RIK International Journal of Science and Technological Research (RIK-IJSTR) *Vol. 7 No. 3, August 2021*
ISSN: 2350-6310 (Print)
ISSN: 2350-543X (Online)

Improved Document Categorization Using Natural Language Toolkit in Python

Orlunwo, Placida Orochi
Department of Computer Science
Ignatius Ajuru University of Education
Rivers State
Nigeria

&

Onuodu, Friday Eleonu
Visiting Scholar
Department of Computer Science
University of Port Harcourt
Rivers State
Nigeria

ABSTRACT

The challenge of selecting, filtering, and managing increased quantities of textual data, to which access is usually crucial, face modern information technologies and web-based services. Text Retrieval is an information subtask that permits users to more quickly navigate across category hierarchies through the collection of texts of their own interests. In the realm of artificial intelligence, machine learning systems automatically develop models from the data to make better decisions. Natural language processing (NLP) offers good performance in statistical problems, including text categorization or sentiment mining, through corpora and learning methodologies. The study is focused on the role of Natural Language Processing in document categorization with the aim of increasing accuracy in the categorization of documents. The specific objectives are to design a model that will optimize accuracy of document categorization, implement the document categorization using python language NLP Toolkit for text processing and naïve bayes method to build a model and evaluate the accuracy of the categorization. The data set consisted of 93239 news reports and six features; the naive Bayes model obtained 95.5% precision after text processing. This is beneficial to office assistants and any other cooperate position in the management of large volume of document.

Keywords: Document Categorization, Natural Language Processing Toolkit, Machine Learning.

Introduction

In numerous organizations, document monitoring has been a source of concern. Various tactics for document monitoring have been established over time, resulting in the creation of software systems for document management. Document dissemination, workflow, email management, and document security are all addressed by these applications. Regardless of these features, we need to improve document classification efficiency and efficacy. The challenge of selecting, filtering and categorization of increasing amounts of textual information, to which access has usually been essential, is addressed through modern computer algorithms as Natural Language Processing (NLP).

Document Categorization

The classification of documents is an example of machine learning (ML) by means of natural language processing (NLP). We classify a text, which makes management easier and sort easy, by assigning one or more classes or categories to a document. It is particularly handy for publishers, news sites, blogs and anyone dealing with much content. Due to the large quantity of documents that are involved, automated approaches are needed to efficiently categorize documents. The objective of the text classification is to classify the text documents automatically into one or more defined categories. Some examples of text grading are: Audience understanding of social media sentiments, Spam and non-spam emails detection, Customer requests are auto-tagged, and Categorization in defined issues of news stories.

The categorization of documents is an information systems or computer science challenge. A document is assigned to one or more classes or categories, either manually or with certain algorithms. Classification can allow an organization to meet legal and regulatory obligations for the collection of certain information within a certain amount of time. However, the data techniques vary widely, since the diverse data types and quantities are generated by each firm.

Types of Document Classification and Techniques

i. Supervised Document Classification

The supervised classification offers the proper information about document classification through an external mechanism (e.g. human feedback).

ii. Unsupervised Document Classification

In unattended categorization of documents, sometimes referred to as document clustering, classification must be done completely without reference to externals. Clustering of documents requires the use of characteristics and the extraction of features. Characteristics are word sets describing the contents of the cluster. The clustering of documents is often regarded as a centralized procedure.

Text Classification is an example of a supervised machine learning job since text documents are marked with a labelled dataset and their labels are used for classification training. There are three basic components of an end-to-end text classification pipe:

1. Preparing Datasets: The first stage is the Dataset Preparation step which involves the loading and the basic pre-processing of a dataset. The dataset is then broken down into train sets.
2. Engineering Feature: The following phase is the Feature Engineering process for converting raw data into flat features that can be used in a machine learning model. In this step, new features from current data are also created.
3. Model Training: The final step is to construct a model, which is trained on a marked data set in the machine learning model.

Natural language processing

Natural language processing is a set of computerized methods and algorithms to detect patterns in text data. By treating word and word clusters as meaningful, NLP more effectively extracts texts from concepts, and relationships than people can do using the representation of bag of words to train the target classification feature, standard statistical models of machine learning. The only features used to learn the statistical model are the words contained in the documents. The development of classification functions is completely overlooked by typical natural language structures such as morphology, syntax and synanthropic. Semantic information from the Text Categorization models is still not being used for the most important natural language applications.

Python Natural Language Toolkit (NLTK)

Steven Bird and Edward Loper of the University of Pennsylvania created the library, which was essential in ground-breaking NLP research. In Python, NLTK is a useful package that helps with categorization, stemming, tagging, parsing, semantic reasoning, and tokenization. It's essentially your major machine learning and natural language processing tool. It now serves as a basis for Python developers who are just getting their feet wet in the industry (and machine learning). NLTK, Python libraries, and other technologies are now used in many university courses throughout the world.

Natural Language Text Processing

Bag of words: For machine learning algorithms, we need a mechanism to represent text data, and the bag-of-words model can help us with that. It's a technique for extracting text features for use in machine learning algorithms.

Tokenization: It divides a chunk of text into distinct words using a delimiter. Different word-level tokens are created depending on the delimiters. Word tokenization includes pre-trained word embeddings like Word2Vec and GloVe.

Stop word removal: The most prevalent words in any natural language are stopwords. These stopwords may not contribute much value to the meaning of the document when evaluating text data and constructing NLP models.

Lowercasing: Words like book and book have the same meaning but are represented as two different words in the vector space model if they are not transformed to lower case (resulting in more dimensions).

Stemming: it is the process of reducing a word to its basic form.

Lemmatization: unlike stemming, reduces the words to a single term that already exists in the language.

Statement of the Problem

In this thesis a study has been carried out on the interactions between natural language processing and document categorization. Since these require a high level of efficiency and precision, we have studied and implemented models with both features. Most experiments on natural language information retrieval techniques have shown the inefficiency of current linguistic processing to improve the categorization of statistical text. It is indeed difficult to process natural text automatically in the original form. In order to put the text into a form suitable for further analysis, tokenization, stop word removal, lowercasing and stemming will be applied to various pre-processing types.

Aim and Objectives

The study is focused on the role of natural language processing in document categorization with the aim of increasing accuracy in the categorization of text. The specific objectives are to design a model that will optimize accuracy of document categorization, implement the document categorization using Natural language toolkit (nltk) for text processing and naïve bayes method to build a model and evaluate the accuracy of the categorization.

Related work

In organisations, text data is prevalent, digitisation and the ease with which information is created online (e.g., e-mail messages etc) help in generating a wide range of everyday documents (Cardie & Wilkerson 2008). Information that can increase our understanding of organizational processes is included in these texts. Organizational investigators are therefore increasingly seeking techniques to organize, classify, identify and extract views, experiences and feelings from the text (Pang & Lee, 2008; Wiebe, Wilson, & Cardie, 2005). Until recently, most text analyzes in organizations relied on time consumption and manual processes which were not feasible and less labour intensive.

A typical purpose of text analysis is to allocate text for present categories, similar to content (Duriau, Reger and Pfarrer, 2007; Hsieh and Shannon, 2005; Scharkow, 2013) and template analysis (Brooks, McCluskey, Turley and King, 2015). Assigning massive text collections to categories is expensive and, due to cognitive overload, it may become erroneous and unreliable. Furthermore, quirks among human coders may creep into the labeling process resulting in coding errors. One solution is to only code part of the body instead of coding all documents. This is to the extent that the important information is maybe omitted which might lead to a bias and decline in the internal or external validity of the results. University students rely increasingly on the World Wide Web (WWW) as an information source to complete course projects including electronic versions of

scholars' and popular media, however research has shown that modern students rely on this information without verification (Barclay, 2017; Metzger, Flanagin, & Zwarun, 2003). Kissel et al. (2016) find that the primary research sources for university students are the Internet, web searches, or magazines and less than 20% of students say they can assess a source's credibility.

Zhang (2016) bidirectional recurring neural networks (BRNN) are used for collecting both past and future data and for encapsulating local data with a convolutionary layer. In this context, the standard RNN has been replaced by two recent RNN changes, known as long-term short-term memory and gated recurring unit (GRU), to increase the efficiency of the new real-time classification architecture. The fundamental advantage is that the experimental model is fully trained and easy to implement without involving humans. Nema et al. (2015) system based on the algorithm of the ANTS and the method of cluster mapping. The ant colony optimization algorithm is used to optimize text data mapping for classification purposes. ACO classification result is better compared to the RSVM AND ML-FRC algorithm in all tree datasets webKB, Yahoo, Rcv1, along F1, BEP, and HLOSS.

In the study by Kulathunga (2017), they employed the method of Words sense disambiguation and assessed two algorithms Sequential Bottleneck Information and K means Algorithm. You used 446 documents from the EMMA repository and the categorized document with the status and purpose of a class label. The methodology proposed has demonstrated better results in terms of cluster purity. Gupta (2017), presents the survey of the various clustering techniques which have been examined and which reflect each algorithm's ascending and downside semantic relationship of the term for clustering not depending on its significance.

NLP was successfully used in extracting data from clinical electronic records and was applied to efficient and accurate chart abstraction in many areas. (Mendonca et al. 2005; Mendonca et al. 2001; Murff et al. 2011; Pakhomov et al. 2007; Salmasian et al. 2013) In some studies information on smoking has been automated as an isolated finding. (Stevens et al. 2013; Hazlehurst et al. 2005; Wicentowski et al. 2008; Lindholm et al. 2010) One study examined the increase in the NLP tool to identify cardiovascular risk. (Khalifa et al. 2014) Another study used the NLP tool MediClass for extracting weight management consulting information during postpartum visits and demonstrated similar extraction capabilities to human abstractors.

Hazlehurst (2014) and team separate study extracted documentation on lifestyle amendment in diabetes management assessment. (Hosomura et al. 2015) Moreover, several previous lifestyle modification studies have used survey data with potential bias and/or generalization concerns. To our understanding, this is the first work to assess changes in hypertension lifestyle in a wide population using automated methods.

In Pratama (2017), system automatically detects complaints by means of supervised and unattended teaching. The supervised learning was used in this regard to classify the subject of complaints, whereas unchecked learning is used for cluster Tweets based on the equivalence of the complaints. They evaluate the accuracy, precision, retrieval,

and F1 scores by using the assessment algorithms like SMO, Naive Bays, Multinomial and Random Forests. For single label classification and Random Forests for multilabel topic classification, SMO provides an average 95 percent precision with a 97,92F1 score. Unchecked learning is used in the evaluation by a clustering index value of the subject clusters.

In Pengfei (2016) neural network methods, several natural language processing tasks have achieved considerable progress. To learn about multiple associated tasks together, use the recurrent neural network framework. Text classification tasks show that the given model goes beyond the performance of a task with other related tasks.

Muhmmad (2018) neural network model, introduced with Bi-LSTM was used to learn deeply Arabic keyword extraction, to extract key phrases from Arabic text. They do not provide a large sufficient dataset, so they are constructing a new dataset of 6,000 abstracts of scientific documents in Arabic. The characteristics here are the same as the English datasets. Compared to the existing system as morphke, KP minor for wiki, the evolutionary result of this method is substantially better.

In the Pacifico (2018) system, the multi label text categorization model combines both the Rocchio algorithm and the Random Forest algorithm. These methods identify a correlation between data set training and the classification required only for related specifications filter. The precision achieved by the classification of hybrid text is slightly greater than individuals.

Behera and Kumaravelan (2019) provided the state-of-the-art ML techniques applied to four benchmark datasets in their paper. SGD, Ridge, Passive Aggressive, MLP, and SVM deliver the best results on the specified dataset when compared to the other classifiers. However, KNN, Rocchio, and Decision Tree classifiers performed poorly compared to other classifiers. To improve these classifiers' adaptability to large datasets, future work will focus on improving them. Deep learning models such multilayer feedforward neural networks, Convolution Neural Networks (CNN), Recurrent Neural Networks, or ensemble deep learning models are therefore a must for further study in this area.

It was determined by Vladimer (2017) and his team of researchers that Text Classiffier (TC) has enormous promise for making text-based organizational research fast, efficient and accurate. If you need to examine vast textual data, TC comes in handy. In some circumstances, it can even retrieve patterns that are impossible for people to notice. Manual qualitative text analysis approaches may be sufficient in the absence of automated methods. Since real issues and current theory can be used to model the development of new methodologies, the increased use of TC in organizational research is likely to benefit both organizational research and TC research.

Hugo's (2020) thesis examined the performance of neural networks versus traditional/classic statistical machine learning approaches in a document categorization setting using Swedish texts," the thesis states. The total performance of the neural network outperforms most classifiers when comparing different methods in traditional/classic statistical areas. This is evident from the test accuracy result. SVM

classifier, on the other hand, performs at the same high level of accuracy as the other performance metrics. There is no difference between NN and SVM when it comes to performance, even with marginal errors. Within the study context of document classification, there are a variety of approaches for varying the classification. Comparing machine learning algorithms and document categorization methods could be an intriguing study topic. The procedure of document classification is another important consideration.

Raghavendra (2018) and his team has utilized an automated text categorization strategy to classify Wikipedia corpora. Two supervised machine learning techniques were utilized in this study to classify Wikipedia articles. First, python packages are used to pre-process the raw Wikipedia articles before they are published. As a result, these pre-processed corpora are classified into separate groups using the pre-defined labels. Based on a kernel that is not linear It is possible to achieve great efficiency in text categorization by combining RBF and the linear kernel of SVM. Nearest K-neighbor linear support vector machine kernel Information about health and banking as well as news articles can be categorized. When it comes to classifying vast volumes of unstructured data, Raghavendra (2018) and his team propose using machine learning methods such as neural networks and decision trees in the future. Unsupervised and supervised learning can be combined to organize text documents in the future, and a large number of structured data sets can be used to test text categorization.

For users to find their chosen papers quickly and efficiently, Sang and Joon (2019) developed a paper classification system that supports the paper classification process efficiently and effectively. As part of the proposed approach, TF-IDF and LDA techniques are used to determine the relevance of each publication, and K-means clustering is used to group papers with related subjects. Users' interesting papers can be classified correctly as a consequence. When doing the experiments to show the performance of the proposed system, we used real-world data from the FGCS Journal. Using keywords taken from abstracts, the suggested system was able to classify papers with related topics. Work on research paper classification has been the primary emphasis of this project. A generic technique requires the effort to be expanded into a variety of datasets, such as documents and tweets, among other things. It is therefore planned to continue working with text mining datasets, as well as to construct even more efficient classifiers for research paper datasets, in the near future.

Research by Kannaiya and coworkers (2020) addressed multiple issues related to multilabel text classification by experimenting with different data sets, KNN algorithms and Naivebayes, feature transformations, as well as the incorporation of label space information. They also discussed the peculiarities of the results and attempted to explain them through an in-depth analysis of undesired results, which they called "unexpected". The KNN Algorithm had an F1-Score of 93 percent and a total precision of 87 percent, whereas the Nave Bayes Algorithm had an F1-Score of 95 percent and 94 percent, respectively. In terms of multilevel text classification, the nave Bayes model is the most effective.

The Reuters-21578 and 20 Newsgroups datasets were chosen by Asmaa and Alok (2020) as examples of our rule-based approach to classifying texts into ten categories. Information retrieval systems were expected to benefit from these results, and this work helped us establish acceptable methods for text classification based on precision, recall, and accuracy approaches. As well as improving the rule-based method, additional forms of machine learning were selected. A rule-based strategy that is more active when dealing with real-time datasets, such as newspaper datasets, was improved in the study. Categorizing content or goods on your website to make browsing easier or to help users find relevant stuff on your site Technological automation can be used by platforms such as E-commerce, news agencies and content curators.

Thailambal and Sheshasaayee (2019) present a model Author Keyword Weightage in Journal Ranking (AKWJR) to help academics select relevant publications from a pool of irrelevant ones. Many keyword ranking programs, such as RAKE and TEXTRANK, compare and rank author-annotated terms. For better outcomes, this research used citation value instead of journal selection because citations change depending on journal accessibility. The documents are accurately inserted, which prevents the researcher from tagging the wrong documents for a given query. In addition to reducing query time, AKWJR produces accurate results. When searching, researchers have restricted access to the complete document, thus this study work provides users with relevant documents that assist their work.

Demeke and Getamesay's (2021) study aimed to demonstrate how the proposed feature selection strategy increases classification accuracy. They ran a number of experiments and comparisons with existing approaches on two separate datasets with varying numbers of categories in order to validate the performance of our proposed feature selection method. This approach, paired with MLP classifier, produces good results. It is therefore appropriate to employ the suggested feature selection approach in multiple contexts, such as document organization and topic extraction in Amharic and information retrieval in Amharic. As the number of characteristics and categories rises, IGCHIDF might use some enhancements to reduce the decrease in classification accuracy that results. Additionally, we want to examine additional categories and datasets in the future. Integration of features is still a topic of interest for future research.

Existing System

Knowledge of customer views on airlines is very important, as customer feedback from Indian Airlines is very difficult to collect. In addition, it is crucial to understand how satisfied customers are for the company's strategy development (Sera et al. 2020). In Fig. 1, Sera et al. (2020) have used twitter dataset to analyse customer satisfaction.

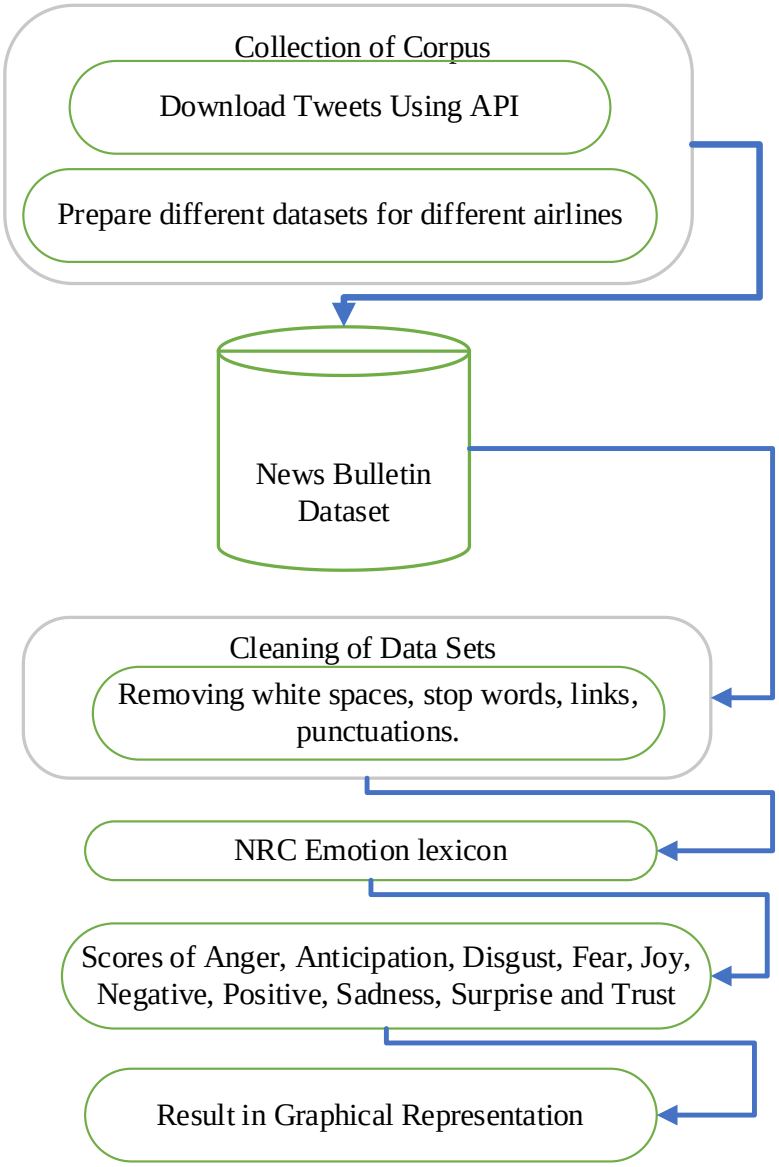

Figure 1: Existing Tweeter Airline Sentiment Analysis Architecture. Source: Sereja et al. (2020)

The study focused on the analysis of customern views on the services offered by Indian Airlines. The data dsiplay shows the feelings, anger, fear, anticipation, surprise, sorrow, joy and discontentment of customers. The views of airlines are obtained with the help of R programming for extraction and analysis purposes. Emotions like afraid, surprise, confidence, disgust, anticipation and rage apart from positive and negative ideas provide the company with useful insights into how it can satisfy its customers and customers. While further work should be carried out to analyze and differentiate positives and negatives, also in the study the level of precision obtained in the classification of Tweets has not been specified.

Methodology

The study in Fig. 2, utilized a text classification approach to natural language processing (NLP), using University of Califonia Irvin (UCI) Machine Learning Repository "News Headline dataset".

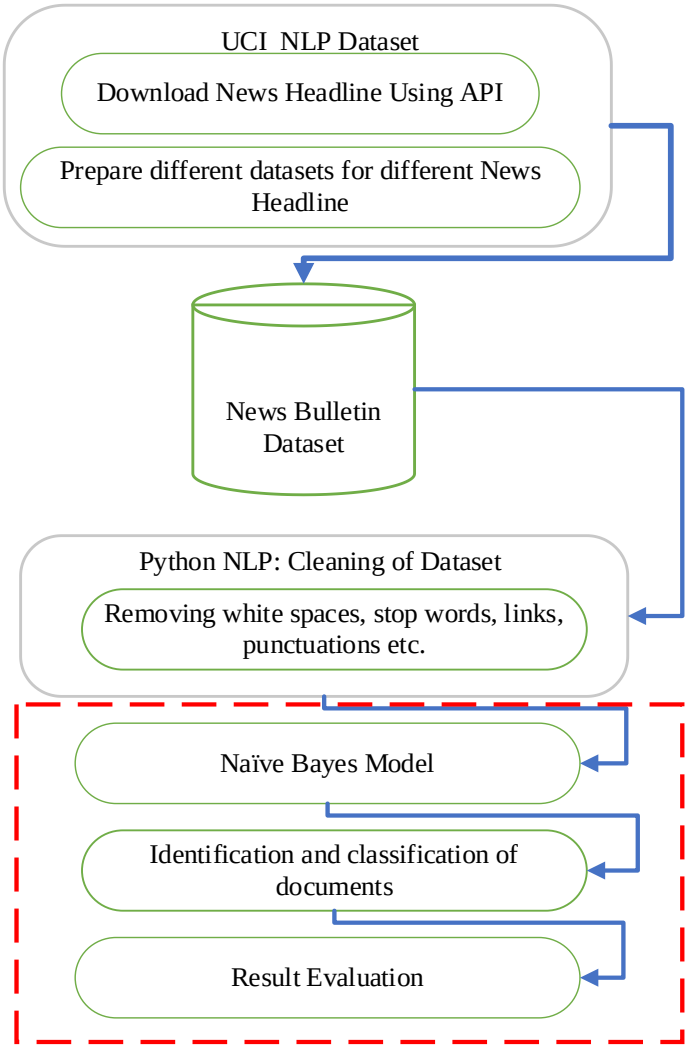

Figure 2: Enhanced Document Categorization Architecture

Both folders contained text movie review files labeled in relation to their overall feeling polarity (positive, negative) or subjective rating. The sentences are mixed with special symbols and URLs, but they are certain. The datasets were mixed but we submitted the classification with Randomforest in Fig. 6 through text mining.

Result and Discussions

There are several problems with the classification of text in connection with the approach to text-mining. The data collection process led to successful phases of pre-

processing, attribute abstraction, and ultimately text mining, was experiential. The type of data moulded collected with dissimilar pre-processing approaches with different investigation results in the Text pre-processing process. To make the dataset user-friendly, the dataset must be arranged via tokenisation, termination, stemming and space document. There are parameters we have considered for our experiments in order to extract features. The number of words called Frequency in machine learning is given in these parameters. In order to improve the text classification process, we selected these parameters for text classification.

The dataset utilized (Figure 3) was obtained from "UCI Machine Learning Repository" and consists of 93239 document observations and Six (6) variables, as listed below:

IDLink: is made up of the unique identification numbers of the documents that have been retrieved.

Title: The title is made up of the titles of the books.

Headline: The headline is a summary of the document that was retrieved.

Topic: is a variable that classifies documents and determines whether or not they are valid.

Document: indicates if the document is valid (Yes) or not, similar to the variable.

'Topic' (No): This is the target variable, which was originally included in the data.

Step 1: The first step is to load the libraries and modules that are required.

Step 2: Loading data and running basic data checks

Step 3: Get the raw text ready for machine learning by pre-processing it.

Step 4: Make the Training and Test Datasets

Step 5: TfidfVectorizer is used to convert text into word frequency vectors.

Step 6: Make the classifier and fit it.

```
(93239, 6)
```

	IDLink	Title	Headline	Source	Topic	Document
0	99248.0	Obama Lays Wreath at Arlington National Cemetery	Obama Lays Wreath at Arlington National Cemete...	USA TODAY	obama	NO
1	10423.0	A Look at the Health of the Chinese Economy	Tim Haywood, investment director business-unit...	Bloomberg	economy	YES
2	18828.0	Nouriel Roubini: Global Economy Not Back to 2008	Nouriel Roubini, NYU professor and chairman at...	Bloomberg	economy	YES
3	27788.0	Finland GDP Expands In Q4	Finland's economy expanded marginally in the t...	RTT News	economy	YES
4	27789.0	Tourism, govt spending buoys Thai economy in J...	Tourism and public spending continued to boost...	The Nation - Thailand's English news	economy	YES
5	27790.0	Intellitec Solutions to Host 13th Annual Sprin...	Over 100 attendees expected to see latest vers...	PRWeb	microsoft	YES

Figure 3: UCI Repository dataset

After loading the data in Figure 3, bar plot is used to check the distribution of the target class. Figure 4 clearly shows that the target variable has more occurrences of 'Yes' than 'No,' indicating a significant difference.

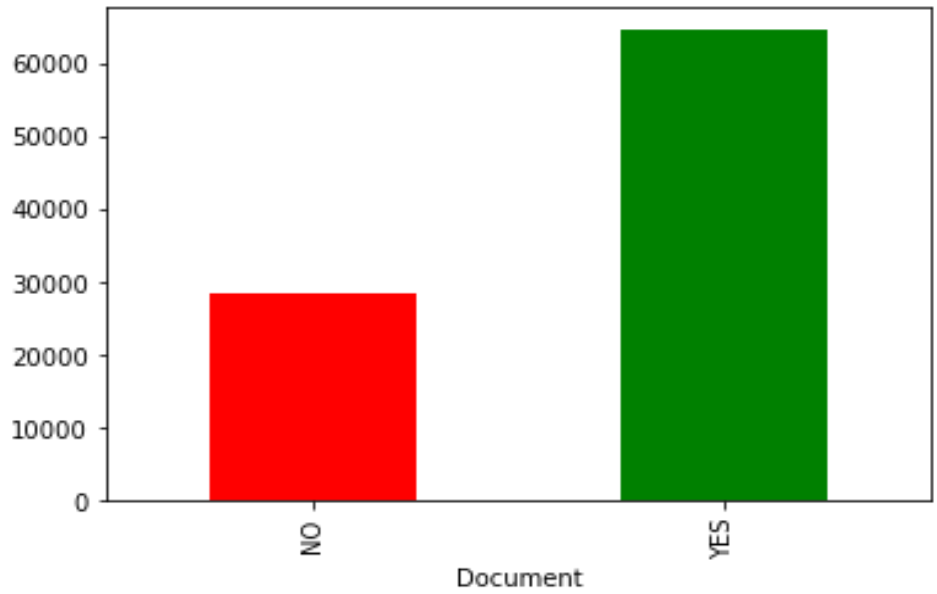

Figure 4: Distribution of the target class

The baseline accuracy is calculated to be 56%. It's calculated by dividing the total number of observations by the number of times the majority Document (i.e., 'No') appears in the target variable. The baseline accuracy is significant since it establishes the minimum accuracy that the model should attain. The dataset used to build the document classifier is raw text, which cannot be used as features on its own. As a result, the content must be pre-processed.

The following are some of the most common pre-processing steps:

Getting rid of punctuation: the general rule is to get rid of everything that isn't in the form x,y,z.

Removing Stopwords: Stopwords, such as 'the,' 'is,' and 'at,' should be eliminated. These aren't useful because the number of stopwords in the corpus is high, yet they don't help distinguish target groups. Stopwords are removed, which decreases the data size.

Conversion to lower case: Words like 'Document' and 'document' must be viewed as one word when converted to lower case. As a result, they've been transformed to lowercase.

Stemming: is the process of reducing the number of inflectional versions of words that exist in a text. This reduces terms like "argue," "argued," "arguing," and "argues" to their common stem "argu." This contributes to the reduction of vocabulary space. Stemming can be done in a variety of ways, the most prominent of which being Martin Porter's "Porter Stemmer" approach.

The load of the python natural language toolkit (nltk) module is required to complete the above-mentioned stages.

The preprocessed in Figure 5, dataset has a new column called 'processedtext' once it has been preprocessed.

```
(93239, 7)
```

	IDLink	Title	Headline	Source	Topic	Document	processedtext
0	99248.0	Obama Lays Wreath at Arlington National Cemetery	Obama Lays Wreath at Arlington National Cemete...	USA TODAY	obama	NO	obama lay wreath arlington nation cemeteri
1	10423.0	A Look at the Health of the Chinese Economy	Tim Haywood, investment director business-unit...	Bloomberg	economy	YES	a look health chines economi
2	18828.0	Nouriel Roubini: Global Economy Not Back to 2008	Nouriel Roubini, NYU professor and chairman at...	Bloomberg	economy	YES	nouriel roubini global economi not back
3	27788.0	Finland GDP Expands In Q4	Finland's economy expanded marginally in the t...	RTT News	economy	YES	finland gdp expand in q
4	27789.0	Tourism, govt spending buoys Thai economy in J...	Tourism and public spending continued to boost...	The Nation - Thailand's English news	economy	YES	tourism govt spend buoy thai economi januari
5	27790.0	Intellitec Solutions to Host 13th Annual Sprin...	Over 100 attendees expected to see latest vers...	PRWeb	microsoft	YES	intellitec solut host th annual spring microso...
6	80690.0	Monday, 29 Feb 2016	RAMALLAH, February 25, 2016 (WAFA) - Palestine...	NaN	palestine	YES	monday feb

Figure 5: Pre-processed Dataset

Converting text to word frequency vectors is required after text processing in order to develop machine learning models. There are a few options, such employing CountVectorizer and HashingVectorizer, but TfidfVectorizer is the most popular.

Term Frequency-Inverse Document Frequency (TF-IDF) is an abbreviation that stands for "Term Frequency-Inverse Document Frequency." In text mining applications, it is employed as a weighting factor.

Term Frequency (TF): This summarizes the document's normalized Term Frequency.

IDF (Inverse Document Frequency): This method minimizes the weight of terms that appear frequently in documents. To put it another way, TF-IDF tries to highlight essential words that appear frequently in a document but do not appear in other documents. For our documents, we'll focus on developing TF-IDF vectors.

The Naive Bayes Classifier algorithm was employed. Figure 6 and 7, to compute accuracy, the 'metrics.accuracy score' function was used. Using test data to evaluate the model's performance. The accuracy of the initial baseline was 55.5 percent.

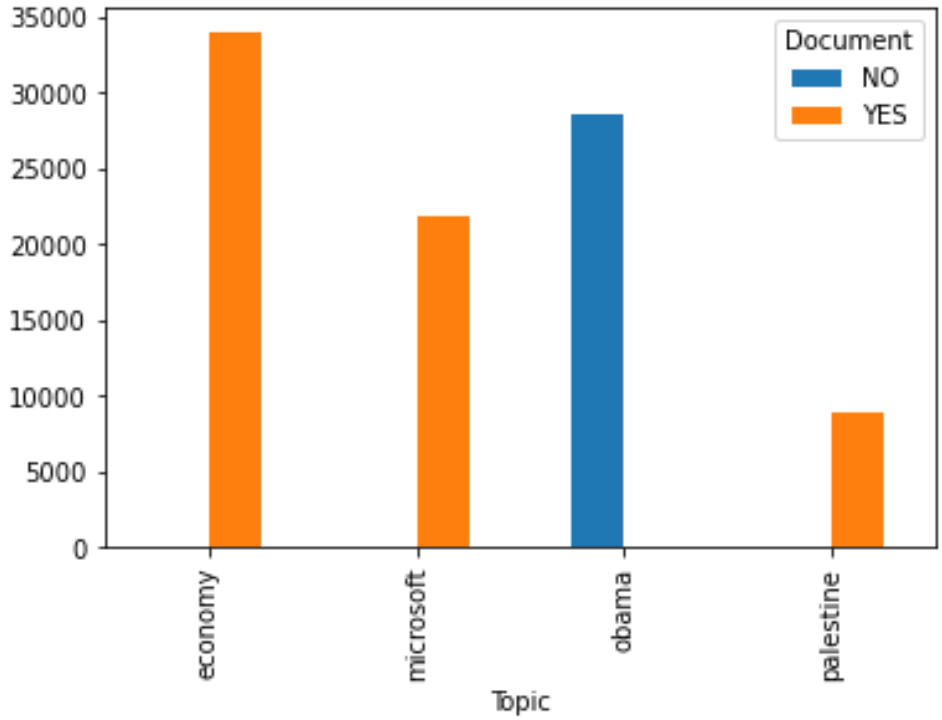

Figure 6: Classified Document

Our Naive Bayes model easily outperforms the baseline model, with an accuracy score of 95.5 percent. This also establishes a new standard for any future algorithm or model improvements.

```
# Calculate the accuracy score: score
accuracy_tfidf = metrics.accuracy_score(y_test, pred2)
print(accuracy_tfidf)
```

```
0.9525239525239525
```

Figure 7: Performance metric

Conclusion

In reviewing the relevant works, we concluded that text classification applications are rapidly increasing in the field of IT and that assessment can be increased if the text of the learning is increased during experiments and simulations. These attributes increase the number of terms that are automatically made up of a good sample categorization class. The main objective of this study is to classify document, to select functions and to evaluate performance using the Python programming language as an experimental tool.

In a supervised master training approach, we extract the textual characteristics from the English language documents. The assessment and comparison were done using a selection of parameters such as accuracy percentage. Finally, the performance of our selected machine learning technique achieved 95.5% on the chosen dataset. It is very clear from our discussion that every technique for learning a machine has its advantages and disadvantages according to the size of the datasets.

Contribution to Knowledge

The study was able to employ the python toolkit for text analysis to categorize the document. A systematic processing of texts utilizing naive bayes supervised machine learning technique for accuracy in the categorisation of documents.

Further Research

The study centred on the use python natural language toolkit in optimising the categorisation of documents. With a supervised machine learning approach, we extracted the text features from the English-based documents. The simulations show very clearly that the strategies of the Naiye Baye machine to learn are extremely precise for the utilization of the data set. It is extremely evident from our explanation that any dataset is capable of performing differently than a certain machinery, as such a deeper examination of other methodologies is urged.

References

ACM SIGMOD Record, 37(3), 40-45. Retrieved from https://sigmodrecord.org/publications/sigmodRecord/0809/p40.jagadish.pdf.

Asmaa M. Aubaid and Alok Mishra. (2020). A Rule-Based Approach to Embedding Techniques for Text Document Classification.

Barclay, D. A. (2017, January 4). The challenge facing libraries in an era of fake news. The Conversation. Retrieved from https://theconversation.com/the-challenge-facing-libraries-in-an-era-of-fake-news-70828 9591.

Berry, M. W., Castellanos, M. (2008). Survey of text mining II—Clustering, classification, and retrieval. Retrieved from http://www.springer.com/gp/book/9781848000452

Berry, M. W., Castellanos, M. (2008). Survey of text mining II—Clustering, classification, and retrieval. Retrieved from http://www.springer.com/gp/book/9781848000452

Bichitrananda Behera, Kumaravelan G. (2019). Towards the Deployment of Machine Learning Solutions for Document Classification.

Brooks, J., McCluskey, S., Turley, E., King, N. (2015). The utility of template analysis in qualitative psychology research. Qualitative Research in Psychology, 12(2), 202–222.

Cardie, C., Wilkerson, J. (2008). Text annotation for political science research. Journal of Information Technology & Politics, 5(1), 1–6.

Demeke Endalie and Getamesay Haile. (2021). Hybrid Feature Selection for Amharic News Document Classification.

Duriau, V. J., Reger, R. K., Pfarrer, M. D. (2007). A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements. Organizational Research Methods, 10(1), 5–34.

Gupta, Aditi, Jyoti Gautam, & Ajay Kumar. (2017). A survey on methodologies used for semantic document clustering. International conference on Energy, Communication, data analytics and Soft computing (ICECDS). IEEE.

Hazlehurst B., Frost R. & Sitting D. (2005). MediClass: a system for detecting and classifying encounter based clinical events in any electronic medical record. J Am Med Inform Assoc.

Hazlehurst BL., Lawrence JM. & Donahoo WT. (2014). Automating assessment of lifestyle counseling in electronic health records. Am J Prev Med.

Hosomura N, Goldberg SI & Shubina M. (2015. Electronic documentation of lifestyle counselling and glycemic control in patients with diabetes. Diabetes Care.

Hsieh, H.-F., Shannon, S. E. (2005). Three approaches to qualitative content analysis. Qualitative Health Research, 15(9), 1277–1288.

Hugo Moritz. (2020). A comparative study of machine learning algorithms for Document Classification.

Jagadish, H. V. (2008). The conference reviewing crisis and a proposed solution.

Jamieson, S. (2016). What the citation project tells us about information literacy in college composition. In B. J. D'Angelo, S. Jamieson, B. Maid, & J. R. Walker (Eds.), Information literacy: Research and collaboration across disciplines (pp. 115-138).

Kannaiya Raja N., Naol Bakala & Barani Sundaram B. (2020). Multi-Label Classification for Afan Oromo Text by using Python Machine Learning Tools.

Khalifa A. & Meystre S. (2015). Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. J Biomed Inform.

Khalifa A. & Meystre S. (2015). Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. J Biomed Inform.

Kissel, F., Wininger, M. R., Weeden, S. R., Wittberg, P. A., Halverson, R. S., Lacy, M., & Huisman, R. K. (2016). Bridging the gaps: Collaborating in a faculty and librarian community of practice on information literacy. In B. J. D'Angelo, S. Jamieson, B. Maid, & J. R. Walker (Eds.), Information literacy: Research and collaboration across disciplines (pp. 411-428).

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., Den Hartog, D. N. (2018). Text mining in organizational research. Organizational Research Methods, 21(3), 733–765.

Kulathunga, Chalitha, & Karunaratne D. D. (2017). An ontology-based and domain-specific clustering methodology for financial documents, advances in ICT for Emerging Regions (ICTER), 2017 Seventeenth International Conference on. IEEE.

Lindholm C, Adsit R. & Bain P. (2010). A demonstration project for using the electronic health record to identify and treat tobacco users. WMJ.

Mendonca EA, Haas J. & Shagina L. (2005). Extracting information on pneumonia in infants using natural language processing of radiology reports. J Biomed Inform

Mendonca EA., Cimino JJ. & Johnson S. (2001). Using narrative reports to support a digital library. Proc AMIA Symp.

Metzger, M. J., Flanagin, A. J., & Zwarun, L. (2003). College student Web use, perceptions of information credibility, and verification behavior. Computers & Education, 41(3), 271-290.

Muhmmad Hemly, R. M. Vigneshwaram, Giuseppe Serra & CarloTasso. (2018). Applying Deep Learning for Abarbic Key Phrases Extraction "2018, The 4th International Conference on Arabic Computational Linguistics (ACLing 2018), November 17-19 2018, Dubai, United Arab Emirates. Published by Elsevier B.

Murff HJ, FitzHenry F. & Matheny ME. (2011). Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA.

Nema, Puneet & Vivek Sharma (2015). Multi-label text categorization based on feature optimization using ant colony optimization and relevance clustering technique. Computers, Communications, and Systems (ICCCS), International Conference on. IEEE.

Pacifico LD., Macario V. & Oliveira JF. (2018). Plant Classification Using Artificial Neural Networks. In2018 International Joint Conference on Neural Networks (IJCNN), IEEE.

Pakhomov S., Weston SA. & Jacobsen SJ. (2007). Electronic medical records for clinical research: application to the identification of heart failure. Am J Manag Care.

Pang, B., Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1–135.

Pengfei Liu, Xipeng Qiu & Xuanjing Huang. (2016). Recurrent Neural Network for Text Classification with Multi-Task Learning. (IJCAI).

Pratama, Timothy, & Ayu Purwarianti. (2017). Topic classification and clustering on Indonesian complaint tweets for Bandung government using supervised and unsupervised learning. International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA), IEEE**.**

Raghavendra S., Lingaraju G. M., Shekar sivasubramanian. (2018). Supervised Learning Based System for Classification of Wikipedia Articles.

Salmasian H, Freedberg DE & Friedman C. (2013). Deriving comorbidities from medical records using natural language processing. J Am Med Inform Assoc.

Sang-Woon Kim & Joon-Min Gil. (2019). Research paper classification systems based on TF-IDF and LDA schemes.

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. Quality & Quantity, 47(2), 761–773.

SreeDevi J., Rama Bai M., Chandrashekar Reddy M. (2020). Newspaper Article Classification using Machine Learning Techniques

Sreeja I, Joel V Sunny & Loveneet Jatian. (2020). Twitter Sentiment Analysis on Airline Tweets in India Using R Language. Journal of Physics: Conference Series.

Stevens V, Bailey S. & Hazlehurst B. (2013). PS1-1: use of CER hub to evaluate outcomes of smoking cessation services, a behavioral treatment. Clin Med Res.

Thailambal G., Ananthi Sheshasaayee. (2019). A Different Text Mining Process for Classifying Journal Databases Using Machine Learning Algorithms.

Vladimer B. Kobayashi, Stefan T. Mol, Hannah A. Berkers, Gábor Kismihók, Deanne N. Den Hartog. (2017). Text Classification for Organizational Researchers: A Tutorial.

Wicentowski R & Sydes MR. (2008). Using implicit information to identify smoking status in smoke-blind medical discharge summaries. J Am Med Inform Assoc.

Wiebe, J., Wilson, T., Cardie, C. (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39(2-3), 165–210. https://doi.org/10.1007/s10579-005-7880-9

Yaakov HaCohen-Kerner, Daniel Miller & Yair Yigal. (2020). The influence of preprocessing on text classification using a bag-of-words representation

Zhang Y, Er MJ, Venkatesan R, Wang N & Pratama M. (2016). Sentiment classification using comprehensive attention recurrent models. neural Networks (IJCNN), 2016 International Joint Conference. IEEE.