CSC 820 Natural Language Technologies

HW #10

Khalid Mehtab Khan

SFSU ID: 923673423

Everyone has their own unique style of writing. Even when using the same English language, people use them differently. By analyzing the properties of text written by someone like how often they use certain words, the number of adjectives, the emotions conveyed by the words, and the average length of sentences, we can understand their writing style. By using metrics like these, given a text, we should be able to estimate the likelihood that it was written by a specific person if we know their writing style. In this assignment, we used NLTK tools to achieve this. We initially trained a model using labeled data from three authors, analyzing the writers' using features like TF-IDF of words, nouns, average length, etc. Later, we used this model to try and predict the author of an unlabeled text.

As we examined the cross-validation scores across different settings, we noticed an increase in accuracy as we increased the number of folds from 2 to 20. As the training data covers all the range of data the model captures better trends.

- **K = 2:** Average accuracy was about 68.6%. This setting gave us the least reliable estimate of model performance due to the fewer number of folds.
- **K = 10:** Average accuracy improved to about 70.0%. This setting provided a more stable estimate of model performance as it averaged the scores over more folds.
- **K = 20:** Average accuracy slightly increased to about 70.2%. With even more folds, this setting is likely the most reliable for estimating how well the model performs across different subsets of the data.

Feature Importance Analysis

Class (EAP):

Words like "**say**," "**Mr**." and "**said**" are heavily weighted because it looks like the authors writing often uses dialogues in his stories, creating narratives that frequently have direct speech and involve words like these. These words reflect his narrative style that often involves interactions between characters, in the style of conversations.

Names are less frequent words in his writings, such as "**Idris**," "**Perdita**," and "**Raymond**," receive negative weights. These are not commonly found as they are more typical of indirect speech.

Class (HPL):

Words like "**old**," "**street**," and "**west**" are indicative of Lovecraft's focus on setting and atmosphere, important elements in his fictious stories. These words help understand the descriptions of environments that are characteristic of his style. This way of style can help distinguish the writer easily among the given set of authors.

On the other hand, emotional words like "**heart**" and "**love**" are less emphasized. Lovecraft's stories typically prioritize cosmic horror and fear over personal relationships or emotional depth, reflecting why these words are negatively weighted.

Class (MWS):

Names and emotions such as "**Raymond**," "**love**," and "**Adrian**," along with "**heart**" are heavily weighted because Shelley's novels often revolve around deep emotional themes and relationships amongst people as she used persons names directly. These features are crucial as they underline the personal and emotional context of her narratives, emphasizing character development and emotional expression.

Whereas common nouns and verbs like "**thing**" or "**things**" and "**say**" are less significant in her works, possibly because her style is more descriptive and focuses on emotional and thematic depth rather than mundane details or casual dialogue.

Features that are prominent and distinctive in an author's writing are given positive weights as they are strong identifiers for that author. Whereas features that are rare or have very less or 0 frequency to an author's typical content receive negative weights, indicating their lack of usefulness in predicting that author's texts. This approach allows the model to focus on the most telling aspects of each author's writing style.

Error Analysis and Trends

Impact of Average Word Length: We observed that variations in the average word length contributed to frequent misclassifications, particularly when texts by EAP were incorrectly identified as MWS's work. This typically occurred when EAP's usually concise word usage unexpectedly expanded, mimicking the more verbose style characteristic of MWS, thus leading to these erroneous classifications.

Misclassification Between EAP and MWS: A recurrent issue was the misclassification of EAP's texts as those of MWS. This trend underscores the difficulty in differentiating between the styles of EAP and MWS when certain textual characteristics, such as the complexity of sentence structures or the usage of specific thematic vocabulary, are more reminiscent of MWS's writing style.

Influence of Nouns and Verbs: Misclassifications were notably pronounced in instances involving MWS's texts, where an overrepresentation of nouns and verbs — given the negative weighting of these features in MWS's profile — skewed the predictive accuracy. This anomaly suggests a sensitivity in the model to fluctuations in the frequency of these parts of speech, which may not align with the typical stylistic markers of MWS's writing.