**News Article Classification Using Natural Language Processing and Analysis of Word Clouds and Machine Learning Algorithms**

**Group – A**

Abhishek Naithani - c0871411

Karishma Pramod Shirsath - c0871245

Kalla Tejaswi - c0852124

Venkata Sai Charan Yerapasetty - c0863224

**Instructor:** **Vahid Hadavi**

**Class: AML – 2304 (Natural Language Processing)**

**Date:** 25-06-2023

ABSTRACT

A vast amount of news is generated online every day. The importance of automated systems for identifying news articles has increased with the quick expansion of digital news and information. This report presents a comprehensive analysis of news article classification methods. One of the methods in Natural Language Processing (NLP), developed for the objective of classifying documents into pre-determined categories based on their contents is text classification or topic spotting. We have explored various techniques of text transformation retrieved from the web and machine learning modeling to categorize the genre of news articles published on the CBC news website.

INTRODUCTION

Everyday news has become an integral part of our lives. Over the past few decades, news distribution and consumption have altered significantly, from conventional media like printed newspapers to consuming news online via websites and smartphone apps in a variety of formats, including social media, news summarizers, and podcasts. Digitalization had influenced the reach of instantaneous news from one corner of the world to another. The advancement in technology has now made it easy for the user to customize their preferences and priorities.

Manually categorizing and classifying the enormous number of publicly accessible articles becomes problematic as the internet replaces print media as the dominant source of news consumption. Thus, automated methods for classifying news articles have become crucial for NLP applications such as content analysis, recommendation systems, and information retrieval. The process of classifying news articles signifies tagging them with predetermined labels or categories in accordance with their content. We can enable effective viewing, filtering, and searching of news items by precisely categorizing articles. This enables the users to personalize their reading experiences.

In this report, we analyze and examine various methods and NLP algorithms for classifying news articles. We will also understand machine learning algorithms like Naive Bayes, Support Vector Machines (SVM), and Random Forests, to build a model that could deliver a finely classified news article into various categories (Business, Politics, Entertainment, Health, Science, and Indigenous). We will go over how to clean and prepare text data, feature extraction techniques like Bag-of-Words, TF-IDF, and word embeddings, compare word clouds, as well as how to evaluate classification models.

METHODOLOGY:

2.1 – Web Scraping:

* BEAUTIFUL SOUP:

Beautiful Soup is a Python library used for parsing web content. By creating these objects, the content of the web pages for each HTML or XML document can be quickly accessed and modified. This might also be helpful for things like web scraping, extracting certain information, or analyzing the data that has been retrieved. For data scraping, we have taken CBC news articles and generated a dataset.

2.2 – Data Preprocessing:

Data preprocessing is often done to clean up the data and convert it into a format that is appropriate for further analysis or modeling. Some of the data-wrangling steps that we followed in this project are discussed below.

* Data Cleaning:

It involves dealing with missing, erroneous, and duplicated values in the data. After generating the dataset, we checked it for missing values and found “none”. There are 2 duplicate values. After performing certain functions and operations we got rid of duplicates and a cleaned dataset is created.

* Data Formatting
* Text Preprocessing with NLP

Text cleaning or processing is the stage where noise removal, standardized representation, and machine-compatible text are formatted. This stage comprises cleaning up the text data of any extra letters, symbols, or noise. As part of the process, Special Characters, HTML elements, punctuations, and any other non-alphanumeric characters that are irrelevant to the analysis are commonly removed. This improves semantic coherence, creates word vectors of higher quality, and more precise outcomes for various NLP tasks.

The Natural Language Toolkit (NLTK) is a Python library to manipulate and format text for easy and effective analysis. It comprises several built-in functions that perform specific tasks such as tokenization (breaking down a sentence into simpler words called tokens), the removal of stop words (those repetitive words that do not carry any significant meaning to the context), punctuations (using regular expression library), Stemming, and Lemmatization (to convert a word in the text to its root form), etc. It is also possible to filter out or eliminate any numerical numbers, dates, or other non-textual components from scraped data if they are not vital for the analysis at this stage.

2.3 - **Word Vectorization or Word Embedding** is a process of converting text or words taken from WordNet into a numeric format for fitting a machine learning model. There are several ways to vectorize words in NLP:

* Bag of Words
* Term Frequency – Inverse Document Frequency (Tf-idf)
* Word2Vec
* GloVe
* Fast Text

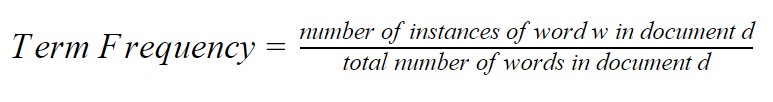
For this project, we have considered BoW and Tfidf to generate and compare the word clouds of both methods.

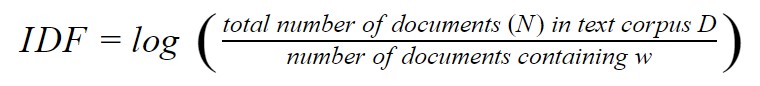
* 1. - Bag of Words:

Bag Of Words or Count Vectorizer is technically a feature extraction method to model text in NLP. It develops a vocabulary of distinct words from the corpus and gives each word a numerical value based on how frequently it appears in a document. BoW vectorization generates a sparse matrix with each row denoting a text sample or document and each column denoting a single word. The dimensionality of this matrix depends on the size of the vocabulary. This method ignores grammar and semantic information and depicts text as isolated words and transforms them into respective vectors for machine readability.

* 1. - Tfidf Vectorizer:

The term frequency is an indication of the number of times a particular word appears in a document. The frequency of a term reveals its importance. Each text from the data is expressed by term frequency as a matrix, where the rows correspond to the number of documents and the columns to the number of unique terms used in each document. Inverse Document Frequency quantifies a word by calculating the total number of documents in the corpus to the documents comprising the word ratio. This vectorization is effective in terms of ignoring stop words during the analysis.





* 1. - Comparing word clouds:

The formatted text is visually represented as a word cloud. The size of terms in a word cloud is influenced by how often they appear and their significance in the content. They are helpful for giving short glimpses into the data that can be looked at further.

A close up of words

Description automatically generated with low confidence

Because it offers both the importance of the words and their frequency in the corpus, TF-IDF is superior to Count Vectorizers in terms of word analysis. The model-building process can be made simpler by removing terms that are less crucial for analysis and so limiting the input parameters.

* 1. – Exploratory Data Analysis and Data Visualisation

EDA and visualization give a deeper understanding of our data and improved confidence in information extraction, creating the required machine-learning models, and an accurate article classification. We have used some plots to understand patterns and relationships of words from the scrapped data.

We can derive significant facts and deeper meaning from the text data by analyzing the frequency and distribution of unigrams, bigrams, and trigrams, allowing for more efficient evaluation and comprehension. The ‘count vectorizer’ from the sklearn library provides various functionalities for feature extraction and for creating bigrams and trigrams.

Bigrams can show word combinations, idiomatic idioms, and phrases that are vital in the corpus's context. Bigram analysis can help us learn more about the syntactic and semantic connections between words, allowing us to gather additional contextual data. In the below figure, we have plotted the frequency vs top-20 Bigrams (conjugated terms). We observed that ‘of-the’ has been repeated mostly and then ‘first-nations’.

A picture containing line, rectangle, colorfulness, design

Description automatically generated

Trigrams are the triplet pair of consecutive terms occurring in the text. It gives an assessment of sentence structures, syntactic structures, and how the language has been used during NLP analysis. We have plotted the following top-20 trigrams and found out that the pairs have similar repetitions overall in the corpus.

A picture containing colorfulness, pattern, rectangle, line

Description automatically generated

The below visualizations are a bar plot showing various news from the scrapped articles and the respective categories they fall into, and a pie chart of percentages of each category.

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing diagram, circle, line, screenshot

Description automatically generated

Insights from the above plots:

* Most of the news from selected articles is of the Indigenous category, with 32 records (19.5%) i.e., the news is related to the Indigenous or native community. Researchers, decision-makers, and people interested in recognizing and dealing with the special difficulties, experiences, and viewpoints of indigenous communities may find this classification useful.
* News from the ‘Politics’, ‘Health’, and ‘Entertainment’ fields share the same percentage (17.1%) with 28 records each in the articles.
* ‘Business’ news has 16.5% with 27 records classified that may be related to stocks, prices of goods, marketing, etc, related to Canadian Business.
* News classified into ‘Science’ is 12.8% with 21 records that may relate to technology advancements, Scientific research, Environmental science, etc, enlightening the public about the newest discoveries and research in a variety of scientific fields.
* All the news is classified from the Canadian perspective into different categories that may interest or concern the government and its people.
  1. – Train and Test Data Split:

Splitting data into train and test sets is a common practice to maintain neutrality in evaluation, identify overfitting, fine-tune hyperparameters, prevent data leaking, and make judgments based on the model's performance on new data with confidence. It aids in the development of trustworthy and dependable machine-learning models. After defining the input parameter (vectorized text data) and target feature (classification\_id), We have used a 70 – 30 split of the data to train the models and validate their performance respectively.

* 1. – Modelling

1. Random Forest Classifier

It is a group classifier called a random forest that makes predictions using a combination of multiple decision trees. It effectively fits a variety of classifiers using decision trees on different dataset subsamples. The news articles are individually categorized by each decision tree using random subsets of attributes, and the final prediction is chosen by the majority. It can successfully classify news articles into several categories by capturing complicated correlations between keywords and categories. We have optimized the ‘n-estimators’ parameter that determines the number of trees to be built for a better classification.

1. XGBoost (eXtreme Gradient Boosting)

It is a widely used supervised machine-learning algorithm for classification problems. Gradient boosting methods are used to recurrently train decision trees that fix the mistakes produced by earlier trees. Additionally, it offers feature importance scores, which are helpful in figuring out the main variables affecting the categorization. This process can discover intricate patterns and connections in the data. The processing competence makes it ideal for handling high-dimensional text data effectively. XGBoost provides several novel benefits for model tuning, computing environments, and algorithm enhancement.

1. Logistic Regression

Logistic regression is a simple, straightforward, and versatile algorithm for binary and multi-class classification problems. To produce predictions based on the input features, the logistic regression model acquires a set of weights for each feature, by minimizing a cost function like logistic loss or cross-entropy loss, it learns the ideal coefficients for each parameter during model training. Usually, optimization algorithms like gradient descent or its variations are used for this.

1. Naive Bayes

It operates on the concepts of Bayesian conditional probability and presumes that features are conditionally independent, given the class variable and its previous likelihood. It determines how likely or frequently each feature will appear within each class. The model requires less training data and is scalable for both continuous and discrete variables. It combines past and future probabilities and assigns a class to a new data input. Multiclass categorization assigns the article to the class that has the highest likelihood of all the classes (for example, politics, sports, and entertainment).

1. Support Vector Classifier (SVC)

SVC is a supervised machine-learning algorithm that primarily handles classification tasks. Finding an ideal hyperplane to divide data points into different classes in a high-dimensional feature space is the basic goal of SVC. The margin, which measures the separation between the hyperplane and the nearest data points for each class, is maximized in the selection of the hyperplane. Support vectors are the closest-lying data points that influence the position of the hyperplane. After training, SVC categorizes new observations by identifying which side of the hyperplane they lie on. The decision boundary can successfully divide data points into their appropriate groupings. Similarly, the news article is assigned to a class with respect to its position along the decision boundary.

* Confusion Matrix
* Classification Report

A classification report is a performance evaluation metric that provides precision, recall, F1-score, and support metrics for each class in the dataset, taking into account the predicted class labels and actual class labels.

**Precision**: It measures the proportion of accurate positive predictions to all of the model's positive predictions. It evaluates how well the optimistic expectations turned out. Fewer false positives are indicative of higher precision.

**Recall**: The ratio of genuine positive predictions to the total number of real positive cases in the dataset is known as recall, also known as sensitivity or true positive rate. Lesser false negatives are indicated by a higher recall.

**F1**: It offers a balanced measurement that considers a harmonic mean of both recall and precision. When the dataset is unbalanced or when both precision and recall are crucial, the F1-score is an effective statistic.

A screenshot of a graph

Description automatically generated with low confidence

* 1. Conclusion

Through this experiment, we discussed the significance and difficulties of news article classification in NLP. We provided effective information extraction, focused analysis, and customized recommendations for content by effectively categorizing news items. Researchers, journalists, and readers may benefit from this classification system's ability to help them quickly find the news stories that are most connected to their interests instead of navigating through massive amounts of information. The best accuracy obtained through modelling is 46%.

* 1. References