FINAL GROUP PROJECT SUMMARY

Tashfeen Ahmed; Jiaxuan Wang; Xiaorong Tian; Zhiming Zhang; Siqi Wang

The Objective of the Analysis

We aim to develop a model that can distinguish text produced by AI tools, such as GPT, from those authored by humans. By conducting this analysis, we aim to enhance the transparency and trustworthiness of the digital information landscape. And hopefully, it could be a tool for decision makers in relevant fields to address the challenges posed by AI technology.

For data collection, we utilized two datasets sourced from Kaggle. The first, named the *DAIGT Proper Train Dataset*, features texts generated by ChatGPT or LIama alongside human-composed texts from official training essays and the Persuade corpus. The second dataset, *ArguGPT*, comprises about 4,000 essays generated by 7 different GPT models. We combined the two datasets and randomly selected a balanced sample of 5,000 AI-generated and 5,000 human-generated texts for our further analysis.

Analytics Approach

During data processing, we initially preprocessed the texts by transforming them to lowercase, eliminating new lines, special characters, certain predefined phrases, stopwords, and leading or trailing spaces. For exploratory data analysis, we start form utilizing the ***nltk*** library to identify and quantify the parts of speech (POS) tags present in the texts. Then, we tokenized and lemmatized the text to reduce words to their base or dictionary form for normalization. We then employed ***TfidfVectorizer*** from ***scikit-learn*** to transform these POS-tagged texts into a TF-IDF matrix. This matrix reflects the importance of each word within the documents and across the corpus, assigning each text a row and each POS tag a TF-IDF score in its columns. This ensures that each POS tag contributes equally to the analysis. It should be noted that TF-IDF features were generated for both non-lemmatized and lemmatized texts to allow for a comparison of their performance in the model. We analyzed these TF-IDF scores by summing them for designated POS tags and categorizing them accordingly. Finally, we conducted t-tests to assess the differences in TF-IDF scores for POS tag categories between AI-generated and human-generated texts. Our findings indicated that human created texts employ a higher frequency of nouns, verbs, determiners, adjectives, auxiliary verbs, coordinating conjunctions, and particles than texts created by ChatGPT. Consequently, these POS tags were selected as features for building our model. In addition to the advanced technique of generating TF-IDF scores for specific parts of speech (POS) tags, traditional TF-IDF analysis was also performed on the preprocessed texts without POS categorization. By doing so, we created a comprehensive set of features that includes both the general importance of words (via traditional TF-IDF) and their grammatical significance (via POS-tagged TF-IDF scores). This dual-feature strategy ensures that our model benefits from the semantic richness of the text as captured by standard TF-IDF, as well as the linguistic and syntactic insights provided by the POS-tagged TF-IDF features.

Following the data preprocessing phase, we employed two different approaches to develop the final model: the K-Nearest Neighbors (KNN) and the Naïve Bayes methods. We prepared two feature sets, one derived from non-lemmatized text and the other from lemmatized text, both augmented with standardized linguistic features. The dataset was then divided into training and

testing subsets for both text versions, allocating 20% of the data to the testing sets. For the KNN model, we utilized the ***KNeighborsClassifier*** from the ***sklearn.neighbors*** module, and for the Naïve Bayes model, we employed the ***MultinomialNB*** from the ***sklearn.naive_bayes*** module. Finally, we assessed each model's performance by calculating accuracy, precision, recall rates, and the AUC.

Expected Impact

We establish this model to differentiate AI-generated text from human-written content. Tools like this can identify AI-generated content, distinguishing the increasing false news or incorrect information online. Additionally, it could also be used in education to maintain academic integrity. Through these application scenarios, our tool could help maintain legal and ethical boundaries, enhancing the reliability and authenticity of information for the public in the long run. From another perspective, clearly distinguishing text sources can effectively eliminate public doubts about AI development, achieving a win-win situation for both humans and AI.

In the area of education, our model could provide educators with an effective tool to differentiate between student-generated work and AI-generated submissions. The application of these tools can effectively prevent the occurrence of plagiarism problems, and help develop an atmosphere of honesty and trustworthiness in the field of education.

From a legal and ethical perspective, such tools can also be applied in document authenticity review processes. This measure can support the reliability of legal documents and ensure that relevant documents are not contaminated by AI-generated content. Such tools can also serve as a deterrence to prevent AI from being used by bad actors to create fraudulent documents. This helps build a more transparent and trustworthy digital environment.

Our current model is still relatively basic, but using it as a foundational framework allows for the development of more complex models. These models, serving as testing tools, aid in identifying subtle differences between artificial intelligence and human text outputs. As such, they can serve as testing tools for artificial intelligence text generation techniques, assisting in overcoming current limitations in mimicking human text production and enabling further research and innovation in the field of natural language processing.

Considering the public's attitudes towards artificial intelligence, similar tools can also have a guiding role in mass media. On one hand, the public expresses high enthusiasm and anticipation for the rapid development of artificial intelligence, while on the other hand, concerns about AI invading and replacing human intelligence also lead to debates filled with both trust and skepticism. Having reliable related models can help the public recognize the complexity of modern information consumption, thus aiding in shaping interaction patterns among digital-age audiences.

In summary, our project develops an AI differentiation model with significant implications for enhancing digital content integrity, upholding academic standards, and confronting legal and ethical challenges. In the long term, the continuous refinement of such models will be instrumental in measuring and advancing natural language processing. Possessing reliable models like these will be crucial in reducing public skepticism and enhancing understanding and engagement with AI.