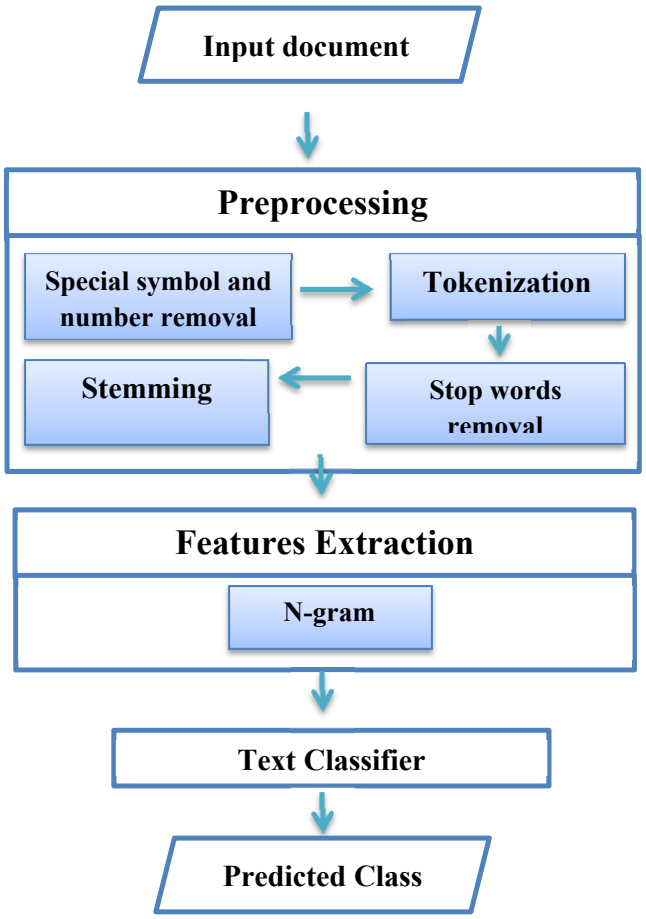
P. P. Ramadhani, S. Hadi. (2021) [10], instagram was divided into organized groups (fashion, food & beverage, technology, health & beauty, lifestyle & travel) with 66,171 comments, and the support vector machine algorithm was evaluated to categorize Instagram comments. TFIDF method was used for feature extraction, the support vector machine algorithm with radial basis function (RBF) resulted in an average F1 score of over 88%.

J. Ahmed, M. Ahmed. (2021) [7], the framework for classifying news articles is explained, as well as some current methods for selecting news online. Various classifiers were tested to reach high accuracy. The experimental method using Bayesian classifier achieved an accuracy rate of 93%.

M. Khan et al. (2021) [11], use of online Urdu news data to train algorithms to automatically categorize the presented information. The results showed that the Bernoulli Naïve Bayes (Bernoulli NB) and Multinomial Naïve Bayes (Multinomial NB) algorithms beat the others in terms of all performance criteria. An NB polynomial classifier have accuracy of 94.278 %, followed by a Bernoulli NB classifier accuracy of 94.274 %. The findings suggest that a brief text for news headlines can be utilized as a text classification input.

4. PROPOSED METHOD

In this paper, news is classified into its different categories using machine learning algorithms. The features are extracted using N-gram algorithm, and this process is considered one of the most important steps used to classify the news because the classification process depends mainly on extracting the features in order to reach the best classification. Figure (1) shows the the basic stages used in classification process



Figure(1): The basic stages for news classification

4.1 Data Collection: The data we wish to categorize is collected from BBC News in 2,225 texts divided into five categories: business, sports, technology, political and entertainment. The news data is distributed as follows:(511 sport ,510 business, 417 politics , 401 tech , 386 entertinment)

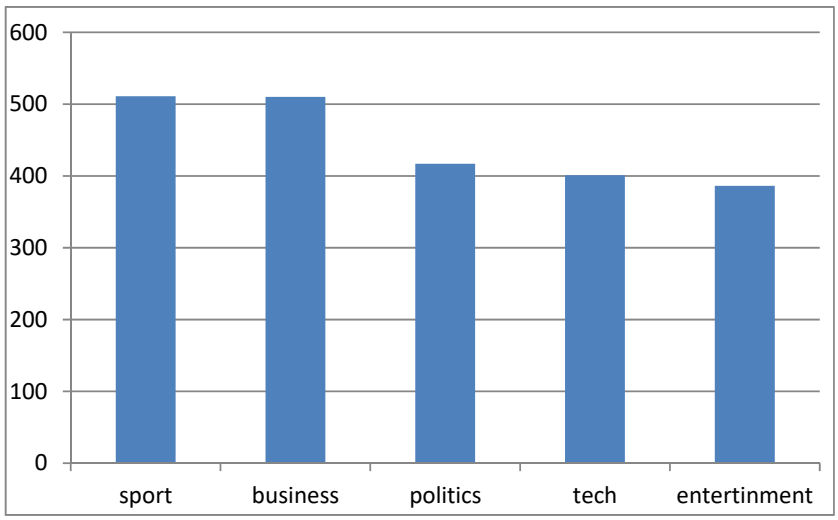


Figure (2): Categories of news dataset

4.2 Pre-Processing Steps

At this stage, several operations are performed on the texts to format them and make them ready for use and to reduce noise in the text for the purpose of obtaining the best performance for the classifier.

A: Special symbol and number removal: Remove any symbol from the set of special symbols (!, :, ÷, ×, º, >, <, \, /, @, #, $,%, ^, &, *,), (, _, -, +, =, ~, ø, [,],", etc) in the text to keep only related words, resulting in a much better feature extraction procedure and higher classification accuracy.

B: Tokenization: This process includes dividing the text into words or sentences, in word tokenization, each word is separated from the other by a separator, and each word is placed in quotation marks in order to facilitate the handling of the text and perform other operations better.

C: Stop word removal: Commonly used many words in the English language, which are never without text, such as (conjunctions, pronouns and prepositions) These words are not related to a specific category, but are present in all categories, stop words do not carry any important information in the text and have a low value, therefore, they are removed and the most relevant and high valuable words are used in the classification process.

D: Stemming: Derivation is a very important step in the preprocessing where the number of words is reduced by returning each word to its root, since each word can come in several images, these images return to one root. The process of derivation is important because it reduces the time to extract features.

4.3 Feature Extraction

The main objective of feature extraction is to define a set of inputs containing relevant information that is used in the text classification process. In this paper, the following algorithm have been used to extract features from text:

A: N-gram

Features can be extracted using n-gram, which contains several types, first type represents dividing the text every one word separately , and this is called 1-gram, and the second type divides every two words separately called 2-gram, and so for the rest of types where the division is according to type of n-gram that is chosen.

An Example " After sleeping for four hours, he decided to sleep for another four"

the tokens in 2-gram are as follows: "After sleeping", "sleeping for", "for four", "four hours", "hours he", "he decided", "decided to" , "to sleep", "sleep for", "for another", "another four".

the tokens in 3-gram are as follows: "After sleeping for", "sleeping for four", "four hours he", " hours he decided", "he decided to", "to sleep for", "sleep for another", "for another four".

4.4 Classification

After the process of extracting the features, the news data set is classified into its specific categories. In our proposed system, we use the MNB algorithm and the DT in classification process.

A: Multinomial Naïve Bayes(MNB)

Is one a form of Naïve Bayes used for multinomial distributed data. It is widely used as a baseline in text classification since it is fast and easy to construct. Furthermore, with proper preparation it may compete with more complex methods such as support vector machines [12]. In order to prevent zero probability of

new words in the testing set that weren't present in the training set, we need to add a smoothing one to the conditional probability when using multinomial naïve bayes [13].

B:Decision Tree (DT)

It's a supervised learning algorithm that's commonly used to solve classification difficulties. It works for both categorical and continuous variables [14]. Decision trees collect attributes by ranking them in accordance with their values. Each tree is made up of nodes and branches; each branch provides a value that the node take, and each node reflects characteristics in a group that need to be categorized [15].

4.5 Experimental Results

The data set is divided into 80% training and 20 % testing, the classification algorithms are trained on the features extracted from N- gram , and then the test features are classified and the accuracy calculated for both MNB and DT algorithms. Implementation requirements which were used to implement the proposed system include:

A. Hardware Requirements

In this proposed system, huawei computer with specifications Intel (R) Core (TM) i5-10210U CPU @ 1.60GHz 2.11 GHz ,RAM of 8 GB has been used.

B. Software Requiremen

The using software is python (3.9) environment Anaconda Navigator (Jupyter) installed on Windows 10 Pro 64-bit operating system. The data set size is 5 MB.

4.6 Evaluation

Newsgroup is classified by Multinomial Naïve Bayes algorithm and Decision Tree algorithm. The proposed model was evaluated and performance measures were calculated for each algorithm.

Some well-known classifier evaluation measures include accuracy (also known as recognition rate), sensitivity (or recall), specificity, precision, and F1 Score [7], [16], [17].

A. Accuracy: Is the number of correct predictions made by the classifier divided by the total data collection. The accuracy is stated mathematically as

$$\text{Accuracy} = \frac{(TP+TN)}{TP+FP+TN+FN} \quad \ldots\ldots\ldots \quad (1)$$

B. Precision: By dividing the number of articles that are correctly recognized (True Positive) by group (True Positive and False Positive). The precision in mathematical form is:

$$\textbf{Precision} = \frac{TP}{TP+FP} \ldots\ldots\ldots\ldots \quad (2)$$

C. Recall: The recall is defined as the proportion of accurately predicted positive articles to group (True Positive and False Negative). The recall is mathematically expressed as

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{..... (3)}$$

D. F1 score: The F1 score is the harmonic mean of precision and recall. When we add Precision and Recall together we get the F1 score. Because it includes both precision and recall, using one value for measurement is more efficient than using two. F1 score is presented as

$$\text{F1- Score} = \frac{2*Recall*Precision}{(Recall+Precision} \quad \text{..... (4)}$$

The following table show the experimental results of the proposed model

Table (1):MNB and DT with N-gram

metrics	MNB with N-gram (2-gram)	DT with N-gram (2-gram)
Accuracy	%98.2	%78.8
Precision	%98.2	%80.2
Recall	%98.2	%78.6
f1-score	%98.4	%79

5. CONCLUSION

There are huge amounts of news text scattered on different news sites, but it is usually unorganized, so a process of classifying the texts was needed to get the structured form. Machine learning algorithms do not have the ability to deal with texts only after converting them to numerical values, so the N-gram algorithm was used in this paper to obtain the features and this is considered the most important step in classification process, and then classification is done using both multinomial naive Bayes and decision tree. The results showed that the MNB algorithm was superior to the DT algorithm with an accuracy of 98.2%. Future work may include the use of a data set that contains all news categories.

REFERENCES:

[1] **K. M. Verspoor and K. B. Cohen,** "Encyclopedia of Systems Biology," Encycl. Syst. Biol., no. June 2018, 2013, doi: 10.1007/978-1-4419-9863-7

[2] **S. Theodoridis,** Machine learning A Bayesian, vol. 53, no. 9. 2019.

[3] **N. Ortiz, R. D. Hernandez, R. Jimenez, M. Mauledeoux, and O. Aviles**,

"Survey of biometric pattern recognition via machine learning techniques," Contemp. Eng. Sci., vol. 11, no. 34, pp. 1677–1694, 2018, doi: 10.12988/ces.2018.84166.

[4] **S. S. Mousavi, M. Schukat, and E. Howley**, "Deep Reinforcement Learning:

An Overview," Lect. Notes Networks Syst., vol. 16, pp. 426–440, 2018, doi: 10.1007/978-3-319-56991-8_32.

[5] **A. Ławrynowicz and V. Tresp**, "Introducing machine learning," Perspect. Ontol. Learn., vol. 18, no. November, pp. 35–50, 2014, doi: 10.1007/978-3-

030-67626-1_8.

[6] **D. Liparas, Y. Hacohen-Kerner, A. Moumtzidou, S. Vrochidis, and I. Kompatsiaris**, "LNCS 8849 - News Articles Classification Using Random Forests and Weighted Multimodal Features," 2014. [Online]. Available: http://www.bbc.com/news/business-25445906.

[7] **J. Ahmed and M. Ahmed,** "Online News Classification Using Machine Learning Techniques," IIUM Eng. J., vol. 22, no. 2, pp. 210–225, 2021, doi: 10.31436/iiumej.v22i2.1662.

[8] **Z. M. Jawad and Z. A. Khalaf**, "The combination of text classification system 1," vol. 1, no. 1, 2020.

[9] **M. M. Rahman, M. A. Z. Khan, and A. A. Biswas,** "Bangla News Classification using Graph Convolutional Networks," Jan. 2021, doi: 10.1109/ICCCI50826.2021.9402567.

[10] **P. P. Ramadhani and S. Hadi**, "Text classification on the Instagram caption using support vector machine," J. Phys. Conf. Ser., vol. 1722, no. 1, 2021, doi: 10.1088/1742-6596/1722/1/012023.

[11] **M. B. Khan,** "Urdu News Classification using Application of Machine Learning Algorithms on News Headline," IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 21, no. 2, p. 229, 2021, doi: 10.22937/IJCSNS.2021.21.2.27.

[12] **S. Xu, Y. Li, and Z. Wang,** "Bayesian multinomial naïve bayes classifier to text classification," Lect. Notes Electr. Eng., vol. 448, no. November, pp. 347–352, 2017, doi: 10.1007/978-981-10-5041-1_57.

[13] **H. M. Ismail, S. Harous, and B. Belkhouche,** "A Comparative Analysis of Machine Learning Classifiers for Twitter Sentiment Analysis," Res. Comput. Sci., vol. 110, no. 1, pp. 71–83, 2016, doi: 10.13053/rcs-110-1-6.

[14] **A. Abdi,** "Three types of Machine Learning Algorithms List of Common Machine Learning Algorithms," no. November, 2016, doi: 10.13140/RG.2.2.26209.10088.

[15] **A. Dey,** "Machine Learning Algorithms: A Review," Int. J. Comput. Sci. Inf. Technol., vol. 7, no. 3, pp. 1174–1179, 2016, [Online]. Available: www.ijcsit.com.

[16] **J. Han, M. Kamber, and J. Pei,** "Data Mining : Concepts and Solution Manual," Data Min. Concepts Tech. Solut. Man., p. 135, 2012, [Online]. Available: https://moam.info/data-mining-concepts-and-techniques- solutionmanual_59894d1b1723ddd1695415f9.html.

[17] **S. Kumar Thapa and S. Pokhrel**, "Nepali News Document Classification using

Global Vectors and Long Short Term Memory."