NIDHI DEEPAK

SRIKANTH BOLA

SRI VARDHAN GOUD PADALA

PRANATHI REDDY MULA

AKHILA DANDUBOYENA

Headline Classification

DSCI 5350.001 – BIG DATA ANALYTICS

**Contents**

[Chapter 1: Executive Summary 4](#_Toc164852122)

[Chapter 2: Project Motivation 5](#_Toc164852123)

[Chapter 3: Dataset Description 6](#_Toc164852124)

[Chapter 4: Data Preparation 7](#_Toc164852125)

[4.1: Feature Selection 8](#_Toc164852126)

[4.2: Handling Missing Values 9](#_Toc164852127)

[4.3: Displaying Non-Truncated Data 10](#_Toc164852128)

[Chapter 5: Data Analysis 11](#_Toc164852129)

[5.1: Category Distribution Analysis 11](#_Toc164852130)

[5.2: Frequent Titles in News Articles 12](#_Toc164852131)

[5.3: Transformation - Regex Transformation 14](#_Toc164852132)

[5.4: Tokenization and Stop Words Removal 15](#_Toc164852133)

[5.5: Refinement of Tokenized Words by Removing Stop Words 16](#_Toc164852134)

[5.6: Encoding Categories for Machine Learning 17](#_Toc164852135)

[5.7: Categorization Frequency in News Data 18](#_Toc164852136)

[5.8: Publisher Representation in News Dataset 19](#_Toc164852137)

[5.9: Article Count by Publication Year 19](#_Toc164852138)

[5.10: Word Frequency Analysis in News Articles Titles 20](#_Toc164852139)

[5.11: Distribution of Articles by Year and Category with Null Values Highlighted 21](#_Toc164852140)

[5.12: Identifying Repeated Titles 22](#_Toc164852141)

[5.13: Visualization of Common Terms in News Titles with Word Cloud 23](#_Toc164852142)

[5.14: Latent Dirichlet Allocation (LDA) 24](#_Toc164852143)

[5.15: Model Training 26](#_Toc164852144)

[Chapter 6: Findings 28](#_Toc164852145)

[Chapter 7: Business Implications 29](#_Toc164852146)

[Chapter 8: Conclusion 30](#_Toc164852147)

[Chapter 9: References 31](#_Toc164852148)

**List of Figures**

[Figure 1: Installing pyspark libraries. 7](#_Toc164852149)

[Figure 2: Values in the dataset. 7](#_Toc164852150)

[Figure 3: Essential Columns for News Classification 8](#_Toc164852151)

[Figure 4: Null Value Analysis in Title and Category 9](#_Toc164852152)

[Figure 5: Full Title and Category View 10](#_Toc164852153)

[Figure 6: Category Count Overview 11](#_Toc164852154)

[Figure 7: Most Frequent News Article Titles 12](#_Toc164852155)

[Figure 8 Bar Chart of Category Distribution 13](#_Toc164852156)

[Figure 9 Bar Chart of top 10 publishers 13](#_Toc164852157)

[Figure 10: Cleaned News Titles for Analysis 14](#_Toc164852158)

[Figure 11: Tokenized and Cleaned Title Data 15](#_Toc164852159)

[Figure 12: Stopword Removal from Tokenized Titles 16](#_Toc164852160)

[Figure 13: Category Encoding for Machine Learning 17](#_Toc164852161)

[Figure 14: News Category Distribution Count 18](#_Toc164852162)

[Figure 15: Top Publishers by Article Volume 19](#_Toc164852163)

[Figure 16: Article Counts by Year 19](#_Toc164852164)

[Figure 17: Word Frequency in News Titles 20](#_Toc164852165)

[Figure 18: Category Distribution with Null Values 21](#_Toc164852166)

[Figure 19: Frequency of Repeated News Titles 22](#_Toc164852167)

[Figure 20: Word Cloud of Trending Topics in News 23](#_Toc164852168)

[Figure 21: NLP Text Processing Stages 24](#_Toc164852169)

[Figure 22: LDA Topic Modeling 24](#_Toc164852170)

[Figure 23: LDA Model's Topic Distribution 25](#_Toc164852171)

[Figure 24: Naïve Bayes Classification Results 26](#_Toc164852172)

[Figure 25: Naïve Bayes Model Accuracy and Test Error 27](#_Toc164852173)

# Chapter 1: Executive Summary

This project by Team Oaks employs advanced NLP techniques to classify 422,937 news articles: Business, Science and Technology, Entertainment, and Health, respectively. The data was collected from various publishers between March and August 2014. We utilized PySpark to clean the dataset, thus helping better manipulate the data for the machine learning model and increasing its accuracy further. Encoding the data with categories and removing stop words when preparing the data for classification will enable businesses to tailor content and ads based on user engagement patterns.

The detailed dataset exploration reveals significant diversity across the article distribution into categories, most belonging to Entertainment and Business. In our data preprocessing, we strictly cleaned the data, removing stop words and numerical characters from titles - being quite harmful in model training - using PySpark encoding for categorical data and text tokenization. It made our classification models' input unambiguous.

This class of algorithms was fine-tuned to adapt to the data's imbalanced distribution, thus ensuring unbiased and accurate predictions. Our project illustrates the process of data cleaning, tokenization, and frequency analysis of the titles used in the model's training phase.

This enabled our project to achieve high classification accuracy, proving the efficiency of the chosen algorithms and pre-processing methods. The achievement lent credence to machine learning's automation potential for categorizing large text corpora, which is immensely needed to help content management strategies within digital publishing.

# Chapter 2: Project Motivation

The project aims to exploit NLP capabilities to understand and classify a large amount of textual data efficiently. With the tremendous and varied nature of news data, automatic classification adds organization to the content and can assist in business trends and the identification of consumer interest. The project will, therefore, aim to demonstrate how, from machine learning, the raw data is transformed into actionable insights that best inform the improvement of content delivery and advertising strategies.

The model was meant to classify news articles under relevant categories to which they belong, support publishers in their content management, and offer targeted content to their customers. Classification also supports the offering of personalized content recommendations as a way of driving increased user engagement and optimizing strategies for advertisements to align with the most relevant content.

Moreover, insights gained from classification results cover broad implications for understanding customer interests and trends, which are crucial for businesses to tailor their offerings to market demands. The high accuracy of our model, as demonstrated in the project's findings, confirms the reliability of using automated systems for such classifications.

With such in-depth analytics of the article distributions and frequency, we were able to fine-tune our approach to handling the diverse data sets effectively. Ideally, this project's success will go a long way in boosting news aggregation and set the pace and standard for many similar applications in other sectors that deal with large-scale text data. This work shows how the field of NLP is transforming in the digital era and how text data will be put into the process and further applied in the respective industries.

# Chapter 3: Dataset Description

This dataset comprises of 422937 news articles collected by a web aggregator from March 10th, 2014, to August 10th, 2014. It consists of headlines, URLs, and different categories of news stories.

The Dataset contains the following variables:

**ID**: Numerical identifier for each article.

**Title**: Headline of article

**URL:** web address of article where it was published.

**Publisher:** Publisher of article

**Category:** Different categories are:

b: Business

t: Science and Technology

e: Entertainment

m: Health

**Story:** This is an alphanumeric identifier of news story that the article discussing.

**Host Name:** Name of the host or webpage where article was published.

**Time Stamp:** Publication time of article provided in Unix time format.

# Chapter 4: Data Preparation

Installed pyspark libraries and imported features, functions, classifiers, evaluators as shown in figure 1.



Figure 1: Installing pyspark libraries.

A screenshot of a computer

Description automatically generated

Figure 2: Values in the dataset.

Figure 2 shows the first few entries of the ‘news\_data’ in PySpark. The data frame is displayed using the show () method, which provides the ID, Title, URL, Publisher, category, Story, Hostname, Timestamp.

## 4.1: Feature Selection

A screenshot of a computer screen

Description automatically generated

Figure 3: Essential Columns for News Classification

Figure 3 shows the new data frame focusing on two essential columns for classification.

**Title:** The headlines of the news articles, which are the text features to be used in the classification algorithms.

**Category:** The target variable for the classification, indicating the genre or topic each article belongs to.

## 4.2: Handling Missing Values

A screenshot of a computer program

Description automatically generated

Figure 4: Null Value Analysis in Title and Category

Figure 4 shows the output for number of null values in both the title and category variables i,e. 389 null values in title and 516 null values in category.

Compared to total number of instances which are 422937 After cleaning the total number of entries is shown to be 422,421. This indicates that some rows have been removed from the Data Frame as part of the data cleaning process.

## 4.3: Displaying Non-Truncated Data

A screenshot of a computer screen

Description automatically generated

Figure 5: Full Title and Category View

Previewing full titles in the dataset. The show () method is used to display the Data frame. Setting the truncate=False ensures that the content of the Data Frame is not truncated in the output, which is particularly useful for viewing the complete text of longer titles.

The data frame shows a list of news articles and their associated categories. Categories are represented as single letters, such as ‘b’, which likely stands for a specific type of news, such as business.

# Chapter 5: Data Analysis

## 5.1: Category Distribution Analysis

A screenshot of a computer

Description automatically generated

Figure 6: Category Count Overview

A distinct count operating reveals that there are 265 unique categories in the dataset. A subsequent ‘groupBy’ and ‘count’ operation followed by an ‘orderBy’ shows the frequency of each category, highlighting that most articles belong to categories ‘e’, ’b’, ‘t’, and ‘m’, which could stand for entertainment, business, technology, and health. This distribution is essential to understand the dataset’s skewness or balance and will significantly impact the performance of any classification model.

## 5.2: Frequent Titles in News Articles

A screenshot of a computer

Description automatically generated

Figure 7: Most Frequent News Article Titles

Data Frame is grouped by the ‘TITLE’ column, counted, and then ordered by the count in descending order. The titles that appear most frequently are listed along with their counts. A message related to missing articles tops the list with 145 occurrences. This insight could be important for data cleaning, as it may influence the decision to exclude titles from the classification.

As shown in the figure 8 category is distributed approximately evenly with economy “e” as the most occurring thing in the dataset and the category ‘m’ is the least occurred value in the dataset

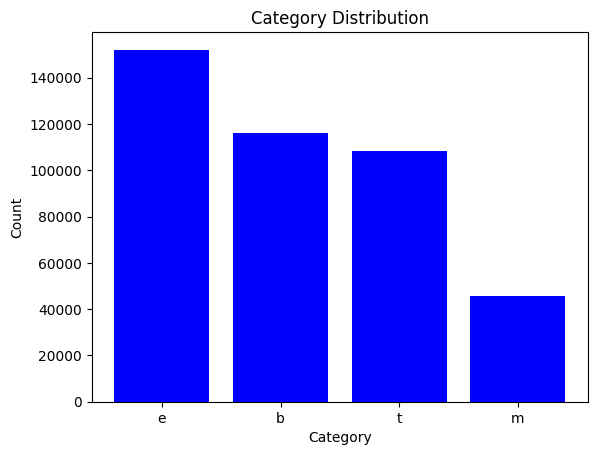


Figure Bar Chart of Category Distribution

Figure 9 explains that reuters are the most frequent publisher for the news and published around 3800 news articles in the current period of time

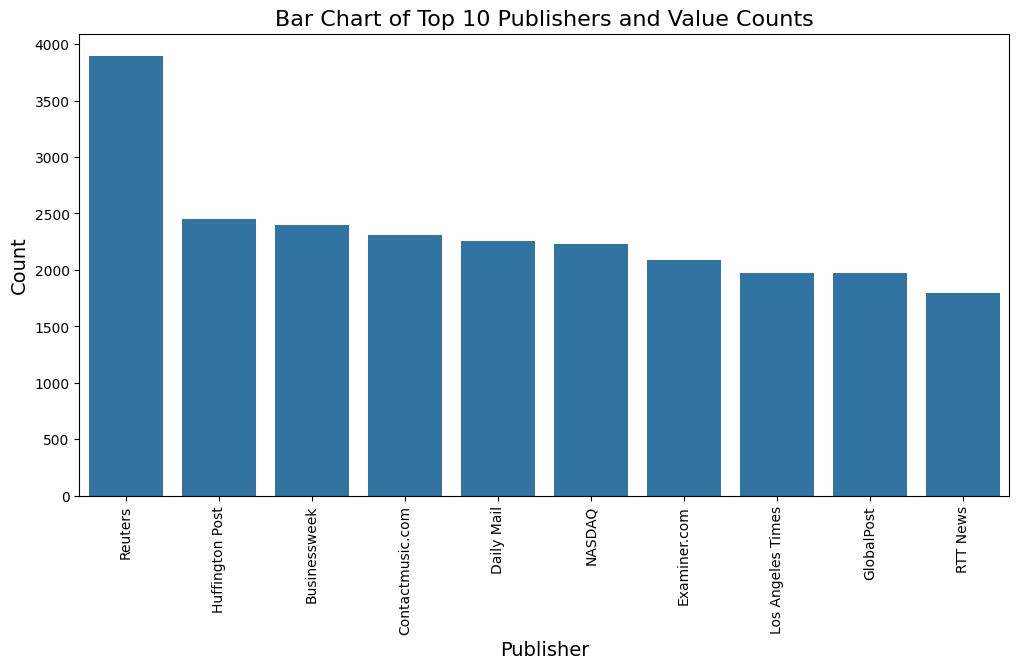


Figure Bar Chart of top 10 publishers

## 5.3: Transformation - Regex Transformation

A screenshot of a computer

Description automatically generated

Figure 10: Cleaned News Titles for Analysis

We cleaned the data from numbers in the titles which are not useful information and could potentially skew the analysis. The result ‘only\_str’ column in the data frame will have titles free from numeric characters, which is often a necessary preprocessing step for text data before it is used for tasks like machine learning modelling.

## 5.4: Tokenization and Stop Words Removal

A screenshot of a computer

Description automatically generated

Figure 11: Tokenized and Cleaned Title Data

The transformation removes numerical characters that might be a part of the news articles titles, such as dates or numbers, which could be irrelevant for modeling tasks.

This regex transformation is a text cleaning step that simplifies the titles and helps standardize the input for subsequent NLP processes. The cleaned titles are stored in the new column ‘only\_str’, which can then be used for further analysis.

## 5.5: Refinement of Tokenized Words by Removing Stop Words

A screenshot of a computer program

Description automatically generated

Figure 12: Stopword Removal from Tokenized Titles

The text data has undergone a tokenization process which splits the titles of articles into individual words. Following tokenization, a stop words removal processing is initiated. Stop words are commonly used words such as ‘the’, ‘is’ ‘at’, ‘which’ that usually do not carry important meaning and are thus removed from the analysis.

The PySpark ‘StopWordsRemover’ is utilized here. It is configured to take the column ‘words’ which contains the tokenized text and produce a new column ‘filtered’ that excludes stop words.

Words columns shows the original tokens from the titles while filtered shows the tokens after stop words have been removed.

## 5.6: Encoding Categories for Machine Learning

A screenshot of a computer

Description automatically generated

Figure 13: Category Encoding for Machine Learning

The PySpark Data frame has undergone a category encoding process. The ‘StringIndexer’ is applied to the ‘CATEGORY’ column of the DataFrame, which contains categorical data representing news article categories.

## 5.7: Categorization Frequency in News Data

A screenshot of a computer

Description automatically generated

Figure 14: News Category Distribution Count

The data is being grouped by the ‘CATEGORY’ column, and the number of occurrences for each category is being counted. The output reveals the distribution of articles across various categories. It indicates labeled ‘e’, ‘b’, ‘t’, and ‘m’ have the highest frequency with ‘e’ being the most common.

## 5.8: Publisher Representation in News Dataset

A screenshot of a computer

Description automatically generated

Figure 15: Top Publishers by Article Volume

It’s evident from the output that Reuters tops the list with 2898 articles, followed by Huffington Post and Businessweek among others. This analysis is key to understanding which news sources are most represented in the dataset.

## 5.9: Article Count by Publication Year

A computer screen shot of a code

Description automatically generated

Figure 16: Article Counts by Year

The result of this operation shows a count of articles published in the year 2014 and highlights that there are some articles with a null year, which indicates missing data in the ‘TIMESTAMP’ column or records that might correspond to a different year outside the range of the dataset.

## 5.10: Word Frequency Analysis in News Articles Titles

A screenshot of a computer code

Description automatically generated

Figure 17: Word Frequency in News Titles

The output Data Frame ‘words\_df’ shows two columns’ words and counts. The most frequently occurring words, which are typically common English stop words such as ‘to’, ‘in’, ‘the’, are listed at the top with their counts.

## 5.11: Distribution of Articles by Year and Category with Null Values Highlighted

A screenshot of a computer

Description automatically generated

Figure 18: Category Distribution with Null Values

The output displayed is a truncated list showing the top 20 rows, revealing that aside from the NULL categories, there is a diverse array of unique categories with a count of one, indicating a long distribution.

## 5.12: Identifying Repeated Titles

A screenshot of a computer

Description automatically generated

Figure 19: Frequency of Repeated News Titles

Displaying the frequent count of titles, performing Group By on the title column. The result table indicates that there are numerous instances where the title is either missing or means that the title doesn’t have a specific title.

## 5.13: Visualization of Common Terms in News Titles with Word Cloud

A screen shot of a computer code

Description automatically generated

Figure 20: Word Cloud of Trending Topics in News

In the word cloud, terms like ‘Kim’, ‘Kardashian’, ‘Samsung’, ‘Galaxy’, ‘China’, and ‘Twitter’ are prominently displayed, showing that they are among the most frequently mentioned or focused topics. Larger and bolder words like ‘Kardashian’ indicate higher occurrences in the text data.

## 5.14: Latent Dirichlet Allocation (LDA)

LDA is an unsupervised machine learning model used to discover abstract topics from a collection of documents. It is particularly popular in text mining. LDA models the text data by assuming that each document is a mixture of a certain number of topics. And each topic is a distribution of words.

The LDA model then learns by going through each document iteratively and assigning each word in the document to one of the K topics. This assignment is not random but based on two probabilities.

A close-up of a computer screen

Description automatically generated

Figure 21: NLP Text Processing Stages

A white sheet of paper with text

Description automatically generated

Figure 22: LDA Topic Modeling

Once the LDA model had finished learning, it gives two key outputs:

* A list of topics identified from the corpus. Each topic is characterized by a distribution over words.
* For each document, a distribution over the discovered topics.

A screenshot of a computer

Description automatically generated

Figure 23: LDA Model's Topic Distribution

Topic 0: Focused on Celebrity news, especially around Paula Patton and Robin Thicke, suggests articles around their personal life and music career development.

Topic 1: Centers around technology and entertainment with references to Apple, Google, and new game releases in 2014.

Topic 2: Mixes updates about Ebola with pop culture elements involving Jay-Z, Beyonce, and Facebook

Topic 3: Discussed cultural entertainment topics, including movies or celebrities with references to Lana Del Rey and Lupita.

Topic 4: Likely focuses on entertainment technology and television with mentions of Google, Android and Tv shows like ‘Mad Men.’

## 5.15: Model Training

The Naïve Bayes algorithm is a probabilistic machine learning model that’s used for classification tasks. The ‘naïve’ aspect comes from the assumption that the features used to predict the target variable are independent of each other.

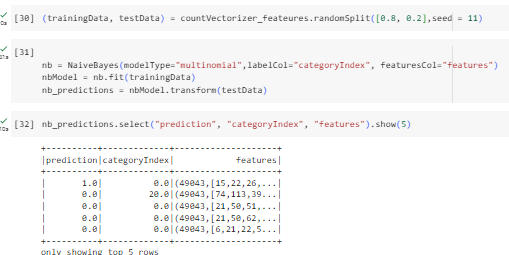


Figure 24: Naïve Bayes Classification Results

The dataset, which has been transformed into feature vectors by ‘CountVectorizer’, is randomly split into a training set (80% of the data) and the remaining 20% for testing the data. This is done to evaluate the model’s performance on unseen data. The model is instantiated with multinomial as the model types and then fitted on the training data, learning the probability distribution of the features. The output shows the top 5 values of the prediction column of the model. The categoryIndex converts the categorical strings to numerical uniquely. For ex, Business, entertainment, other categories available are indexed as 0,1,2 etc. The features are produced by Countvectorizer where it creates vector representing the document based on word counts.

**Accuracy:**

A screenshot of a computer code

Description automatically generated

Figure 25: Naïve Bayes Model Accuracy and Test Error

The output indicated that the Naïve Bayes model achieved an accuracy of 92.6%. This suggests that the model correctly predicted the category index for 92.6% of the cases in the test data. The test error of about 0.074 of the predictions were incorrect.

# Chapter 6: Findings

The project uncovered several key insights from the dataset of the news articles. The categorization analysis revealed that most of the articles fell under category ‘e’ which is Entertainment and ‘m’ which is Health had the least number of published articles. We analyzed that the Reutters publications was leading the highest numbers of published articles. Through LDA model insights, it discovered topics within the articles, with certain words like ‘economy’, ‘market’, ‘health’, and ‘technology’ being more prevalent. These topics are the collective focus of the news articles and can serve as measures of public interest. The Naïve Baye model’s performance, with an accuracy of 92%.

Topic model generated the 10 topics which classified different words into each topic and the top 5 topics are describing each category and will be helpful to find whether the given text heading is categorized among those 4 values.

# Chapter 7: Business Implications

* The analysis and categorization of news articles can significantly enhance content personalization strategies for media platforms such as Twitter and Instagram by understanding their interested topics and engagement.
* The insights from the text classification can inform ad targeting by aligning advertisements with the most relevant article categories.
* Insights from topic modeling and article categorization can guide teams in making data-driven decisions about which types of articles to prioritize to promote based on current trends and reader interests.
* Understanding the most frequently discussed topics and keywords can aid in search engine optimization, enabling news outlets to rank higher in search engine results and attract more traffic.

# Chapter 8: Conclusion

The project effectively demonstrates the utilization of PySpark and Spark NLP in classifying a massive set of news articles into meaningful categories. The initial EDA provided valuable insights into the distribution and characteristics of the data, which informed the subsequent modeling phase. By using robust Machine Learning and NLP for automated news categorization and topic modeling. The Naïve Bayes classifier proved to be a useful tool, exhibiting high accuracy of 92% in categorizing news articles into distinct groups. We also used an unsupervised algorithm, Latent Dirichlet Allocation (LDA) model that provided a large perspective on the dataset. These findings support the use of these models in digital news platforms to enhance content and deliver personalized user experiences. The methodologies applied can be useful for publishers to map out content strategies.

In topic analysis most of the topic words are picked as celebrity names, music names, and their personal life happenings which indicates that those words are most important to classify a heading as an entertainment topic.

# Chapter 9: References

**[1]** M. I. Rana, S. Khalid, and M. U. Akbar, "News classification based on their headlines: A review," 17th IEEE International Multi Topic Conference 2014, Karachi, Pakistan, 2014, pp. 211-216, Doi: 10.1109/INMIC.2014.7097339.

[2] Garbarino, E. (2023, Nov 13). “Using NLP to Categorize News Headlines”

<https://garba.org/posts/2022/news_analysis/>

[3] Katharina, M. (2022, Jan 3). “Labeling Text Data for News Article Classification and NLP” <https://www.omdena.com/blog/labeling-text-data-for-news-article-classification-and-nlp>

[4] Varun Dogra, Sahil Verma, Kavita, Pushpita Chatterjee, Jana Shafi, Jaeyoung Choi, Muhammad Fazal Ijaz, "A Complete Process of Text Classification System Using State-of-the-Art NLP Models", *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 1883698, 26 pages, 2022. https://doi.org/10.1155/2022/1883698