Model Development Phase Template

Date	1 November 2024
Team ID	SWTID1726834817
Project Title	Fake News Analysis in social media using NLP
Maximum Marks	5 Marks

Feature Selection Report Template

In the forthcoming update, each feature will be accompanied by a brief description. Users will indicate whether it's selected or not, providing reasoning for their decision. This process will streamline decision-making and enhance transparency in feature selection.

Feature	Description	Selected (Yes/No)	Reasoning
Text	The actual content of the news article	Yes	Essential for NLP analysis, as the text will be used for vectorization and classification.
label	Indicates if the news is real or fake	Yes	This is the target variable needed for training and evaluation of the model.
text_length	Length of the text	Yes	Can help in identifying patterns, as fake and real news articles may have different lengths.

word_count	Total number of words in the text	Yes	Helps to capture the verbosity of the article, which may vary between real and fake news.
avg_word_length	Average length of words in the text	Yes	Provides insights into vocabulary usage, potentially helping to differentiate between classes.
stop_words (from CountVectorizer & TfidfVectorizer)	Common words that carry less meaning (e.g., "the", "is")	Yes	Excluding stop words can improve model performance by focusing on more meaningful terms.
Tfidf score	Term frequency-inverse document frequency score	Yes	Reflects the relevance of terms within the text, which aids in capturing important features.
alpha	Smoothing parameter in Naive Bayes model	Yes	Used in tuning the Naive Bayes model to optimize accuracy in prediction.
max_depth	Maximum depth of Decision Tree model	Yes	Hyperparameter that controls tree depth, preventing overfitting or underfitting.
min_samples_split	Minimum samples required to split a node in Decision Tree	Yes	Controls tree structure, helping to improve model generalization.

These features will help capture both the linguistic structure of the articles and statistical patterns that can differentiate fake and real news.