Project Proposal

Date: October 06,2024
Student Name/s: Viswateja Adothi – T00736529
Khadiza Tannee – T00729053
Solomon Maccarthy – T00734513
Professor : Dr. Quan Nguyen

1. Project Title: News Category (using SVM to detect Writing styles)

3.1. Introduction

The rapid evolution of digital media has transformed how we consume news, with readers now accessing content across multiple platforms and expecting varied storytelling approaches. As a result, understanding the differences in writing styles across different news categories has become crucial for publishers, content creators, and readers. The writing style of a news article greatly affects how readers engage with and interpret the content. For example, news reports focus on providing clear, unbiased information, while opinion pieces often express personal viewpoints and arguments.

This data science project aims to explore and classify the writing styles used in various news categories by using the News Category Dataset. The primary goal is to automatically identify the writing style of each article, focusing on categories such as informative, opinion, descriptive, and narrative. Analyzing these styles will provide insights into how different news categories shape storytelling and how these styles connect with various audiences.

By using Natural Language Processing (NLP) techniques like Support Vector Machine (SVM) and, if necessary, Long Short-Term Memory (LSTM) networks, this project aims to classify news articles based on their writing styles. The outcome of this project will provide valuable insights into the variation of writing styles across news categories and help us better understand the connection between writing style and storytelling.

3.2. Data Science Techniques and Description

In this project, we will start by using a **Support Vector Machine (SVM)** model to classify the writing styles in the News Category Dataset. If the SVM model does not perform as expected, we will explore a more advanced model, such as **Long Short-Term Memory (LSTM)**. These models

are suitable for this project because they have been effective in handling tasks like text classification, sentiment analysis, and recognizing writing styles.

Support Vector Machine (SVM) in This Project:

The use of SVM in writing style detection benefits from its capability to handle both linear and non-linear relationships, allowing for the extraction of nuanced stylistic characteristics from textual data, such as sentence length, word frequency, and punctuation usage. Furthermore, studies have shown that SVM can outperform traditional statistical methods, particularly when the feature set is carefully selected to represent the unique attributes of each author's writing style [1].

Long Short-Term Memory (LSTM) in This Project:

LSTM is a type of neural network that works well with text because it remembers the order of words and their context. If SVM doesn't perform well, we will use LSTM to better capture the flow and structure of longer texts, making it easier to detect more complex writing styles.

In contrast, LSTM networks, a variant of recurrent neural networks (RNNs), offer a powerful alternative for detecting writing styles, particularly due to their ability to capture temporal dependencies in sequential data. This characteristic is particularly beneficial when analyzing text, as it allows for the modeling of long-range dependencies that are often present in writing [2].

- SVM is simpler and faster, while LSTM is better for understanding more complicated patterns in the text.

1. Text Classification:

- **SVM** is highly effective for **text classification tasks**, including distinguishing between different **writing styles**. In this context, writing styles might include categories like **informative**, **opinionated**, **descriptive**, etc.

2. Capturing Subtle Differences in Writing Style:

- Writing styles can vary in small ways, like word choice, tone, sentence structure, and phrasing. By using **SVM with a non-linear kernel (like Radial Basis Function or RBF)**, we can detect these subtle differences and effectively distinguish between writing styles.

3. Feature Extraction:

To classify writing styles, we need to turn the text into features that the SVM model can understand. Here are some common methods we might use:

- **Bag of Words (BoW)**: This method counts how many times each word appears in the text. It helps us understand the most common words used.
- **TF-IDF (Term Frequency-Inverse Document Frequency)**: This method gives more importance to words that are unique to certain articles. It helps highlight words that are significant in the context of the entire dataset.

- **Word Embeddings**: This method converts words into numerical vectors that capture their meanings and relationships (for example, using Word2Vec or GloVe).

For this project, we may prioritize TF-IDF for SVM because it effectively highlights important words and reduces the influence of common terms. This can help SVM better identify differences in writing styles. However, we might also consider combining methods (like using both TF-IDF and BoW) to enhance our model's performance by capturing different aspects of the text.

Using SVM with these feature extraction techniques will allow us to identify patterns in word usage, phrase structures, and sentence lengths that define different writing styles.

4. Application to Writing Style Features:

- Beyond basic word features, you can extract more sophisticated features for SVM to detect writing styles, such as:
    - **Readability scores** (e.g., Flesch-Kincaid, Gunning Fog Index).
    - **Sentence structure**: Sentence length, use of passive voice, or punctuation.
    - **Tone and sentiment**: Using **sentiment analysis** to detect whether a piece is neutral, positive, or negative can help in differentiating writing styles.
    - **POS tagging**: Parts of speech distributions (e.g., nouns, verbs, adjectives) may vary between styles like **narrative** and **descriptive**.
- SVM can handle these features and combine them to make more nuanced classifications.

Challenges:

- **Data Imbalance:** In datasets where certain categories or writing styles are underrepresented, **SVM** may struggle to effectively classify these minority classes.
- **Overfitting on Noisy Data: SVMs** can be sensitive to **noisy data**. In a dataset like the **News Category Dataset**, where headlines and descriptions may include irrelevant or ambiguous information, the SVM may overfit to noisy patterns rather than general trends.

Dataset Overview:

The News Category Dataset consists of approximately 200,000 news articles from The Huffington Post, categorized across multiple sections such as Politics, Business, Entertainment, and Sports [3]. Key features of the dataset include:

Field/Feature	Description
ID	A unique identifier (typically a MongoDB objectID)
link	The URL of the news article on the website.
headline	The title or headline of the news article, providing a brief and often compelling summary of the article's content.
category	The section or category to which the article belongs
short_description	A brief, one or two sentence description of the article's content
authors	The name(s) of the article's author(s)
date	The date when the article was published, allowing for time-based analysis

This project will utilize key concepts from our data science courses, such as Data Collection and Storage, Data Cleaning and Preprocessing, Exploratory Data Analysis (EDA), and Machine Learning. These techniques will support our objective of deriving meaningful insights and improving model predictions.

Success Criteria for This Project

1. **Model Effectiveness**: The model should accurately classify writing styles in news articles.
2. **Practical Use**: The model should help publishers and content creators tailor their content based on writing styles.
3. **Adaptability**: The model should work well with both current and future articles.
4. **Clear Documentation**: There should be thorough documentation of the model's performance and processes for transparency.

Evaluation Metrics:

Accuracy:

- **Definition**: The percentage of articles that the model correctly classifies into their writing styles.
- **Why**: It shows the overall performance of the model in getting the classifications right.

Precision:

- **Definition**: The measure of how many of the articles classified as a specific style are actually that style.
- **Why**: It helps us understand how many correct classifications are made, minimizing errors.

Recall:

- **Definition**: The measure of how many actual articles of a specific style were correctly identified by the model.
- **Why**: It ensures the model captures most of the articles belonging to that style.

F1-Score:

- **Definition**: A combined measure of precision and recall that balances the two, especially useful for uneven categories.
- **Why**: It provides a balanced view of the model's performance, especially with uneven categories.

Confusion Matrix:

- **Definition**: A table that shows how well the model performed, displaying the correct and incorrect classifications for each writing style.
- **Why**: It helps visualize how well the model distinguishes between different styles and identifies areas of error.

Cross-validation Results:

- **Definition**: A technique used to check how well the model will work on new data by testing it on different subsets of the dataset.
- **Why**: It checks if the model can generalize well to new data and helps prevent overfitting.

3.3. Timeline

Task	Start Date	End Date
Data Collection and Preprocessing	Sep 3	Sep 10
Exploratory Data Analysis (EDA)	Sep 11	Sep 24
Model Development (Baseline - SVM)	Sep 25	Oct 15
Model Fine-tuning (LSTM, if needed)	Oct 16	Oct 29
Evaluation and Metrics Analysis	Oct 30	Nov 12
Dashboard and Visualization Setup	Nov 13	Nov 19
Final Review and Documentation	Nov 20	Nov 27

3.4. References

[1] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

[2] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[3] Huffington Post. (n.d.). *News Category Dataset*. Kaggle https://www.kaggle.com/datasets/rmisra/news-category-dataset