Computational Data Analytics
Project Proposal

Multi-Label Genre Classification for Books Using BERT and Explainable AI

AS (Group 197)

I. Overview (Problem Statement)

A. **The Plight of Self-Publishing Authors:** Accurate genre classification of books can be essential for discoverability and sales, especially for the first-time authors of over 1 million self-published titles released annually in the US alone. Misclassification can cause titles to be poorly marketed or listed in irrelevant categories, limiting their visibility on platforms like Amazon and Goodreads, while proper metadata like genre tags can increase discoverability by as much as 55%. In fact, authors could greatly benefit from accurate multi-genre classification, such as a label of both "romance" and "mystery", in order to benefit from improved searchability, recommendation accuracy, and targeted advertising.

B. **Facilitating Multi-Label Classification:** I will build a multi-label classification model for predicting multiple literary genres given a book's synopsis. This task is nontrivial, as overlapping labels and the nuance of distinctions between genres can present complexity; for instance, a title could belong to multiple similar genres such as "mystery" and "thriller".

   1. **Addressing Label Overlap and Nuance:** To address this problem, I will leverage BERT (Bidirectional Encoder Representations from Transformers) to build a context-aware model, rather than one that takes each word at face value. For instance, utilizing BERT, my model may be able to understand, based on context, whether the word "murder" refers to a crime in a mystery novel, or an odd verbiage choice for a grouping of crows in a classic title. The "Transformers" aspect of BERT facilitate positional encodings that will maintain sequential information.
   2. **Binary Vector Formatting:** Additionally, to address the fact that there are multiple correct outputs for each book, I will transform labels into a binary vector format. For training purposes, I will need to ensure that the model balances precision and recall across the genres considered.
   3. **Analyzing Genre Signatures:** Performing analysis on this kind of dataset has interesting potential to indicate which keywords most significantly drive a genre classification.
   4. **Explainability:** Finally, given the complexity and high-dimensionality of the BERT model, I'll use XAI (Explainable AI) techniques like SHAP (SHapley Additive exPlanations) to understand global feature importance,

the features that the model thinks are important; local interpretability, an explanation for why specific classifications are made; and characteristics of flagged features, with regard to whether they positively or negatively influence the prediction, as well as how features interact with each other.

C. **Application Outside of the Book Domain:** I'm interested in working on this project to cultivate skills in natural language processing (NLP), but also for its utility in other domains, which include the following:

   1. **Legal:** Technology-assisted review, which is projected to reach an $8.4BN market by 2026, deals with the categorization of legal documents into multiple relevant categories, like specific legal subareas, for the purpose of managing large-scale document discovery processes. A more accurate and explainable classification process can contribute to saving law firms time and money.
   2. **Healthcare & Finance:** Similarly, accurate and explainable classification of medical documents and risk types in finance, such as market and credit, can contribute greatly to the decision-making processes in the $5.5BN healthcare NLP market and the $26BN AI-driven financial services market.

D. **Building Upon Existing Literature:**

   1. **Label Attention Mechanisms:** Label attention mechanisms are a method by which we can give more attention to the most relevant parts of a text, and change our approach based on the label (genre) considered; for example, given a "sci-fi" label, we may consider weight to the words "spaceship" and "alien". While the existing literature appears to have leveraged this method for overlapping conditions in patient records (Liu, Cheng, & Klopfer et al.), I plan to test its efficacy in accounting for interconnected labels (such as "mystery" and "thriller") outside of the healthcare domain and in the more subjective domain of genre classification.
   2. **XAI in Multi-Label Text Classification:** While Explainable AI techniques have been leveraged widely for single-label text classification and sentiment analysis (Yuan, Xu, & Sun), where they've been very effective for highlighting important features in predictions, their use for multi-label classification has been limited. Exploring XAI in the multi-label context will be interesting, as I'll be challenged to explain how certain features, like words or phrases, contribute to multiple labels simultaneously.
   3. **KDE for Label Overlap:** While Kernel Density Estimation is often used for visualization of single-label feature distributions (Liu, Cheng, & Klopfer et al.), it has been underutilized for multi-label analysis. KDE will

be a useful tool in this project in order to observe how keyword distributions in a given genre may overlap with those of a different genre, elucidating relationships between categories and visualizing label co-occurrences.

II. Methodology (More Thorough Breakdown)

A. Dataset

1. I'll be working with this dataset compiled by Ishika Johari on Kaggle, which scrapes 10,000 of the most recommended books on Goodreads and their synopses.

B. Preliminary Exploratory Data Analysis (EDA): I will do an initial EDA on raw data to explore genre distribution and flag any patterns or outliers that may influence pre-processing.

C. Data Pre-Processing:

1. **BERT Tokenizer for Synopses:** The BERT tokenizer will be used to convert the raw text of synopses into tokens, ensuring that each word is effectively represented in embeddings to capture contextual meaning. Each sequence will be padded to the same length.
2. **Text Cleaning:** While BERT handles most raw text effectively, I will perform minimal preprocessing to remove unnecessary characters like excessive whitespace and special characters.
3. **Binary Vector Formatting for Genres:** I'll convert genre labels into a binary vector format, where each position corresponds to a genre, to ensure that each book can be associated with multiple genres.
4. **Dealing with Potential Class Imbalances:** If certain genres show up less than others, I'll apply resampling methods like oversampling of underrepresented genres or undersampling of overrepresented genres.

D. Advanced EDA: I will use several methods from class to investigate the embeddings produced.

1. **K-Means Clustering & Spectral Clustering:** Both K-Means and Spectral Clustering will be applied to the embeddings to identify natural groupings. For instance, spectral clustering will help capture non-linear relationships, revealing if certain genres are inherently more connected.
2. **Gaussian Mixture Models:** GMMs will be used to identify soft clusters, whereby books can belong to multiple clusters with varying probabilities. This will elucidate which genres tend to be connected, and which may cause confusion for the classification model.
3. **PCA for Dimensionality Reduction and Visualization:** I will also attempt PCA for better understanding of the synopsis structure.

E. Model Building

1. **Class Weights:** These will be applied to penalize misclassification of rarer genres to ensure that all genres are equally considered by the model.
2. **Benchmarking through Logistic Regression and Naive Bayes:** The use of BERT is complex and computationally expensive, but provides context-dependent analysis of words, non-linear relationships, the facility to determine a word's meaning based on a different word that appears much earlier in the text, as well as pre-trained knowledge of language semantics prior to fine tuning. Assessing simpler models that do not integrate these benefits will help to assess whether the use of BERT is justified.
    a) **Logistic Regression (LogReg):** LogReg will learn a linear relationship between the features (words and phrases) and the genre labels.
    b) **Benchmarking through Naive Bayes:** Using a strong assumption of feature independence, Naive Bayes will probabilistically classify genres given features.
    c) **Performance Evaluation:** Both baseline models will be evaluated using precision, recall, and F1-score.
3. **Deep Learning with BERT**
    a) **Fine Tuning for Multi-Label Classification:** While BERT is already trained on general language features, I will replace the final layer of BERT, which typically deals with hidden states for each token that consider context, with a task-specific layer that will use a sigmoid activation function to produce probabilities for each genre.
    b) **Mitigating Overfitting:** I will utilize the AdamW optimizer, an iteration upon the Adam optimizer that balances new learning with previous learnings, and handles weight decay and prevents overfitting by regularizing the model coefficients. I'll also utilize a manipulated dropout rate, which will randomly set some neurons to zero and force the network to be versatile in the pathways that it relies upon.
4. **Label Attention Mechanisms:** While BERT utilizes self-attention mechanisms, which allow each word to be interpreted in the context of its relationship with other words in the sequence, I will implement additional post-BERT layers that understand the weights of relationships between genres (labels), allowing the adjustment of predictions accordingly. For example, "thriller" texts may share similarities with "mystery" texts.

F. **Explainability with XAI**

1. **SHAP (SHapley Additive exPlanations):** SHAP will be leveraged to understand global feature importance across the dataset, indicating how each word or token contributes to classification decisions. I note here that SHAP tends to struggle to isolate individual feature effects when dealing with complex and high-dimensional data, so I am interested to see how it performs here.
2. **Evaluation of Explainability:** Using SHAP results, I'll visualize the model's explanations for classifications and compare them with intuitive explanations; for example, if the model identifies "murder" and "detective" as key indicators of the "mystery" genre. I'll additionally assess SHAP's consistency and fidelity through verification of whether the explanations stay stable across similar inputs.

G. Evaluation Strategy & Model Validation

1. **Performance Evaluation:**
    - a) **Tuning:** Given the multi-label classification nature of the project, I'll tune probability thresholds for each genre using techniques like Youden's index.
    - **b) Precision, Recall, & F-1 Score:** The model will be evaluated on its ability to identify relevant genres without missing key categories (recall) or misclassifying the text with unrelated genres (precision), assessing the model across each genre individually as well as with micro/macro averages.
    - **c) Additional Metrics:** I'll also assess the model with Hamming Loss, i.e. on the basis of the fraction of labels that are incorrectly predicted, providing greater insight into the model's performance.
    - **d) Model Calibration:** I'll utilize the Brier score metric and visualize calibration plots, to ensure that the model's predicted probabilities correspond accurately to genre likelihood.
2. **Cross-Validation Techniques:**
    - **a) K-Fold Cross-Validation:** This will be done to ensure that the model can generalize across different portions of the dataset and avoid overfitting.
    - **b) Bias-Variance Tradeoff Analysis:** The model's complexity will be adjusted to manage the bias-variance tradeoff. In the event of BERT overfitting, I can use dropout and early stopping to find the optimal balance.
3. **Error Analysis:** I'll conduct a detailed error analysis to identify genres or synopses where the model tends to make systematic errors. Through analysis of the SHAP values of these misclassifications, I hope to understand whether my model misinterprets specific keywords or

contexts, in order to guide improvements to feature engineering or model architecture in future projects.

H. Visualization & Analysis

1. **Kernel Density Estimation (KDE) for Genre Visualization:**

    a) **Understanding Feature Distributions:** KDE will be used to visualize the distribution of embeddings for different genres, assisting in the identification of genre overlap and distinctions. Here, similar to the Professor's example of the "signature of crime", we can obtain the "signature of a genre".

    b) **Exploring Genre Relationships:** Significant overlap between certain genre pairings (e.g. "romance" and "drama"), as elucidated by KDE and the above methods, can explain challenges in classification.

2. **XAI Visuals:** I'll generate (1) heatmaps to visualize which parts of a book contribute most to the genre classification as well as (2) attention maps from BERT to provide clarity regarding the prioritization given to different parts of the text in the model's classification decisions.

III. Notes on Feasibility

A. **Clustering & GMM:** Should K-Means or Spectral Clustering prove to be computationally infeasible, I will prioritize Gaussian Mixture Models to identify probabilistic overlaps between genres.

B. XAI: I will initially focus on the implementation of SHAP to understand global feature importance. If feasible, I will incorporate LIME for local explanations.

IV. Specific Feedback Solicited: In peer reviews, other than general feedback about this report's components, I would also appreciate any insights on the items below. Thank you in advance!

A. I wanted to challenge myself with this project, but do appreciate that due to computational expense or pushing outside of the boundaries of the syllabus, the project may not work out. Are there any "down-scaled" versions of this project that you think would still be effective for the purposes of this assignment?

B. If encountered, challenges & solutions faced when dealing with multi-label classification.

C. If relevant, other domains that you think this kind of model could be useful in.

D. Whether anything in my methodology or evaluation seems insufficient or "overkill", as I'm eager to leverage topics in the syllabus but need not be repetitive.

E. If you have worked with imbalanced datasets in multi-label contexts, any challenges & solutions you found.