Project Proposal

Date: 26th September, 2024

Student Name/s: Viswateja Adothi

Khadiza Tannee

Solomon Maccarthy

Professor : 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 nuances of writing styles across different news categories has become crucial for publishers, content creators, and readers. The writing style of a news article significantly impacts reader engagement and interpretation. For instance, factual articles are more objective, while opinion pieces tend to include subjective arguments.

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

By employing 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 enhance our understanding of the complex relationship between writing style and storytelling.

3.2. Data Science Techniques and Description

We will commence the project by employing an SVM model to achieve the targeted classification objectives. Should the model fail to meet the predefined performance criteria, we will proceed with the implementation of an LSTM model. Both models are highly appropriate for analyzing the

News Category Dataset, owing to their robust language understanding capabilities and their proven effectiveness in tasks such as text classification, sentiment analysis, and writing style detection.

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 often involve subtle differences in **word choice**, **tone**, **sentence structure**, and **phrasing**. SVM, especially with a **non-linear kernel** like **RBF (Radial Basis Function)**, can capture these subtle differences in how texts are written.

3. Feature Extraction:

- To classify writing styles, text needs to be transformed into features that SVM can process. Common methods include:
  - ➢ **Bag of Words (BoW)**: Converts text into word frequency counts.
  - ➢ **TF-IDF (Term Frequency-Inverse Document Frequency)**: Weighs words based on how important they are in the dataset.
  - ➢ **Word Embeddings**: Converts words into dense vectors representing semantic meaning (e.g., Word2Vec, GloVe).
- SVM, combined with these feature extraction techniques, can learn to separate different writing styles by focusing on patterns of word usage, phrase structures, and sentence lengths.

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. 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.

Evaluation Metrics:

- Accuracy: The percentage of correctly classified articles.

- Precision, Recall, F1-Score: To better evaluate the performance on imbalanced categories.

- Confusion Matrix: To visualize the model's performance across various writing styles.

- Cross-validation: To ensure the model generalizes well to unseen data and does not overfit.

3.3. Timeline

Task	**Start Date**	**End Date**
Data Collection and Preprocessing	Week 1	Week 2
Exploratory Data Analysis (EDA)	Week 2	Week 3
Model Development (Baseline - SVM)	Week 4	Week 6
Model Fine-tuning (LSTM, if needed)	Week 7	Week 8
Evaluation and Metrics Analysis	Week 9	Week 10
Dashboard and Visualization Setup	Week 10	Week 11
Final Review and Documentation	Week 11	Week 12

3.4. References

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

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