Digital Innovations
& Contemporary Research in SCIENCE, ENGINEERING & TECHNOLOGY

Vol. 10. No. 4, December 2022 – Special Issue on Cryptography & Security Mechanisms

Article Citation Format
Obiorah, P., Onuodu, F. & Eke, B. (2022):
Topic Modeling Using Latent Dirichlet Allocation & Multinomial
Logistic Regression. Journal of Digital Innovations &
Contemporary Research in Science, Engineering & Technology.
Vol. 10, No. 4. Pp 99-112
DOI: dx.doi.org/10.22624/AIMS/DIGITAL/V10N4P11a

Topic Modeling Using Latent Dirichlet Allocation & Multinomial Logistic Regression

Obiorah Philip, Onuodu Friday & Eke Batholowmeo
Department of Computer Science
University of Port Harcourt
Choba-Port Harcourt, Nigeria
E-mails: philip.obiorah@outlook.com; friday.onuodu@uniport.edu.ng;
bartholomew.eke@uniport.edu.ng

ABSTRACT

Unsupervised categorization for datasets has benefits, but not without a few difficulties. Unsupervised algorithms cluster groups of documents in an unsupervised fashion, and often output findings as vectors containing distributions of words clustered according to their probability of occurring together. Additionally, this technique requires human or domain expert interpretation in order to correctly identify clusters of words as belonging to a certain topic. We propose combining Latent Dirichlet Allocation (LDA) with multi-class Logistic Regression for topic modelling as a multi-step classification process in order to extract and classify topics from unseen texts without relying on human labelling or domain expert interpretation in order to correctly identify clusters of words as belonging to a certain topic. The findings suggest that the two procedures were complementary in terms of identifying textual subjects and overcoming the difficulty of comprehending the array of topics from the output of LDA.

Keywords: Natural Language Processing; Topic Modeling; Latent Dirichlet Allocation; Logistic Regression

1. INTRODUCTION

Due to its extensive capabilities, topic modelling has recently become more popular in data mining. When applying such topic models to different activities, one of the most prevalent and difficult challenges is determining the precise meaning of each topic. Unsupervised classification for datasets has some benefits, and challenges as well. Unsupervised algorithms cluster groups of documents in an unsupervised manner, and they often provide results in the form of vectors containing distributions of words, grouped according to their likelihood of appearing in the same set of documents. As a result, human or domain expert interpretation is required in order for this approach to accurately identify clusters of words belonging to a certain subject.(Kotu and Deshpande, 2022)

There are two distinct ways of detecting topics: supervised and unsupervised. Domain experts are required to train text documents on defined conceptual subjects and then supervised approaches can be used to make predictions on topic labels for unknown data items. On the other hand, unsupervised approaches classify text documents into discrete groups based on their content similarities without requiring domain experts to get documents with the same or equivalent subjects. (Hui, Dong, & He, 2016) .Latent Dirichlet Allocation (LDA) is a technique of unsupervised learning that does not require manually labelled data. The latent Dirichlet allocation (LDA) model of a corpus is a probabilistic generative model (Blie, Ng, and Jordan, 2003). LDA presupposes the creation of documents based on a range of topics. After that, the topics produce words depending on their likelihood of dissemination. Given a collection of documents, LDA recursively attempts to deduce the topics that created those documents in the first place. The training of LDA (as well as other unsupervised topic models) is truly independent of supervisory information since the goal of LDA is to infer the optimal set of latent topics that may explain the document collection rather than divide various classes. With this paradigm, classification tasks are severely restricted (Guo et al., 2009).

However, even though LDA clusters similar groups of documents together in an unsupervised manner, we still require the assistance of a domain expert to identify the hidden topic for which LDA produces a vector showing distributions as a cluster of words based on the likelihood of words occurring together. In response to this issue, we presented a strategy in which the unsupervised output of LDA models is utilized as an input to a classification model, which was successfully implemented. In this scenario, Logistic Regression is useful in training a classifier that ingests the LDA output that has not been provided by the LDA. Logistic regression evaluates the connection between several independent factors and a categorical dependent variable and calculates the likelihood of an event by fitting data to a logistic curve (Park, Hyeoun-Ae, 2013). Logistic regression translates its result into a probability value using the logistic sigmoid function. Several of the major features of LR include its ability to automatically generate probabilities and its extension to multi-class classification problems (Maalouf, 2011). This paper proposes combining Latent Dirichlet Allocation (LDA) with Multinomial Logistic Regression for topic modeling as a multi-step classification process in order to extract and classify topics from unseen texts without relying on human labeling or domain expert interpretation in order to correctly identify clusters of words as belonging to a certain topic.

2. LITERATURE REVIEW

Topic modeling is a method for automatically detecting topics within a text item and extracting hidden patterns from a corpus of text. As a result, making decisions becomes simpler. A few of the most often used topic modeling methods have been Latent Semantic Analysis (LSA), Non-Negative Matrix Factorization (NNMF), Probabilistic LSA (PLSA), and Latent Dirichlet Allocation (LDA) (Kherwa et al. 2018) To extract topics and summarize a document corpus, Blei et al.(2003) introduced the Latent Dirichlet Allocation

2.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is an unsupervised probabilistic generative model for discovering latent semantic topics in a corpus using huge sets of discrete data, such as the words in a set of articles(Blei et al. 2003). LDA is a topic model that generates topics from a collection of documents based on word frequency.

LDA is notably useful for determining reasonably accurate topic mixtures within a given document. The basic idea is that documents are represented as random mixtures of latent topics, with each topic defined by a word distribution (Blei et al. 2003).

The LDA is based on a "bag of words". The premise that each document is a frequency of word counts, regardless of its sequence of occurrence. Words are distributed independently and equally among topics, and topics are interchangeable indefinitely throughout the document. LDA assumes that documents are created on a variety of subjects. Then, depending on their probability distribution, these subjects produce words. Given a dataset of documents, LDA recursively attempts to determine the subjects that generated those documents in the first place. It employs dirichlet priors for the document- and word-topic distributions, allowing for improved generalization (Xu, 2018). The basic concept behind LDA is to produce a discrete distribution of words within a topic and discrete distribution of words across topics within a text. While LDA is creative enough to expose subjects inside a text, it lacks a mechanism for labeling its learning method (Basher, & Fung 2014)

The input of LDA comprises (1). Multiple Documents (*Corpus),* (2) *k* - The number of topics we want to find from our documents, while the output of LDA includes (1) Topics (2)Topic Distributions. The LDA process is as follows: For each document, LDA tries to guess the topics and topic distributions and then creates a fake document using that guessed information and the words from the original document. It then compares the fake document to the input document to see if they match if the fake document matched the input document then the guess is correct and we have found the correct topics and topic distributions.

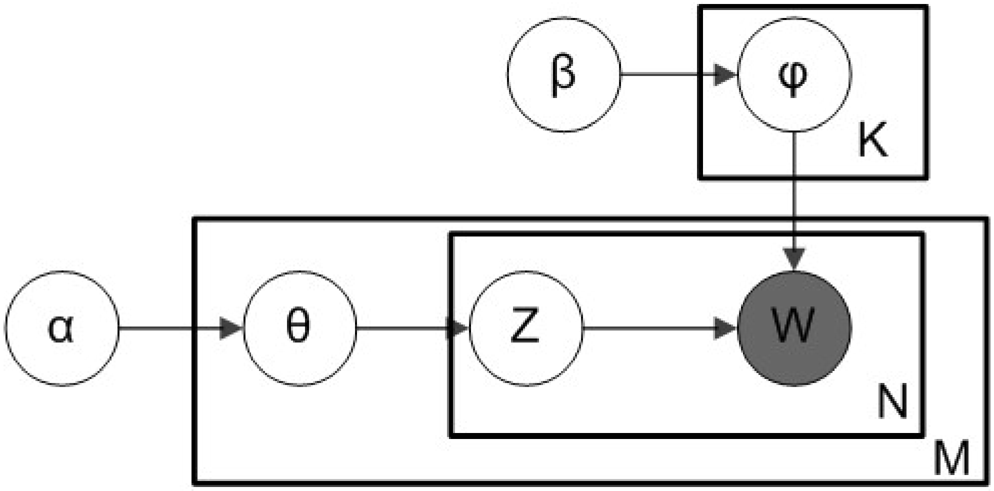


Figure 2.13 Model of Latent Dirichlet Allocation(LDA) (Source: Xu, 2018)

Where
- N is the document's word count.
- M is the quantity of documents to be analyzed
- α is the Dirihlet-prior parameter defining the concentration of the per-document topic distribution
β is identical to the per-topic word distribution parameter.
φ(k) is the term used to describe the spread of a document i

z(i,j) is the topic assigned for w(i,j)
w(i,j) is the j-th word in the i-th document
φ and θ are Dirichlet distributions, z and w are both polynomials.

Alpha and Beta Hyperparameters — alpha denotes the density of document topics, whereas beta denotes the density of topic words. With a higher alpha value, documents include more topics; with a lower alpha value, documents have fewer topics. On the other hand, topics with a high beta value are composed of a big number of words in the corpus, while those with a low beta value are formed of a few words.

2.3 Multinomial Logistic Regression

Multinomial Logistic Regression is a regression model that generalises logistic regression to classification problems with more than two possible outcomes. Multinomial logistic regression is used to predict categorical placement in or the probability of category membership on a dependent variable based on multiple independent variables. The independent variables can be dichotomous (i.e., binary) or continuous (i.e., interval or ratio in scale). Multinomial logistic regression is a simple extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. Multinomial logistic regression, like binary logistic regression, employs maximum likelihood estimation to determine the likelihood of categorical membership. (Starkweather & Moske, 2011). A multinomial regression model is a model with multiple equations. The multinomial regression model estimates k-1 logit equations for a nominal dependent variable with k categories.

2.3 Related Works

Rahman et al. (2013) described a topic search approach for evaluating chat log archives in order to identify logs containing criminal activity. They proposed an extension of the Latent Dirichlet Allocation (LDA) model for extracting topics, calculating the authors' contributions to these topics, and examining the topics' temporal transitions. The outcome reveals that this is crucial for research since it sheds light on the authors' engagement with certain issues. Experiments on two real-world datasets demonstrate that the proposed method is capable of detecting concealed criminal subjects and their author distribution (Rahman et al., 2020).

Dai et al. (2012) focused on topic discovery from tweet replies between Twitter users in order to bring a novel way to studying personal subject interests in tweets between pairs of persons and suggested a generative model for topic discovery among groups of Twitter users. Experiments on a collected dataset of tweets demonstrate that this model is good at detecting subjects in tweet replies. The model is generalizable to various forms of social communication, as long as the premise of a single topic per message stays true. It is not sufficient to just consider twitter messages as documents in order to find the topics of their tweets. As Dai et al. (2012) notes it is thus necessary to take into account both the talking party (or the sender) and the listening party (or the recipient) when determining the subject of twitter responses or any other kind of communication

Bruggermann et al. (2016) examined the problem of finding and tracking stories across time by analyzing news text corpora. They devised a method that employs a dynamic form of Latent Dirichlet Allocation (DLDA) over discrete time steps and allows for the identification and tracking of topics within stories as they occur. The dynamic version (DLDA) enables the analysis of topic distributions throughout time, revealing their changes and evolutions. However, the limitation of this method is the absence of a formal mechanism for grouping related topics into a plot. Additionally, the DLDA's output, which consists of subjects, is unlabeled.

Thielmann et al. (2021) illustrate the combination of web scraping, one-class Support Vector Machines (SVM), and Latent Dirichlet Allocation (LDA) topic modeling as a multi-step classification strategy that avoids human labeling. The findings indicate that unsupervised document classification implemented using web scraping, one-class Support Vector Machines, and Latent Dirichlet Allocation topic modeling produces very accurate classification results for a variety of data sets. However, this strategy would still need the assistance of a domain expert to identify the LDA topics retrieved.

From the foregoing, in the majority of current research on topic modeling, researchers either use the top terms in the distribution as primitive labels (Blei et al., 2003; Ramage et al., 2009; Ramage et al., 2011) or manually develop more relevant labels in a subjective fashion (Mei et al., 2006; Mei and Zhai, 2005). However, identifying top words is insufficient to comprehend a topic's cohesive meaning (Mei et al., 2007). Meanwhile, manually created labels involve considerable human work and are prone to skew toward the user's subjective beliefs. (Mao et.al., 2016). Thus, to extract and classify topics from unseen texts, we propose combining Latent Dirichlet Allocation (LDA) with Multinomial Logistic Regression for topic modeling as a multi-step classification process. This eliminates the need for manual topic extraction and classification, as well as the need for human labeling or domain expert interpretation, and allows for more accurate identification of clusters of words as belonging to specific topics.

3. MATERIALS AND METHOD

3.1. Dataset

Data scrapped from Narialand Forum website (source: https://www.nairaland.com/) form 2016-2022 and the 2004-2005 BBC news dataset (source: http://mlg.ucd.ie/datasets/bbc.html) has been used for this experiment. The Narialand dataset comprises of 25785 while the BBC news website dataset consists of 26512 documents relating to articles in five thematic categories.

3.2. System Architecture

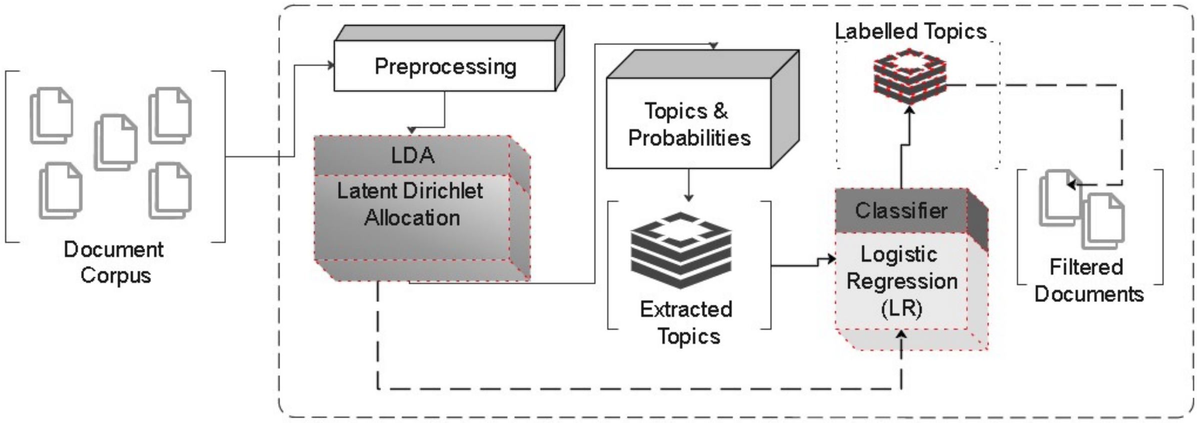


Figure 3.1: Architecture of the Proposed Model

Figure 3.2 shows an Architecture of The Proposed System and how it interacts with different components. The Document Corpus comprise of the dataset scrapped from the Nairaland website. The Pre-processing stage of the proposed system include the following processes. *Noise Removal* - which comprises deleting unneeded text segments by stopword deletion, as well as handling with capitalization, characters, and other minor issues. *Stop-Words Removal* - We omit words with fewer than three characters. The terms that are most often used in a language. In English, stop words include "a", "the", "is", and "are".We also extended the stop words to include Nigerian Pidgin stop-words such as 'na', im', 'dey', 'na', 'us', 'get', 'go'. . *Lemmantization*: Third-person pronouns were converted to first-person pronouns and past and future tense verbs to present tense. *Tokenization*.- The text is broken down into sentences, and the sentences are further broken down into words. We synthesize words from a continuous stream of text.

Simple-to-use tokenization and feature extraction algorithms are included in the scikit-learn library. We utilize the CountVectorizer class from the scikit-learn library to tokenize and produce a vocabulary of often used terms, as well as to encode new documents using that vocabulary. *Lowercasing:* To aid in the preparation and subsequent stages of the natural language processing application's parsing, we convert all text data to lowercase. After preprocessing LDA model is then applied. The output of LDA results to Topic & Topic Probalities or distributions. At this point the topics and topic probabilities are used as input to a multinominal Logistic classifier which would result in Labelled topics that would eventually be useful to categorize unseen documents.

3.3. LDA Use Case

Technically, the LDA model presupposes that topics are established prior to data generation. LDA views documents as mixtures of topics that spew out words based on their probability of occurrence. LDA attempts to guess the topics and topic distributions for each document, then creates a fake document using that guessed information and the words from the original document. The fake document is then compared to the input document to see if they match.

If the fake document corresponded to the input document, your guess was correct, and you discovered the correct topics and topic distributions. In this example, we have a collection of documents related to the post from a chat group. We are interested in identifying underlying topics that structure the collection. Each document, we suppose, has a variety of diverse topics. Additionally, we believe that a topic may be thought of as a set of terms with varying probabilities of appearing in a post debating the topic. For instance, one topic may contain references to "mercy," "lord," "thank," and "pray." Consider the following scenario: we have a corpus of documents from which we desire to examine 'K' number of topics. We will thoroughly examine each document (d). Following that, the following will be computed for each word (w) in the document: - X = p(topic | document) = the proportion of words in document d that are categorized as belonging to topic (t) at the moment. Y = p(word w | topic t) is the proportion of assignments to topic (t) in all documents containing the word w. (Because the same term may exist in many documents, it has a wide scope across them.) Because X * Y is basically the likelihood that topic 't' caused the word 'w,' it makes sense to utilize this probability to resample the current word's topic.

After a significant number of repetitions of these procedures, we should have pretty accurate topic assignments that create descriptive terms for the text. We assume that all topic assignments are valid except for the current word in question, and we use our model to update the assignment of the current word on each iteration.

3.3 . 1 LDA Sample Output

LDA sample output is essentially comprised of (1) Topics (2) Topic Distributions. (1) A topic is a group of semantically similar words. For instance Topic 1 : comprise ['buhari', 'jonathan', 'just', 'governor', 'vote', 'dickson', 'bello', 'bayelsa', 'party', 'kogi', 'people', 'election', 'state', 'pdp', 'apc'] . Topic 4 consist of ['hate', 'make', 'said', 'jesus', 'did', 'country', 'life', 'man', 'like', 'nigeria', 'just', 'don', 'know', 'people', 'god']. (2)-A topic distribution- is how much a document is made up of each topic. 50% of a documents can be about topic 1 and 70% of other documents can be about topic 2

Table 3.1 Showing hightest probability words per topic

TOPIC #1	['buhari', 'jonathan', 'just', 'governor', 'vote', 'dickson', 'bello', 'bayelsa', 'party', 'kogi', 'people', 'election', 'state', 'pdp', 'apc']
TOPIC #2	['win', 'election', 'apc', 'bello', 'polling', 'just', 'lol', 'man', 'kogi', 'lga', 'result', 'unit', 'results', 'votes', 'dino']
TOPIC #3	['time', 'rice', 'know', 'south', 'come', 'say', 'make', 'money', 'nigeria', 'business', 'like', 'just', 'don', 'dey', 'na']
TOPIC #4	['hate', 'make', 'said', 'jesus', 'did', 'country', 'life', 'man', 'like', 'nigeria', 'just', 'don', 'know', 'people', 'god']

We can see the first and second topic group seems to have identified word co-occurrences for Nigerian politics, and the third topic group seems to have identified business experiences. The fourth topic is not clear-cut but generally seems to touch on religion. Another challenging task is analyzing the findings of LDA. The "topics" simply spit out a random string of words. This is one of the algorithm's disadvantages, as it requires extensive trial and error. It's difficult to derive insights from outputs. Often, we would require feedback from a domain expert to determine whether the results are better suited for a particular use case or not. We present a method for training a supervised model on the output of LDA.

As a result, we are able to automatically classify latent themes into their appropriate categories. Our classifier is built using Multinomial Logistic Regression. Multinomial logistic regression is used to determine the probability of a variable (y) being in more than two classes; we want to know the chance of y being in each possible class. After passing through the classifier, the final output of our model would be a collection of Label topics. The purpose of this step is to eliminate the requirement for a domain expert to intervene and determine if a topic is related to a specific subject or not. An automated technique for labeling or categorizing the retrieved subjects becomes critical in the context of the proposed system. In this scenario, the classifier's output is a list of labelled themes that may be used to filter documents.

3.3. Process Diagram

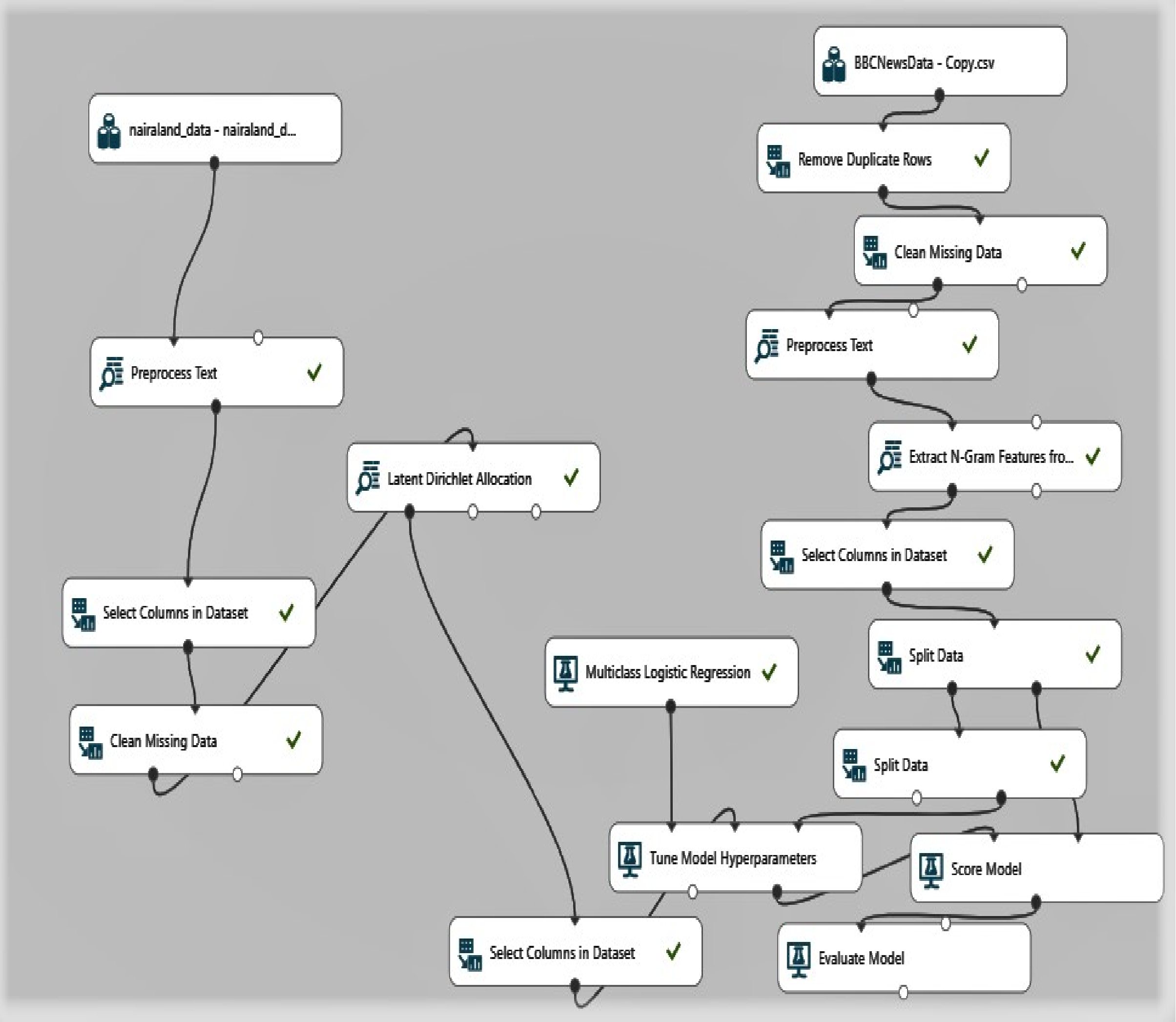

BBCNewsData - Copy.csv
nairaland_data - nairaland_d...
Remove Duplicate Rows ✓
Clean Missing Data ✓
Preprocess Text ✓
Preprocess Text ✓
Extract N-Gram Features fro... ✓
Latent Dirichlet Allocation ✓
Select Columns in Dataset ✓
Select Columns in Dataset ✓
Split Data ✓
Multiclass Logistic Regression ✓
Clean Missing Data ✓
Split Data ✓
Tune Model Hyperparameters
Score Model
Select Columns in Dataset ✓
Evaluate Model

Figure 3.3: Process Diagram

The design process is shown in Figure 3.3. We import the Narialand dataset, identify sentences, remove stop words, do lemmatization and lowercasing, remove numerals, remove special characters, email addresses, URLs, extend verb contractions, convert backslashes to slashes, and divide tokens on special characters. Following that, we choose columns of interest from the dataset, in this instance the pre-processed text. The dataset is then cleaned of missing data, then LDA is performed.

3.4 Classifier

In order to categorize the output of LDA we utilize the BBC dataset to train our Multinomial Logistic Regression model classifier which allows us to categorize the output of our LDA into numerous categories. The multinomial logistic regression model was trained using label datasets from the following industries: entertainment, business, politics, sports, and technology. Our goal is to create a system that can accurately categorize previously unseen posts or messages. We do the necessary pre-processing and turn each remark or post to a feature vector, a list of numerical values indicating some of the text's qualities, as machine learning models cannot analyze raw text and must instead deal with numerical values, as we did with the Nairaland dataset. In this scenario, we employ the bag of words model, in which the presence and frequency of words are considered for each post or remark, but the order in which they appear is ignored. Specifically, for each term in our dataset, we will compute Term Frequency and Inverse Document Frequency (tf-idf). In this case, we consider a word's frequency as a proxy for its importance: for example, if "health" appears 30 times in a document, it may be more essential than if it appears only once. We also utilize document frequency (the number of documents that include a certain word) to determine how common the word is. This reduces the impact of stop-words like pronouns and domain-specific language that adds little information.

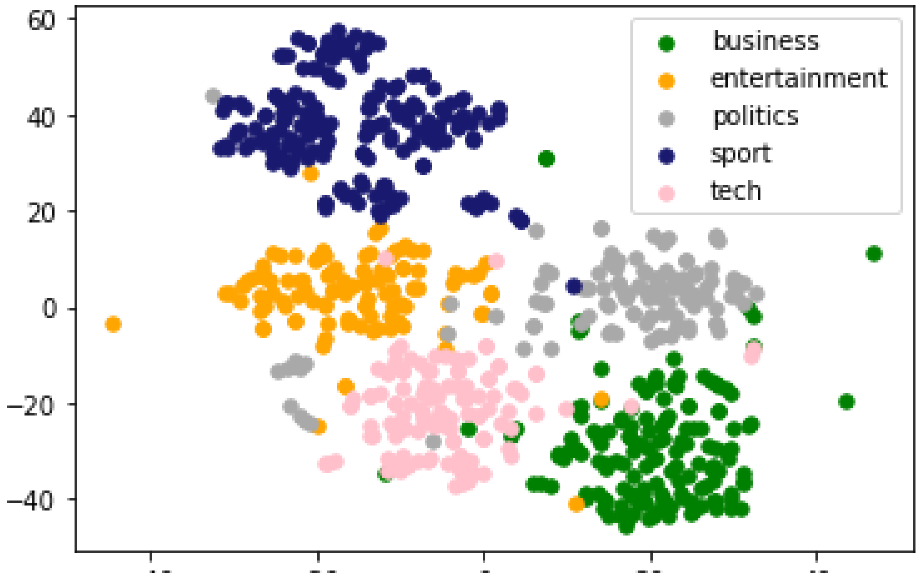
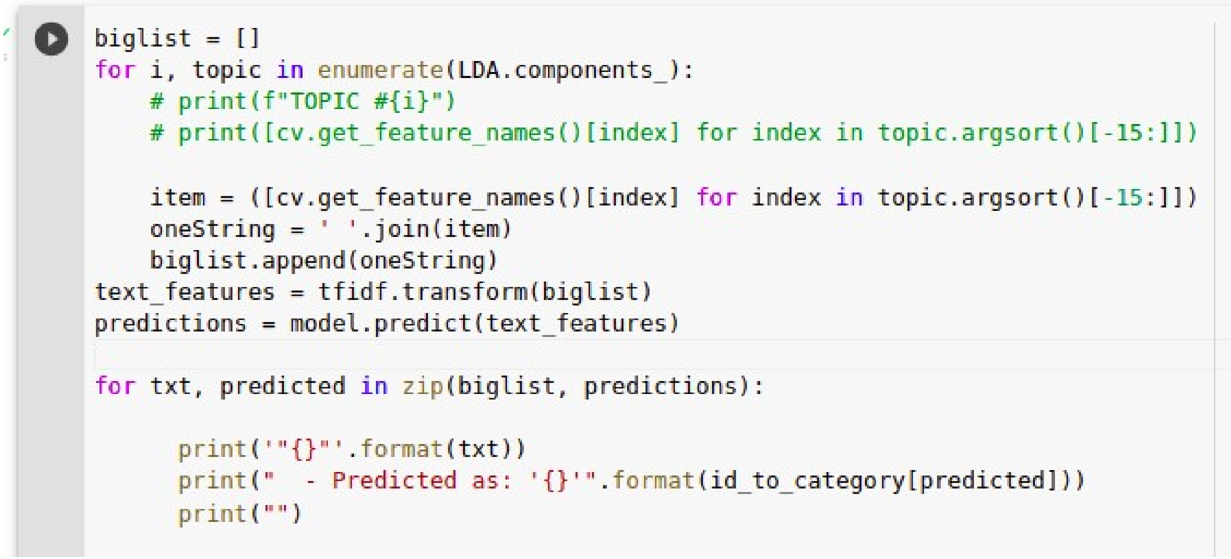


Figure3.4: tf-idf feature vector for each article, projected on 2 dimensions

We import our BBC news Dataset, remove duplicate rows, clean missing data and perform pre-processing techniques such as removing stop words, perform lemmatization, and lowercasing, removing numbers, removing special characters, removing email addresses, removing URLs, expanding verb contractions, normalizing backslashes to slashes and split tokens on special characters. Next, we extract N-Grams features from the pre-processed text. Here we create a vocabulary of 3-Grams size, using Term-Frequency- Inverse Document Frequency as the weighting function and specifying the minimum word length of 3 and maximum word length of 25. We then normalize the n-gram feature vectors and use filter-based feature selection and Chi-Squared as feature scoring method.

We then perform a randomized split of the resulting dataset into a fraction of rows. 70% in the first output dataset to be used for our training. The remaining 30% is to be used for model testing and evaluation. We then create a multiclass logistic regression classification model and then Tune the model hyperparameters. In this step, we run the model through a parameter sweep to discover the best parameter values. We use Accuracy as the Metric for measuring performance for classification. We used 70% of the dataset and use top Topic distributions directly as feature vectors from the LDA model in the supervised model training. Next, we Score the predictions of our trained classification model. Finally, we evaluate our scored classification model with standard metrics for evaluation.



```
biglist = []
for i, topic in enumerate(LDA.components_):
    # print(f"TOPIC #{i}")
    # print([cv.get_feature_names()[index] for index in topic.argsort()[-15:]])

    item = ([cv.get_feature_names()[index] for index in topic.argsort()[-15:]])
    oneString = ' '.join(item)
    biglist.append(oneString)
text_features = tfidf.transform(biglist)
predictions = model.predict(text_features)

for txt, predicted in zip(biglist, predictions):

      print('"{}"'.format(txt))
      print("  - Predicted as: '{}'".format(id_to_category[predicted]))
      print("")
```

Figure 3.5 Code Snippet: The output of LDA is used as input to Classifer.

In figure 3.5 for every topic and probalities from LDA, we utilize the Term-Frequeny Inverser Document Frequency to extract the text features. The text features then become input to a classifer model for predictions.

Table 3.2 Classifier automatically labels the output of LDA

LDA	TOPIC #1 ['buhari', 'jonathan', 'just', 'governor', 'vote', 'dickson', 'bello', 'bayelsa', 'party', 'kogi', 'people', 'election', 'state', 'pdp', 'apc']
LDA+LR	"buhari jonathan just governor vote dickson bello bayelsa party kogi people election state pdp apc" - Predicted as: 'politics'

4. RESULTS DISCUSSION

We conducted experiments with the following models: Multinomial Logistic Regression, Naive Bayes, and Random Forest. Each model is evaluated using the K-fold cross-validation technique. We Iteratively train the models on subsets of data and validate against the held-out data. An accuracy of 0.979782 was obtained from Multinomial Logistic Regression MultinomalNB: 0.970783; Random Forest 0.830570. The Random Forest model results in a huge variance which indicates that the model is overfitting to its training data. Multinomial Logistic Regression has a tiny edge with a median accuracy of roughly 97% over Multinomial Naive Bayes, although both work exceedingly well. However, we discovered an improvement in the overall classification accuracy of Multinomial Logistic Regression with LDA(LDA+LR) with accuracy results 98%.

Table 4.2: Classification Result

Category	Precision	Recall	F1-score
Tech	0.985	0.985	0.985
Business	0.982	0.958	0.97
Sports	0.995	1	0.997
Entertainment	0.992	0.984	0.988
Politics	0.938	0.968	0.952
accuracy			0.98
macro avg	0.978	0.979	0.978
weighted avg	0.98	0.98	0.98

Table 4.3: Result (sample) of Extracted Features and weights form the BBC dataset

Feature Weights					
Feature	business	entertainment	politics	sport	tech
Singer	-0.0920872	1.92495	-0.105108	-0.343511	0
Its	1.6857	-0.0301531	-0.229728	-1.31905	0.370104
Film	-0.25126	1.67053	-0.13449	-0.32231	0
Secretary	0	-0.114275	1.56434	-0.121951	0
Said	0.360848	-0.485533	0.764444	-1.49362	0.373949
Election	0	-0.0957465	1.48658	-0.183256	0
said:	-1.01805	0	1.39944	0	-0.334676
Match	0	-0.181719	-0.0340802	1.38201	0
Firm	1.37182	-0.85532	-0.335993	-0.486694	0.786102
Leader	0	-0.158923	1.36572	-0.135006	0
Computer	-0.0396927	0	-0.0521469	0	1.35031
Company	1.34729	0	-0.38233	-0.392605	0
Users	-0.252235	0	0	0	1.34243
Coach	0	-0.0668491	-0.0593724	1.29961	0
Band	0	1.2693	0	-0.0815981	0
Using	-0.144265	-0.308019	0	-0.265101	1.26486
Show	-0.0548785	1.26062	-0.049953	-0.216559	0
Minister	0	-0.336867	1.25034	-0.188247	-0.156481
Commons	-0.0110677	0	1.24534	-0.0426169	0
Software	-0.0600765	0	-0.0485117	0	1.24414
Tony	-0.0816883	0	1.23142	-0.082112	0
Bank	1.18893	-0.04012	-0.00766209	-0.0474986	0

Table 4.3 Displays the Extracted Features and Weights for the Model. The weights indicate the probability of a specific term pertaining to a specific topic

5. CONCLUSION

This research demonstrates that automated topic modelling can be successfully applied to a wide variety of applications. To extract topics from unstructured text, we created Latent Dirichlet Allocation (LDA) models. Additionally, we tested with four different types of classification tasks in order to make a more informed choice of a model for classifying the output from our LDA as well as previously viewed text into any of the five categories.

Our dataset is comprised of data scraped from the Narialand Forum website (source: https://www.nairaland.com/) and the BBC news dataset from 2004 to 2005 (source: http://mlg.ucd.ie/datasets/bbc.html). The Narialand dataset has 25785 documents, whereas the BBC news website dataset contains 26512 documents, all of which correspond to stories from 2004 to 2005. Business, Entertainment, Politics, Sports, and Technology make up the five categories of news. Another challenging task is analyzing the findings of LDA. We present a technique for training a supervised model on the LDA output. As a result, we are able to classify latent subjects automatically. Our classifier was built using Logistic Regression, a well-known machine learning approach.

In our research, we utilize multinomial logistic regression since our target variable, y (output of LDA), spans more than two classes and we want to know the likelihood that y belongs to each possible class. Each unigram was assigned a term frequency–inverse document frequency (TF–IDF). To extract 32,768 hashing characteristics, a bit size of 15 bits was specified. This experiment utilized the top 5000 closely related traits This model was subjected to an experiment, and the model performed with an accuracy of 98%. The resulting classifiers had a macro-average precision of 0.978, a weighted-average f1-score of 0.980, and a macro-average recall of 0.979. The results show that the two strategies were complementary in terms of identifying textual topics and labelling the output of LDA

References

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). “Latent Dirichlet Allocation”. Journal of Machine Learning Research 3 (993-1022)
2. Bruggermann, D., Hermey, Y., Orth, C., Schneider, D., Selzer, S., & Spanaki, G. (2016). Storyline detection and tracking using Dynamic Latent Dirichlet Allocation. *Proceedings of 2nd Workshop on Computing News Storylines* (pp. 9-19). Austin, TX: Association for Computational Linguistics.
3. Dai, B., Lim, E. P., & Prasetyo, P. K. (2012). Topic Discovery From Tweet Replies Workshop on Mining and Learning with Graphs (MLG-2012). *Proceedings of the 10th.* Edinburgh: Research Collection School Of Information Systems. Retrieved from https://ink.library.smu.edu.sg/sis_research
4. Guo, J.-C & Lu, Bao-Liang & Li, Z. & Zhang, Liqing. (2009). LogisticLDA: Regularizing latent dirichlet allocation by logistic regression. 1. 160-169.
5. Hyung, Z., & Lee, K. (2013). Recommending Music Based on Probabilistic Latent Semantic Analysis on Korean Radio Episodes. *Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (pp. 472-476). Beijing, China: IEEE.
6. Kherwa , P., & Bansal, P. (2017). Latent Semantic Analysis: An Approach to Understand Semantic of Text. *International Conference on Current Trends in Computer, Electrical, Electronics and Communication (ICCTCEEC).* Mysore, India: IEEE.
7. MA Basher, A. R., & Fung, B. (2014). “Analyzing topics and authors in chat logs for crime investigation”. Knowledge and information systems, 39(2), 351-381earch, 3(Jan), 993-1022.
8. Maalouf, M. (2011). Logistic regression in data analysis: An overview. *InternationalJournal of Data Analysis Techniques and Strategies*, 281-299.

9. Mao, X. L., Zhao, Y. J., Zhou, Q., Yuan, W. Q., Yang, L., & Huang, H. Y. (2016). “A novel fast framework for topic labeling based on similarity-preserved hashing”. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 3339-3348).
10. Mei and C.X. Zhai. 2005. “Discovering evolutionary theme patterns from text: an exploration of temporal text mining”. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pages 198–207. ACM.
11. Mei, C. Liu, H. Su, and C.X. Zhai. 2006. “A probabilistic approach to spatiotemporal theme pattern mining on weblogs”. In Proceedings of the 15th international conference on World Wide Web, pages 533–542. ACM.
12. Rahman, A., Basher, M. A., & Fung, B. C. (2013). Analyzing Topics and Authors in Chat Logs for Crime Investigation. Canada.
13. Rahman, Shadikur & Hossain, Syeda & Arman, Md & Rawshan, Lamisha & Toma, Tapushe & Rafiq, Fatama & Biplob, Khalid, 2020).
14. Ramage, D., C.D. Manning, and S. Dumais. 2011. “Partially labeled topic models for interpretable text mining”. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining pages 457–465. ACM.
15. Ramage, D., P. Heymann, C.D. Manning, and H. Garcia-Molina. 2009. “Clustering the tagged web”. In Proceedings of the Second ACM International Conference on Web Search and Data Mining, pages 54–63. ACM.
16. Kotu, V. and Deshpande, B., 2022. *Data Science Concepts and Practice*. 2nd ed. Elsevier Inc, pp.221-261.
17. Starkweather, J., & Moske, A. (2011). Multinomial Logistic Regression. Retrieved 1 October 2022, from https://it.unt.edu/sites/default/files/mlr_jds_aug2011.pdf
18. El-Habil, Abdalla. (2012). An Application on Multinomial Logistic Regression Model. Pak.j.stat.oper.res.. 8. 10.18187/pjsor.v8i2.234.