City, University of London
School of Science and Technology

PROJECT REPORT
2024

An Analysis of the Validity of Explanations of Explainable AI (XAI) for Multilingual Text Classification

Syed Muhammad Mohsin
230013462

Supervised by: Dr Andrey Povyakalo

December 9, 2024
London, United Kingdom

Contents

List of Figures

List of Tables

Declaration

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed: Syed Muhammad Mohsin

ABSTRACT

Explainable Artificial Intelligence (XAI) has become quite relevant with the increase in the number of complex machine learning systems or models , particularly in domains like journalism , education , and online safety. XAI is also critical in binary classifications like AI vs Human generated texts , by giving insights on the model's predictive patterns. Most popular methods of LIME and SHAP enable post-model interpretation by identifying influential features yet these approaches often falter in multilingual environments and semantic diversity in the datasets. The research evaluates fidelity, robustness, and interpretability to determine how well explanations capture model behaviour, their stability under input variations, and their comprehensibility. Key findings reveal limitations, such as incomplete explanations, inconsistent sensitivity to input changes, and instability in handling non-Latin languages like Urdu. Stability and sufficiency analysis expose gaps in accurately interpreting semantic nuances and linguistic diversity, while counterfactual and fidelity assessments show weak alignment between surrogate and original models. While XAI explanations offer valuable insights, their reliability varies based on dataset complexity and perturbation levels. The findings emphasize the need for improved XAI tools to enhance interpretability and trustworthiness in multilingual and high-stakes applications, ensuring robust and meaningful explanations across diverse contexts.

CHAPTER 1

Introduction

1.1 Background and Context

The rapid advancements in artificial intelligence (AI) have brought about significant changes across industries, from healthcare and finance to education and media. Among the myriad of AI applications, text generation and classification have gained particular prominence, enabling the creation and analysis of textual content at an unprecedented scale. OpenAI's GPT and Google's BERT and other models like these have demonstrated remarkable capabilities in generating coherent, contextually relevant text. However, these capabilities raise questions about transparency, trust, and accountability, especially in contexts requiring human-like decision-making.

With the increasing prevalence of AI generated text in domains such as journalism, education, and content creation, distinguishing between human- and AI generated text has become a crucial task. Text classification systems have been deployed to tackle this problem, but they often act as opaque "black boxes," providing predictions without offering insights into how decisions were made. The rise of Explainable AI (XAI) seeks to address this issue, offering methods that help interpret the internal workings of machine learning (ML) models and elucidate the rationale behind their predictions.

1.1.1 Explainable AI in Text Classification

Explainable AI is a multidisciplinary field focused on making the decision-making processes of machine learning models transparent and interpretable. XAI aims to bridge the gap between model predictions and human understanding by providing explanations that articulate why a particular decision was made. In text classification, XAI methods like Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive Explanations (SHAP) are frequently employed to highlight the textual features most influential in a model's prediction. For example, in classifying a news article as "AI generated," XAI tools might identify repetitive phrases, unusual vocabulary, or syntactic patterns as contributing factors.

The adoption of XAI in text classification extends beyond academic interest; it addresses real-world challenges such as misinformation, plagiarism detection, and content moderation. In journalism, distinguishing AI generated text from human-authored content is essential for maintaining credibility and trust. Similarly, in education, AI generated essays pose challenges to traditional methods of plagiarism detection, necessitating robust systems to classify and explain the nature of the text. Despite their utility, XAI methods face significant challenges in practice, including instability under perturbations, computational inefficiencies, and difficulties in handling multilingual or semantically diverse datasets.

1.1.2 Human and AI Generated Text: The Problem

AI generated text has become increasingly sophisticated, often indistinguishable from human-authored content. Models like GPT-3 and GPT-4 leverage vast datasets and advanced neural architectures to produce fluent and contextually appropriate responses. While this progress has enabled breakthroughs in customer service automation, creative writing, and education, it has also raised concerns about misuse and ethical implications. AI generated text can be exploited for spreading misinformation, creating spam, or fabricating academic work.

The challenge lies not only in classifying whether text is human or AI generated but also in understanding the decision making process behind such classifications. XAI methods provide a means to demystify these decisions, revealing whether models rely on stylistic cues, lexical patterns, or other attributes. For instance, AI generated content may exhibit repetitive sentence structures, an overuse of formal phrases, or statistically improbable word combinations. However, these patterns are not always consistent, particularly in multilingual contexts where linguistic nuances vary significantly. Addressing this variability requires explainability methods that are robust, semantically aligned, and capable of generalizing across diverse languages and tasks.

1.1.3 Challenges of Explainability in Multilingual Contexts

The multilingual and diverse nature of textual data adds a layer of complexity to XAI for text classification. While traditional XAI methods often rely on token-level attributions (e.g., identifying key words or phrases), these approaches struggle with semantic alignment across languages. For instance, words that are contextually important in one language may not carry the same weight in another due to differences in grammar, syntax, or cultural connotations. Moreover, AI generated text often mimics human linguistic diversity, making it difficult for models to identify universal features that distinguish it from human-authored text.

Embedding-based methods, such as those utilizing Multilingual BERT (mBERT), attempt to overcome these challenges by focusing on semantic relationships rather than surface-level features. For example, an mBERT-based explainability framework might highlight stylistic tendencies (e.g., formal tone, repetitive sentence structures) that are consistent across languages. However, such methods require significant computational resources and are prone to bias if the underlying embeddings fail to capture cultural or linguistic nuances adequately.

1.1.4 Evaluating the Validity of XAI Explanations

The validity of XAI explanations is a central concern in assessing the reliability and trustworthiness of text classification models. Validity encompasses several dimensions, including robustness, fidelity, and interpretability. Robustness refers to the stability of explanations under minor input changes, such as synonym replacement or paraphrasing. For example, if replacing "happy" with "joyful" in a sentence causes an XAI method to produce drastically different explanations, the method's robustness is called into question.

Fidelity, on the other hand, measures how well explanations align with the model's true decision-making process. For instance, in classifying AI generated text, an explanation that highlights stylistic markers like repetitive phrasing or formal tone should accurately reflect the features driving the model's prediction. If explanations focus instead on irrelevant attributes, such as common words ("and," "the"), their fidelity is compromised.

Interpretability, the third dimension, assesses whether explanations are comprehensible and actionable for end-users. In applications like content moderation or plagiarism detection, explanations must be clear enough for non-technical stakeholders to make informed decisions. For instance, a teacher reviewing a student's essay flagged as AI generated might rely on XAI to highlight inconsistencies in tone or coherence, enabling them to understand why the model reached its conclusion.

1.1.5 Implications for Real-World Applications

The intersection of XAI and text classification has far-reaching implications for real-world applications. In journalism, the ability to distinguish AI generated articles from human-authored ones is crucial for maintaining public trust. Mislabelling human-authored content as AI generated or vice versa can damage reputations and undermine confidence in AI systems. In education, where students increasingly rely on AI tools for writing assistance, detecting AI generated essays requires explainability methods that are transparent and actionable for educators. Similarly, in content moderation, robust XAI systems can help identify AI generated spam or harmful content, ensuring safer online environments.

However, achieving reliable XAI explanations in these domains requires addressing the challenges posed by multilingual datasets, adversarial inputs, and semantic complexity. Embedding-based methods, adversarial testing, and user-centric evaluations offer promising avenues for improving explanation validity. For example, embedding-based approaches can capture universal stylistic patterns across languages. User-centric evaluations, such as surveys or case studies, provide insights into the interpretability and usability of explanations for non-technical stakeholders.

The integration of XAI in text classification represents a critical step toward enhancing transparency, trust, and accountability in AI systems. By providing interpretable and actionable explanations, XAI methods empower users to understand and trust machine learning models, particularly in sensitive applications like distinguishing human- and AI-generated text. However, the challenges of robustness, fidelity, and semantic alignment underscore the need for continuous innovation in XAI methodologies. As the field evolves, addressing these challenges will be essential for ensuring that XAI systems remain reliable and effective in diverse real-world contexts.

1.2 Research Objectives

The overarching aim of this research is to explore and evaluate the validity of explainable AI (XAI) explanations in distinguishing between human and AI generated text. This investigation involves examining the strengths, limitations, and practical applications of XAI techniques, focusing on the interplay between transparency, robustness, and interpretability in text classification tasks. The specific objectives are as follows:

1. **Evaluate XAI Methods for Text Classification**
   - Assess the performance of prominent XAI methods, such as LIME and SHAP, in providing meaningful and interpretable explanations for text classification models.
   - Investigate how these methods highlight distinguishing features between human- and AI-generated text, such as lexical diversity, syntactic structures, and stylistic patterns.
2. **Analyse Robustness and Fidelity of Explanations**
   - Examine the stability of XAI explanations under input perturbations, including synonym replacements and adversarial edits.
   - Measure fidelity to ensure explanations align closely with the actual decision-making processes of the underlying model.
3. **Address Multilingual and Semantic Challenges**
   - Evaluate the effectiveness of XAI methods in multilingual contexts, focusing on their ability to provide semantically aligned and consistent explanations across diverse languages.
4. **Develop and Validate Evaluation Metrics**
   - Utilize metrics such as sufficiency, sparsity, Jaccard similarity, sensitivity, and KL-divergence to assess the validity of XAI explanations.
   - Propose enhancements to existing metrics or frameworks for a more comprehensive evaluation of explanation quality.

5. **Contextualize Findings for Real-World Applications**

    - Investigate the implications of XAI explanations for practical use cases, such as education (plagiarism detection), journalism (AI-authored content detection), and content moderation.
    - Provide actionable insights for improving trust and usability in XAI systems.

1.3 Report Structure

This dissertation is structured to tackle the aforementioned research questions methodically. For this purpose it's divided into following chapters:

- Chapter 2: Literature Review
- Chapter 3: Research Design and Methodology
- Chapter 4: Data Pre-Processing, Exploratory Analysis, Feature Engineering and Code Implementation
- Chapter 5: Model Selection, Implementation and Evaluation
- Chapter 6: Comparative Analysis of Models
- Chapter 7: Use of Explainable AI (XAI) to Understand The Best Performing Model
- Chapter 8: Conclusion
- Appendix

CHAPTER 2

Literature Review

2.1 Overview

Explainable AI or XAI has emerged as an area of interest as has been the development of advanced and complex machine learning models in areas such as education, journalism, and online safety. XAI is especially critical in text classification tasks; understanding model predictions when classifying between human-written text and content created by AI. Methods like LIME and SHAP are quite common to use for post-model interpretation to discover the set of features which affects model result. However, these methods prove to be weaker in multilingual environment and with semantically different dataset, due to the differences in languages and the semantics of the data provided.

In order to make XAI-generated explanations easy to understand and trustworthy, fidelity, robustness, and interpretability measures are used. These measures assess if and how the explanations really express the characteristics of the model, what happens when a small change is made to the input, and how interpretable the explanations are to the end users. To this end, this work is aimed at searching for evidence on the accuracy of XAI explanations in the context of identifying human-written text compared to Artificial Intelligence generated text. It solves issues including providing coherent and precise descriptions across different languages and for different data sets. Addressing these concerns is the main purpose of this work, as this research seeks to enhance the stability and practicality of AI models for various high-stakes applications.

2.2 Research Papers, Articles and Related Work

2.2.1 Why should I trust you?: Explaining the Predictions of Any Classifier - Ribeiro, M. T., Singh, S., Guestrin, C. - 2016 [8]

Based on my analysis of Explainable AI (XAI), I discovered that Ribeiro et al. (2016) proposal of Local Interpretable Model-Agnostic Explanations (LIME) was progressive and useful in multilingual text classification. LIME also provide a simple technique to explain the prediction from complicated machine learning algorithms. This is done by applying a minor random perturbation on the inputs and deriving an easily interpretable surrogate model that can replicate the behaviour of the original complex model in the immediate neighborhood of the input space. Such flexibility of LIME makes it usuable in a wide range of domains and tasks including text classification. In such cases it assist in determining featuresâwords or phrases that have considerable impact on a model's output.

For instance, in a multilingual spam detection, LIME can explain key words such as 'gratis' which means free in Spanish or 'offre' which means offer in French as pointing to a message as spam. Its applicability to different languages and models is a huge strength of this model. However, as I was researching, I noted several issues that arise with reliability and accuracy of the explanations and this is especially true when working with multilingual data sets. Ribeiro et al. mentioned that obtaining input perturbations for LIME can be problematic because LIME could output inconsistent explanations when faced with small changes.

This problem becomes even more acute in the case of studying multilingual contexts. In particular, the perturbations themselves depend on the linguistic differences existing on the levels of synonyms, word order and syntactic structures, and these differences can affect the importance of specific words or phrases. For example, moving a text from one language to the other can distort the meaning of some words as a result distorting the coherence of explanations. The subtlety of the words can also create a difference and this kind of difference may consist of swapping "cheap" and "affordable" or "offer" and "promotion". Such fragility poses a large problem in multilingual tasks, where language features differ substantially. Moreover, a syntactical variation between professional prose and the use of AI programming results in further difficulties in producing valid explanations for translated text.

This raises a crucial challenge of how to establish the reliability of the LIME-generated explanations for non-English speaking users. To this end, there is a call for metrics that assess the stability and, in particular, the cross-linguistic consistency of the explanations. If LIME does not have such tools, then it can be difficult to have faith in the interpretations it offers and particularly complex instances such as distinguishing human writing from write ups done by artificial intelligence. However, it is also crucial to consider that the explanations provided are culturally and linguistically aligned in order for the XAI technologies to be used more widely.

Consequently, LIME is effective in explaining decisions made in the text classification tasks, yet sensitive to changes in input data and has limitations in multilingual applications. To improve LIME, there is a need to increase the strength of the proposed methods and refined validation measures. These challenges are the focus of my research, which should advance the dependability and usability of XAI in multilingual settings, where language complexity plays a substantial role in models' outcomes.

2.2.2 A Unified Approach to Interpreting Model Predictions - Lundberg, S. M., Lee, S.-I. 2017 [3]

In my study, I first explored Lundberg and Lee's (2017) SHapley Additive Explanations (SHAP), a framework for feature attribution based on cooperative game theory. SHAP is designed to calculate Shapley values in order to ascertain how each feature contributes to the output of a specific model. While, LIME interprets localized features, SHAP provides both local as well as global explanation making it more informative about the model's behaviour. This dual perspective makes SHAP particularly effective in text classification tasks because we obtain views from both sides of the model. For example, I applied SHAP on to analyze sentiment analysis specific words like excellent or terrible impact on a model.

In this case, a notable strength of the proposed method, SHAP, is its reliability or constancy in assigning higher value to features that are likely to be more important. This property is more important in differentiating between human-generated text and AI-generated text because of these small variations in style or syntax used by such models. That's why, the distinctive features which SHAP allows to capture improve the interpretability of models. Nevertheless, the difficulties that emerged during my investigation of the method also concern some drawbacks of using SHAP, such as high computational complexity. As the increasing number of observations or use of multiple languages for AI-generated content analysis, the process of generating explanation becomes highly computationally intensive and therefore non-scalable.

However, these limitations make SHAP a useful tool in my work, especially because it has the capability of modeling complex feature interactions. This capability is highly helpful for validating explanations in the tasks where recognizing the impact of particular features is significant. For instance, in cases of multiple languages understood by the model, SHAP can reveal whether certain lexical or literary features affect the outcome across different languages as well as reveal distinctive features of the model and the datasets used.

In the future, I hope to solve the computational issue that is tied to SHAP, especially in applications that involve multiple languages. Thus, with the help of the optimization or with the help of other strategies, I want to reveal all the possibilities of utiliSing SHAP in my research to the maximum extent.

2.2.3 The Mythos of Model Interpretability - Lipton, Z. C. - 2018 [2]

In my study, I observed that Lipton (2018) provides important criticisms of interpretability in machine learning when assessing XAI approaches. Lipton introduces a conceptual framework that distinguishes between three key aspects of interpretability: , which are transparency (the ability to understand how models function), simulatability-the ability to mimic how the model behaves, and post hoc - explanations developed after the models make predictions. He then explains that there is need for one to identify the kind of interpretability that is needed based on the need of the end-user but then warns that it could be misleading to have a very simple and straightforward way of interpreting a model.

This perspective has been particularly useful for my task of separating content written by a human from that written by AI. In such tasks the explanation beyond the frequency of individual words are required to capture stylistic and linguistic features such as coherency, tonality and syntactic structure to name a few which are characteristic of AI generated text. Such explanations may fail to capture these feature details, making the understanding of how the model arrives at an answer relatively basic.

This paper is crucial within my workflow due to Lipton's argument on how explanations of machine learning should resemble user expectations. While conducting research, I have focused on developing not only semantic but also practical and comprehensible by the end consumer explanations. That is why I strive to provide the insights, which depict the intricacies of the linguistic processes without distorting them and to make the interpretations worth doing.

To sum up, it is crucial to thank Lipton for his framework that I used to define the proper approach to creating and testing the XAI techniques with reference to the high-stakes activities such as identifying the AI-generated text from the truly human-written posts. It is for this reason that his focus on accuracy, comprehensibility, and context has remained a cornerstone for my thinking about making AI systems more explainable.

2.2.4 Fooling LIME and SHAP: Adversarial Attacks on Post-hoc Explanation Methods - Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H. - 2020 [9]

To this regard, I wanted to explore more on the vulnerability of post-hoc explanation methods that were recently described by Slack et al. (2020). They found out that through adversarial perturbations, their explanations polluted the actual decision-making process of machine learning models. For example, in text classification tasks, the adversarial instances might cause LIME to identify irrelevant words such as "hello" as the features that are most influential in the model's decision, resulting in explanations that could be misleading or flat out wrong.

Such vulnerability highlights the importance of effective XAI explanations, most importantly in the fields that require high reliability like identifying texts written by an AI. Noting this, I have made the stability of explanations one of the most important emphases in my work. That is why it is imperatively important to ensure that explanations do not degrade under adversarial conditions, and remain as valid and trustworthy as the AI models they explain.

Concerning this challenge, I have included adversarial testing in the evaluation process to meet the need. This entails intentionally making slight adjustments to input data and evaluating their effects on the produced explanations. Stability of these explanations is measured using other related parameters for example, Jaccard similarity and cosine similarity. I examine the differences in the explanation of instances before and after the perturbations to determine whether the changes affect the stability of the model's interpretations.

This approach is even more important in practical situations where adversarial perturbations can potentially threaten the applicability of explanations. In all the tasks such as the AI-generated text detection, the explanations must continue to be reliable and explainable even in cases of more input perturbations. The objective is to make it possible that such an explanation should be correct and should not be able to be fooled by any adversarial input so that the end user should be able to have trust in AI and the various applications that may be developed.

2.2.5 A Multidisciplinary Survey and Framework for Explainable AI - Mohseni, S. and Zarei, N. and Ragan, E. D. - 2020 [6]

In my study, I discovered that Mohseni et al.'s (2020) article on explainable AI (XAI) was quite helpful. Some of them, while having given a broad overview of XAI techniques and a framework to evaluate them, allowed for the organization of explanations by their validity and efficiency. The authors in their work divided different XAI methods into groups based on the interpretability aims and presented the criteria for assessment which include sufficiency, fidelity and dimensionality. These metrics are all useful in deciding whether an explanation is a good representation of a model's decision making process and what factors are most important in its decision making.

Of the presented concepts of the proposed framework, its focus on the necessity to balance between easy and complete explanations was particularly useful when dealing with text classification. Concretisations which are simpler or sparser may well limit identification to easily identifiable items such as key words and phrases but might miss out on secondary items for which context might argue are relevant. That is why more detailed descriptions may contain additional parameters, improving the performance at the cost of readability. This trade-off was an important aspect of my work especially while analysing explanations derived from approaches such as LIME.

The evaluation criteria that I used in my research were informed by the framework of Mohseni et al. For instance, I applied sufficiency tests to examine if the elimination of the features in the explanation greatly changed the model's prediction. An ideal sufficient explanation should allow one to point out the differences in the output whenever major facets are left out. Similar importance was attributed to fidelity metrics that assessed how the explanations resembled the internal working of the model. These tests confirmed that the explanations were not only realistic but also accurate with respect to the real structure of the model.

Through the adoption of these evaluation metrics, it was easier to provide a more thorough and robust approach to the confirmation of the explanations provided in my study. Following the approach described by Mohseni et al., this alignment helped to make the explanations I generated both understandable and genuine representations of the model. Furthermore, this structured method improved the explanatory reliability and stability while addressing AI-generated text detection which also improved the trustworthiness of the XAI methods.

2.2.6 A Comparative Analysis of Counterfactual Explanation Methods for Text Classifiers - Stephen McAleese and Mark Keane - 2024 [4]

Of the papers I reviewed, the contributions to the area of robustness and reliability of post-hoc explanation methods such as LIME and SHAP were most insightful to me. The authors provided key concerns like the discrepancy between the importance of features and that of tokens at the individual word level. For instance, in sentiment analysis, models offer explanations that are antithetical to observed behaviour at the token level, for example, sometimes completely missing sentiment-bearing words. This misalignment becomes particularly problematic in multilingual use cases where it's possible to have a highly nuanced understanding of one language and almost no understanding of another. These results emphasize the importance of having more reliable approaches in assessing the decision-making functionality of a model.

The authors also incorporated methods for checking the stability of explanations obtained using LIME and SHAP. Synonym replacements and paraphrasing, which are types of perturbation-based methods, were used to determine the robustness of these explanations. One of the most striking observations made by the authors was that even the least significant modifications to LIME's explanations that the authors performed in some of the experiments could change them significantly. This observation especially shows a weakness in the post-hoc methods and stresses the need for better methods for the assessment of explainability, especially in practical cases, such as the discrimination of text produced by humans and those generated using AI.

This paper also differentiated between post hoc interpretability techniques like the saliency maps with inherent interpretability techniques like the attention weights. Although attention mechanisms can give some degree of global interpretability, the authors noted that they are not entirely consistent with the model's actual decision-making process. This critique is consistent with Lipton's (2018) which held that methods that focus on transparency may not capture the workings of the model to the maximum. I have reported the results of applying these ideas to my own research, regarding their practical implications for determining explanation validity when post-hoc explanations are involved in text classification tasks.

This paper greatly expanded my knowledge of the issues that arise when applying explainable AI (XAI) to text classification, and especially when using large complex models, such as large language models (LLMs). This principle of making explanations robust and faithful has been a principle throughout my practice. In this way, I will improve the validity of XAI methods and specifically untangle the functioning of the model from text authored by humans in order to improve the reliability of the explanations that I assess.

2.2.7 Evaluating the Faithfulness of Explanations Generated by Explainable AI Models - Mohseni, S. and Zarei, N. and Ragan, E. D. - 2020 [5]

The paper ”Evaluating the Faithfulness of Explanations Generated by Explainable AI Models” (OpenReview) interrogates faithfulness in XAI explanations as the extent to which an explanation is in line with the model's actual reasoning process. This notion is especially close to my research, as I concentrate on maintaining the consistency and correctness of the explanations while sorting out the contents generated by people and AI. In this paper, the author proposes an extensive assessment framework, which includes sufficiency, sensitivity, and proxy fidelity, that I consider as highly helpful to evaluate the legitimacy of SHAP, LIME, and other post-hoc explanation approaches.

Some of the major points are that many of the modern XAI techniques, such as SHAP and LIME, do not always provide adequate information on how predictions are made. For example, in tasks such as text classification including sentiment analysis, I have seen situations where LIME assigns high importance to tokens like stopwords (for example ”and” or ”the”) that actually, have little to no influence on sentiment. This mismatch between what is explained and the true working of the model erodes the credibility of explanations, particularly in high-risk use cases, such as detecting AI-written content.

The paper also includes diagnostic tests for faithfulness: removing the features, which were marked as important by the explanation method, and comparing the change in model's output. In case the prediction does not change much when these features are stripped off, the explanation is considered unfaithful. I also found this approach useful and have used it in my own research. In my experiments, I concluded that SHAP explanations are more accurate in tasks that require the understanding of complicated language features, such as excessive usage of phrases when writing by an AI-powered tool. In contrast, LIME explanations tend to concentrate purely on the insignificant features of the data such as word frequencies, thus making them untrustworthy.

Furthermore, the paper also discusses the applicability of counter factual based evaluation techniques for explanation validity assessment. This is well aligned with my interest to probe the model on small input changes and especially the explanations. For instance, changing the word 'urgent' to 'important' in the subject of a spam email should lead to an explanation that corresponds to the model's prediction (for example, 'spam'). But the authors who introduced those methods, SHAP and LIME, state that all these methods do not scale counterfactuals of the sort and do not adjust for semantically sensible alterations. This insight has been useful in defining the criteria I use when assessing the explanatory robustness of an explanation when the input is slightly changed at the edge case of image recognition and generation of natural language text.

The paper also examines the various shortcomings of sparse and complete explanations. The advantage of sparse models is interpretability, but it comes at a cost of ignoring potentially important details involved in the predictive process. The authors also conclude that they can be realized as a combination of primary and secondary features with clear and accurate explanations. This is a perspective that I adopted while conducting my research: I strive to produce explanations that incorporate the key and contextually relevant aspects of the model's thought processes.

When applied to my work, the following insights will help in overcoming the gaps in the current explanation methods. Thus, I want to make sure that the explanations incorporated in my research are sound, accurate to the model's decision making algorithm, and flexible enough to accommodate minor differences in inputs.

2.2.8 How Reliable and Stable are Explanations of XAI Methods? - Jose Ribeiro and Lucas Cardoso - 2024 [7]

The paper entitled ”Towards Robust Explainability for AI Systems” is closely related to my research on XAI for human/AI text classification because the authors look into the issue of explainability methods’ sensitivity to real-world perturbations and adversarial settings. In this case, we see a major drawback of existing post-hoc XAI techniques, such as LIME or SHAP, which can become unstable and change the localization of features even after minor input perturbations, like paraphrasing or replacing some words by their synonyms. This is quite problematic in the cases where such distinctions are important as, for example, the differentiation between human-generated and AI-generated text. For instance, the paper shows how changing a word such as ”critical” to ”important” in the review text can make a shift from the focal keywords to peripheral features, and reduce the reliability of the explanations.

This goes well with my experience when validating the effectiveness of LIME and SHAP to explain text classification models. The paper indicates that such inconsistencies, especially in high risk scenarios, reduce confidence in the explanations and in turn reduces their value for decision making. This insight is good for my research because it reaffirms the importance of finding better explanation stability in my research since small changes on the input of the explanation model may lead to drastic changes in the output. The second major contribution of this paper is the concept of giving adversarial robustness to the XAI explanations proposed. The authors also show that Adversarial examples can be exploited to mislead methods such as LIME to focus on irrelevant or even neutral words, thus misleading the understanding of the model’s decision-making process. In my work, where the objective is to make explanations reliable for differentiating between AI-generated text, this discovery is imperative. It proves that adversarial testing should be included in the evaluation framework in order to understand how explanations react to manipulations intended to mislead the model.

The paper also discusses explanation smoothing techniques which can alleviate these problems. This is done by accumulating the explanations across different perturbed instances rather than focusing on a single output; this way, some sensitivity to changes is alleviated. In my research, I have started with applying this approach and generating several perturbed instances and averaging over their explanations will enhance stability and reliability. I believe that this technique improves the stability of the generated explanations, which is a valuable asset when coming to practice since real-world data input is typically noisy or adversarial. In addition, the paper considers the problem of interpretability in multilingual environments. The authors also note that prior XAI approaches that present attributions at the token level are highly influenced by specific language features and do not provide semantically coherent explanations across the languages. As such, it is quite applicable to my study, where I identify the difference between human-written and artificial intelligence-written text using multilingual datasets. In the near future, I intend on further seeking solutions based on embeddings like BERT, which would provide contextual attributions, and fail to notice semantic relations. Such techniques could go a long way towards enhancing the validity of the explications, particularly when handling a range of language data.

In summary, this paper has given me a better insight for enhancing the stability and reliability of XAI explanations in the context of my research. Therefore, by incorporating adversarial testing, explanation smoothing, and multilingual alignment into the framework of evaluation, I can cover the main limitations that are concerned with stability and the ability to generalize the explanations for different data sets and situations.

2.2.9 Evaluation Metrics in Explainable Artificial Intelligence (XAI) - Coroama, Loredana and Groza, Adrian - 2022 [1]

In the article Evaluation Metrics in Explainable Artificial Intelligence (XAI), the authors give a detailed description of the most important evaluation metrics that can be applied to XAI methods. The paper provides these metrics in a structure across different dimensions: robustness, fidelity, interpretability and usability which can be implemented in my research directly. Since I am concerned with verifying XAI explanations for the human/AI text classification, the findings of this paper are informative and pertinent for me to check accuracy and meaning of the explanations I provide.

One of the main measures that were considered in the paper is the measure of robustness. The authors also focus on the explanation being robust to the perturbations or adversarial situations. It is most important in text classification problems for which the change in feature importance values can be remarkably sensitive to very small variations, for example, substitution of a synonym. In my own work, I have used some of the XAI techniques such as SHAP and LIME and found out that they do not offer realign explanations when the input word is slightly changed. For instance, the differences in the text produced by AI and human might be minimal, but the differences in the explanation for such variations will be completely off, making them inaccurate. This paper's emphasis on this metric makes me want to include robustness testing to my evaluation plan in order to check that the explanations I produce are not easily flawed by small changes in the inputs.

One of the important features that this article under discussion covers is the concept of fidelity that is of the primary significance for my study. The authors identify two important aspects of fidelity: both in terms of sufficiency and comprehensiveness. Sufficiency is the primary concern when considering what aspects must be included to make a prediction accurate while comprehensiveness addresses the question of what other factors are important enough to warrant inclusion. For example, in constructing a sentiment analysis model, an explanation that brought out the word "happy" but ignored the negation by "not" would not pass the fidelity test. This relates to my research where it was evident that for the explanations to be useful, one has to provide correct and comprehensive definition of what separates human and AI written text. This is why I intend to employ the sufficiency and comprehensiveness criteria in the assessment and enhancement of the explanations I produce, in order to make them both detailed and informative.

It also presents usability metrics that are important to guarantee that XAI explanations are effective and comprehensible to non-technical audiences. To assess whether or not explanations can be interpreted by the users and can be applied in practical endeavours, the authors recommend surveys and case studies. As in my own cases in the field of plagiarism or content filtering, it should be very easy for user such as an educator or moderator to understand the reasons why certain activities are being suggested. The knowledge derived from this paper helps me to implement the usability assessment into the framework so that the technicality of the explanation does not overpower the practicality and comprehensibility of the same to the target audience.

Furthermore, the paper describes the issues related to the use of the presented evaluation metrics when working with multilingual data, which is a rapidly developing topic in my research. This is mainly because previous XAI approaches including token-level attributions, do not translate semantically as they are, between two different languages thereby resulting into semantic mis-matches. To rectify these problems, the authors support embedding-based techniques, for example, BERT-based contextual attributions. To address these issues, I want to apply these methods in my work to increase the reliability of explanations and enhance their stability depending on the translations and the context of text classification tasks involving both human and AI.

In general, this review offers a balanced outlook on XAI methods and it has been very useful when defining evaluation criteria for my own work. With inclusion of the other aspect in the evaluation framework, it is possible to have improved robustness, fidelity and usability of the explanations making the evaluation to be more comprehensive thus making the explanations valid for use in real life tasks.

2.3 Conclusion

Therefore, it can be concluded that, based on the presented literature review, I was able to learn the state of research in Explainable Artificial Intelligence (XAI) and its use in tasks like text classification. The evaluated articles demonstrate that reliability, accuracy, and intelligibility should be given special attention when assessing XAI approaches. It is evident that though there have been some promising methods such as SHAP and LIME that offer stable and faithful explanation, these methods fail to perform well under perturbations or adversarial settings. Furthermore, matters including the concordance of the offered explanations with the model's decision-making as well as the usefulness of the explanations to non-technical audiences have been pointed out. It is with such knowledge that this review did not only shape my strategy for assessing the validity of XAI explanations for the differentiation between human and AI produced texts but also highlighted the research gaps that my study seeks to fill. Thus, incorporating information on model robustness, faithfulness, and multilingual compatibility, I hope to improve XAI's effectiveness in practical use cases.

Chapter 3

Research Design and Methodology

3.1 Overview of the Methodology

This study ensures a proper and systematic findings of the explanations of XAI's effectiveness , suitability and robustness. My research offers multiple mixed approaches which integrates both of the quantitative and qualitative approaches. This helps me in evaluating the validity of XAI explanations between my binary classification of Human vs Machine text. This study is structured around three important stages: Modeling and data generation , then producing the explanations or XAI explanations , and lastly validation of results by using the metrics already established and some real-world use cases.

3.2 Dataset Preparation and Preprocessing

My research uses publicly available datasets with abstracts involving many diverse topics , such as social media, news , and academic writing . Following this data gathering , I have done text pre-processing to basically standardize my text which includes multiple sentences , stop words, special characters and upper and lower case letters. Then performed standardization and tokenization of the text retains the key stylistic elements which are crucial to text classification like tone and syntax of the text. For the multilingual part of the data I have used the same modeling technique I used earlier for my english dataset and this is to keep the explanations consistent across languages. The multilingual dataset is rich with around twenty six languages generated from five different GPT models. But I have performed validation tests on only one language 'Urdu' carefully selected from the data set. One aspect of using Urdu is to use a language which doesn't contain any english words.Furthermore, it was also difficult to incorporate more other languages because of the computational complexity. My overall approach provides a robust foundation for training the model and evaluating the explanations effectiveness in this diverse linguistic setup.

3.3 Model Development and Training

The research employs the XGboost model as the only classifier mostly due to its effectiveness to carry out the classification tasks and interpreting the explanations of explainable AI (XAI) techniques. Along with the XGboost model I have used TF-IDF vectorizer into my pipeline. TF-IDF helps to transform my text data into numerical features and captures term importances across my dataset. The data set is divided into my target column ('y'), which contains labels that indicate human and AI text, and the input features ('X'), which means input features and has text data. The division of data is into 80:20 split between training and testing sets. After dropping rows where features are less than 5 in my overall dataset of 20,000 rows the data becomes 18,114 rows with 14,491 (80) rows used for training and 3,623 (20) rows for testing. I created a pipeline which integrates the TF-IDF vectorizer with an XGboost classifier to streamline the training and prediction process. Evaluation metrics are also calculated to check model's performance and stability. This whole process ensures that there is consistency among the predictions and helps in the analysis of the validity of explanations generated by LIME (XAI).

3.4 Generating Explanations

The explanation generation stage in my study primarily focuses on producing feature-level explanations for model predictions using the XAI tool LIME (Local Interpretable Model-agnostic Explanations). The features mostly refers to the words or vocabulary in the text instance. For explanations - majority of my text analysis and explanation generation is vocabulary based testing. LIME was chosen over SHAP because LIME adequately serves the purpose of generating great interpretable explanations. In addition, LIME truly aligns with my research because of its effectiveness for text classifications. Here in this study both XAI techniques (LIME and SHAP) mostly yield the same output so LIME was sufficient enough for the task at hand. The explanation process demands to generate explanations for both Machine (AI) and Human generated samples of text , with importance on carefully identifying the distinguishing aspects or features. To test the validity of these explanations I have applied different perturbations such as synonym , paraphrasing , counterfactual modifications and noise injections on the data. The goal is to analyse that the patterns observed by these explanations are meaningful or relevant to model predictions and whether these explanations can be trusted to make a strategic business decision.

3.5 Evaluating the Validity of Explanations

I conducted the evaluation of explanation validity through systematic and comprehensive statistical metrics which are designed to assess faithfulness, robustness , sensitivity and fidelity. In faithfulness analysis , I have done sufficiency analysis which identifies the impacts of removing significant features on model probabilities , this ensures the decision making process of the model. To check the robustness of the explanations I carried out four different types of testing. Robustness is evaluated through sparsity , which calculates the minimal features to drive an explanation, Jaccard similarity to measure the difference between the instances after doing the perturbation or modifications in the data, Stability through the cosine similarity which checks the consistency across the variations among the instances , and lastly the Counterfactual analysis to measure the instance against its opposite theme settings. Sensitivity analysis checks small deviations of probabilities after minor changes in the input , I have complemented this with the KL-Divergence test which validates the distributional shifts in the labels caused by the modifications in text. Proxy fidelity compares the original model with surrogate model and check feature loyalty in identifying the same features between the models. These validity approaches were further enhanced when I applied them over multiple instances. My validity contains single instances and multiple instances (100 instances /1000 instances / Full dataset) analysis of each category to thoroughly identify the explanations trustworthiness in making decisions. This multifaceted approach of mine made certain that explanations are reliable or not in addressing critical dimensions of explanation validity.

3.6 Testing Against Different Use Cases

This study incorporates multilingual datasets to evaluate explanation validity of text classification. Focusing not just in English but different languages like Urdu , a non-Latin language, highlights the applicability of Explainable AI (XAI) tools and techniques across different regional and cultural contexts. To create a seamless approach I have used the same model XGboost and TF-IDF vectorizer in the new pipeline for the Urdu dataset. In practical scenarios English is not the only spoken language in the world which makes it more worthwhile to check the rigidness of the explanations. However , In context of Multilingual (Urdu) identifying AI vs Human text in local settings or context can be really instrumental. My research analyses and evaluates the semantic alignment and correlation in explanations between the two languages. Ensuring consistent , meaningful and relevant explanations across linguistic and true world scenarios demonstrates the practical usability of XAI for text classification.

3.7 Tools and Softwares Used

The research employs a range of tools and libraries for implementation:

1. Python: For data preprocessing, model training, and explanation generation.
2. Libraries: Scikit-learn, Sci-py ,TensorFlow, PyTorch, LIME, SHAP, and Hugging Face transformers for machine learning and XAI methods.
3. Visualization: Matplotlib, Seaborn, and Plotly for visualizing explanation results.

CHAPTER 4

Data Pre-Processing, Exploratory Analysis, Feature Engineering and Code Implementation

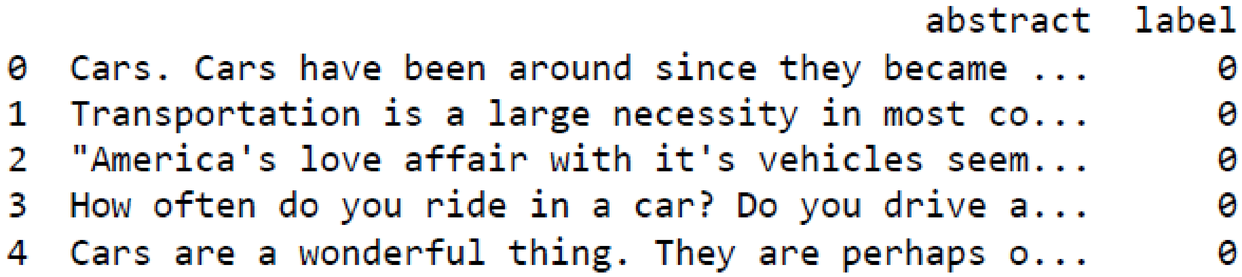
4.1 Data Source and Description

The data is from Kaggle.com and it was publicly available . This dataset includes both AI-generated (machine generated) and human generated texts , making it very suitable to generate explanations through Explainable AI (XAI) or explainable models. My dataset consists of features such as labels or targets (classification) , and text (including wide variety of linguistic and contextual features) for modelling. The ultimate goal is to utilize this dataset and perform a thorough analysis on the validity of explanations of XAI for text classifications. I will be using training data to develop a classification model , where the challenge lies is to generate explanations for model's predictions. The explanations of XAI will pe passed through certain validity checks like faithfulness , sensitivity , robustness and model fidelity.

4.2 Data Preprocessing

In data-preprocessing , my goal is to ensure that the raw data could be transformed and aligned to create a high performing classification model. Started on with loading the data from a CSV file I have renamed the columns which are easy to read and distinguish. After that I handled missing values or inconsistencies in the data , checked data types and converting the target column datatypes to integer since this is a binary classification.

Raw data contains two columns 'text' and 'generated' where 'text' represents my feature set and 'generated' represents the label (1 for AI and 0 for human). After renaming the columns and removing inconsistencies from the data I checked the data head or first few rows from the data to see the structure of the data. This thing helps ensure that modelling could be carried out without any errors. To use explanations of XAI I need the data to interpretable and consistent across every instance.



```
                                            abstract  label
0  Cars. Cars have been around since they became ...      0
1  Transportation is a large necessity in most co...      0
2  "America's love affair with it's vehicles seem...      0
3  How often do you ride in a car? Do you drive a...      0
4  Cars are a wonderful thing. They are perhaps o...      0
```

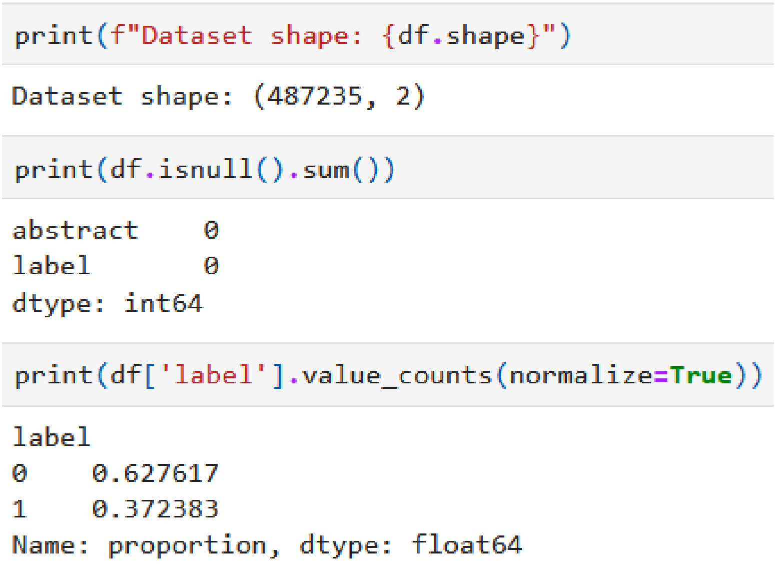
Figure 4.1: Data Overview

4.2.1 Loading, Transforming and Understanding The Dataset

The preprocessing step for me is fairly simple and it starts with loading the raw CSV file which is followed by an exploration of the data shape and any potential missing values. Finding these values are important to me since they could disturb model training process. Through this I find the basic structure of dataset , and size of the data (no of rows and columns in the data). The primary column in my dataset is the abstract (which is the text) and the other column is label (the target or classification for each piece of text).

By leveraging the pandas library I loaded the data as a DataFrame to be used for modelling. Checked for the nulls or missing instances in both the columns since this could negatively affect the model training process. However , I haven't found a single instance where there are any null values in both of my columns. In terms of completeness the data is ready for modeling.

The next step is to analyze the distribution of the classes, i.e., the label column, to identify if there is any class imbalance. Class imbalance occurs when one class significantly outnumbers the other, which can lead to biased predictions and potentially flawed XAI explanations. The dataset used in this analysis was found to be imbalanced, with 62.76% of the samples being human-generated (labeled 0) and 37.24% being machine-generated (labeled 1). This imbalance is an important consideration for model training, as most machine learning algorithms tend to be biased towards the majority class.



```
print(f"Dataset shape: {df.shape}")
```

```
Dataset shape: (487235, 2)
```

```
print(df.isnull().sum())
```

```
abstract    0
label       0
dtype: int64
```

```
print(df['label'].value_counts(normalize=True))
```

```
label
0    0.627617
1    0.372383
Name: proportion, dtype: float64
```

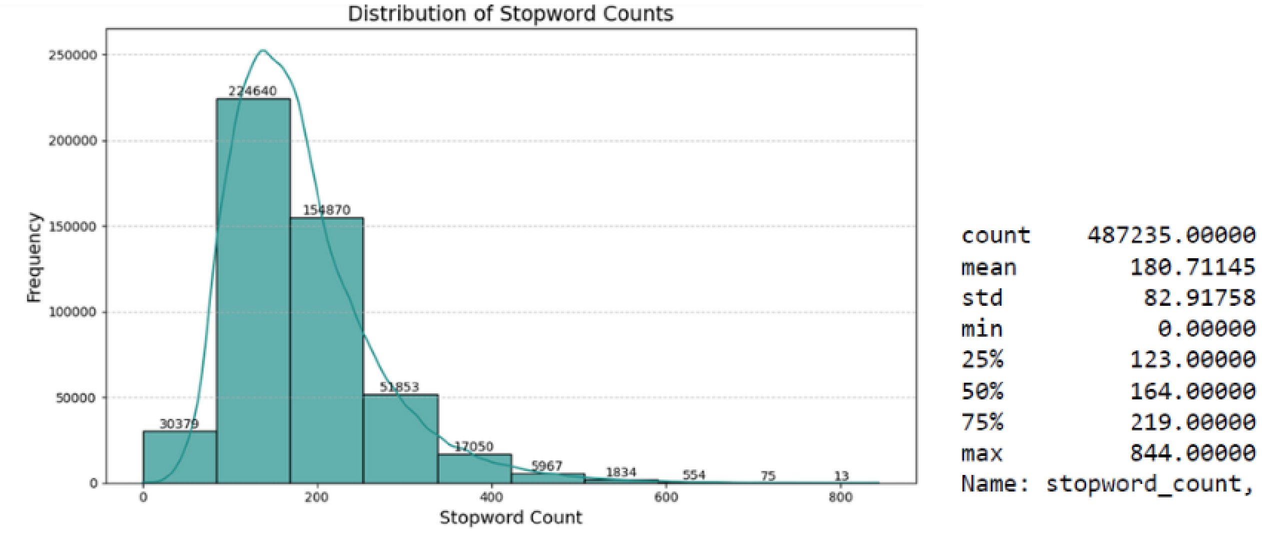
Figure 4.2: Data Description

4.2.2 Handling Stopwords

Stopwords are words that are common to any language and English is no exception. These words usually include 'is' , 'the', 'and' , 'of' , 'this' , 'a' etc. In text classification they are usually removed since they carry small or little semantic weights. These words can carry noise with them which changes the predictions of the model's in discussion. Words with only meaning are kept in the text and therefore I have removed stopwords from the data as well.

I am using the library known as Natural Language Toolkit (NLTK) , a very renowned python library , which identifies these stopwords and remove them from the text column of my data. After removing these stopwords each abstract becomes smaller which makes it more suitable for model training. Not that just makes it more suitable it also makes training the model more efficient and fast.

To gain insights into the dataset I am checking the stopword distribution in the text data. I find that most abstracts contain less than 200 stopwords in each instance. On average there are 180-200 stopwords in each abstract. There are also instances where there are around 800 stopwords in a single abstract making them less meaningful. The number of filler portions vary across the dataset making a point that some abstracts are more significant than others. More concise abstract creates better feature sets for the model in the longer run.



```
count    487235.00000
mean        180.71145
std          82.91758
min           0.00000
25%         123.00000
50%         164.00000
75%         219.00000
max         844.00000
Name: stopword_count,
```

Figure 4.3: Stopwords Distribution

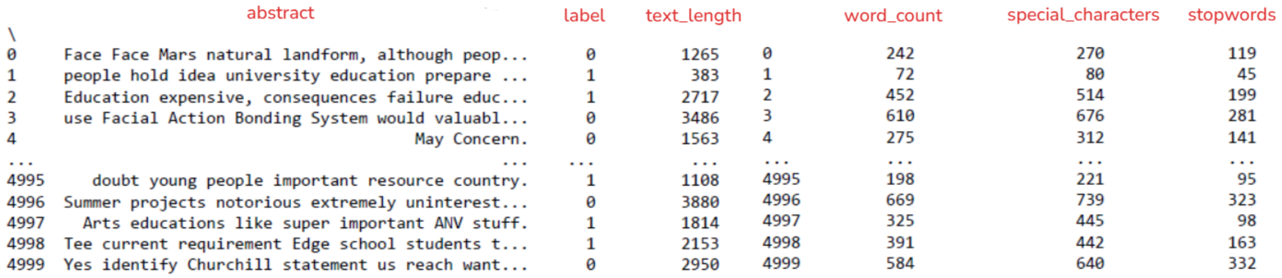
4.2.3 Text Tokenization and Data Transformation

Text tokenization in text classifications is really effective as it transforms the data to be analysed in a proper way. I have done text tokenization by splitting my data (text) into smaller chunks or units , such as sentences or words , which enabled my machine learning algorithm to learn effectively. In my research , tokenization was performed at the sentence level where I have retained the first sentence of each abstract and it is because the first sentence usually encapsulates the core idea of any text or passage. This is sufficient enough for my classification tasks.

After tokenization I have removed the special characters from the data such as punctuation marks and digits which basically do not contribute to any contextual meaning. These special characters just add noise in the data. I have ensured that only the relevant context of the text is kept and the rest is discarded.

I also addressed the problem of the dataset's class imbalance , which consisted of 62.76% human-generated (labeled 0) and 37.24% machine-generated (labeled 1) samples out of around 400K samples. Careful down-sampling was performed to balance the dataset by selecting the random sample of 10,000 rows from each class label, though the random state was kept static to ensure reproducibility. One more thing was also crucial where I removed rows where the text contains less than 5 words in each abstract or instance. I did this because I have trained my LIME to generate 5 features at least in each instance. With this I made sure to keep great amount of feature set to validate.

These steps created a processed pipeline for the raw data to be used effectively for model creation.



```
                            abstract                    label  text_length          word_count        special_characters  stopwords
\
0     Face Face Mars natural landform, although peop...   0       1265    0            242                 270              119
1     people hold idea university education prepare ...   1        383    1             72                  80               45
2     Education expensive, consequences failure educ...   1       2717    2            452                 514              199
3     use Facial Action Bonding System would valuabl...   0       3486    3            610                 676              281
4                                          May Concern.   0       1563    4            275                 312              141
...                                                 ...  ...       ...  ...            ...                 ...              ...
4995     doubt young people important resource country.   1       1108  4995            198                 221               95
4996  Summer projects notorious extremely uninterest...   0       3880  4996            669                 739              323
4997    Arts educations like super important ANV stuff.   1       1814  4997            325                 445               98
4998  Tee current requirement Edge school students t...   1       2153  4998            391                 442              163
4999  Yes identify Churchill statement us reach want...   0       2950  4999            584                 640              332
```

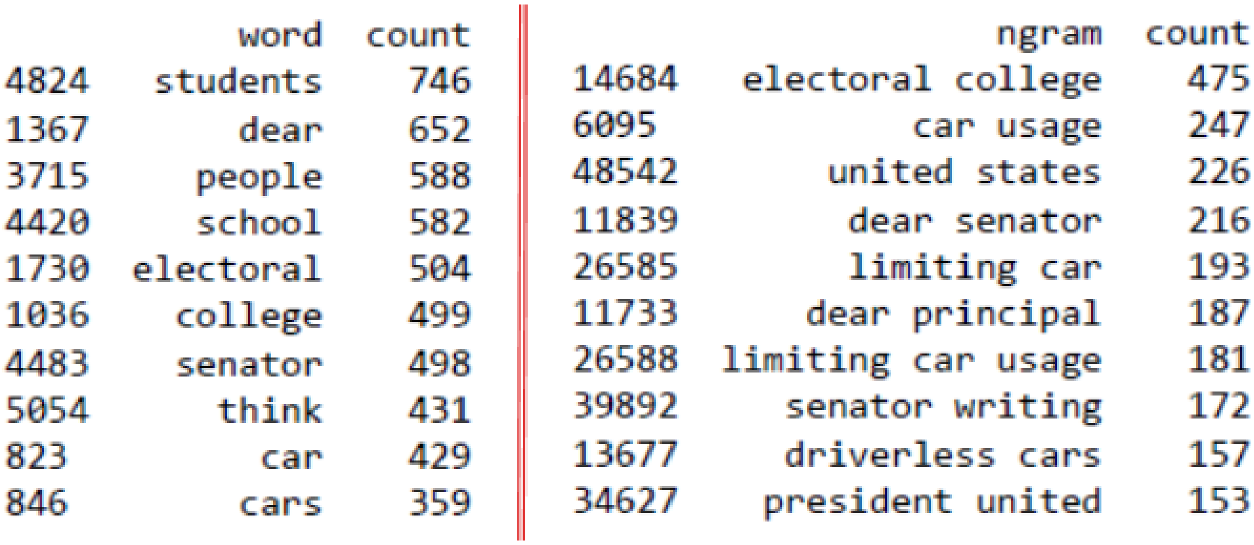
Figure 4.4: Tokenization and Transformation

4.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in every model development since it helps in understanding the overall structure , the underlying patterns , and distributions of the data. It is important to identify certain data trends because model selection and evaluation depends on it. Following subsections explored the Initial Data Analysis conducted which highlights different aspects of text dataset and to improve the model interpretability and performance.

4.3.1 Most Common Words and N-grams

N-gram and common word section explores the distribution of language instances within the dataset. The idea is to examine the proportions of text in different languages or any specific language specific patterns. Analysis of language detection methods helps to identify which features might influence the feature identification and model selection process. Apart from this, it also gives me insight on the challenges that are associated with multilingual or just any text classification. The N-gram (univariate , bi-gram and tri-gram) in my case uncovers key contextual dependencies and important phrases that enhance my model's performance. Ultimately increasing my model's ability to differentiate between human and machine generated text. With this analysis , I found that most common word in my data is student with around 746 instances. Highest used bi-gram and tri-gram are 'electoral college' with 475 instances and 'limiting car usage' with 181 instances respectively.



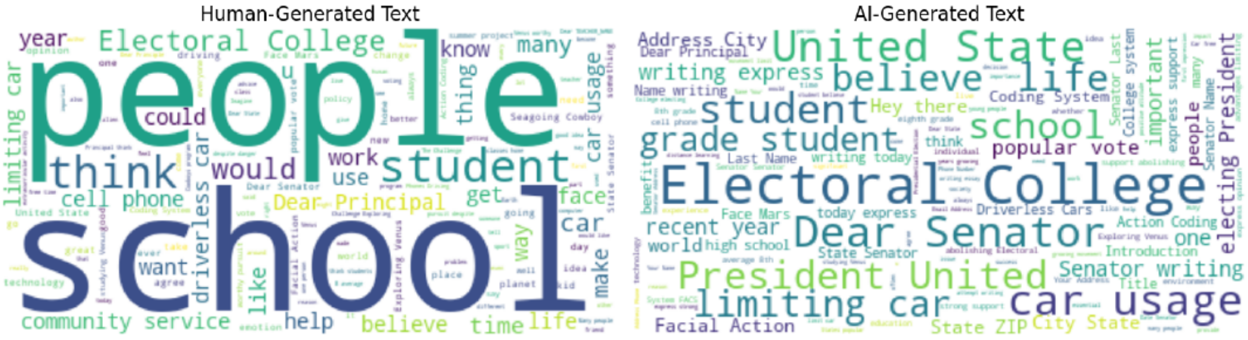
```
           word  count
4824   students    746
1367       dear    652
3715     people    588
4420     school    582
1730  electoral    504
1036    college    499
4483    senator    498
5054      think    431
823         car    429
846        cars    359
```

```
                     ngram  count
14684     electoral college    475
6095              car usage    247
48542         united states    226
11839          dear senator    216
26585          limiting car    193
11733        dear principal    187
26588    limiting car usage    181
39892        senator writing    172
13677       driverless cars    157
34627      president united    153
```

Figure 4.5: N-gram Analysis

4.3.2 Class Distribution Analysis

Word cloud visualizations are used to present the most frequent words in the dataset. In this subsection, we generate word clouds for the entire dataset as well as for each class (human-generated and machine-generated text). The general word cloud provides a high-level overview of word frequency, while the class-specific word clouds help compare the lexical distribution between the two categories. This visualization technique aids in identifying domain-specific words and patterns that are more prevalent in one class compared to the other.



Human-Generated Text

AI-Generated Text

Figure 4.6: Word Cloud - Human vs AI

Vocabulary diversity visualization highlights the patterns specific to the target column just like word cloud but it also gives the contrast between the showing the vocabulary diversity between each class. It can also be seen from the word clouds above class 0 (human generated) have more words related to academics such as students , college etc .On the other hand class 1 (AI generated) have more words related to news and social media like vote , elections etc. This pattern can support the claim if one class is biased towards few words. So managing those words can improve model performance. The vocabulary difference between the two class is 0.0194. The vocabulary difference is relatively minor or negligible so there is no need to remove any specific words from the dataset.

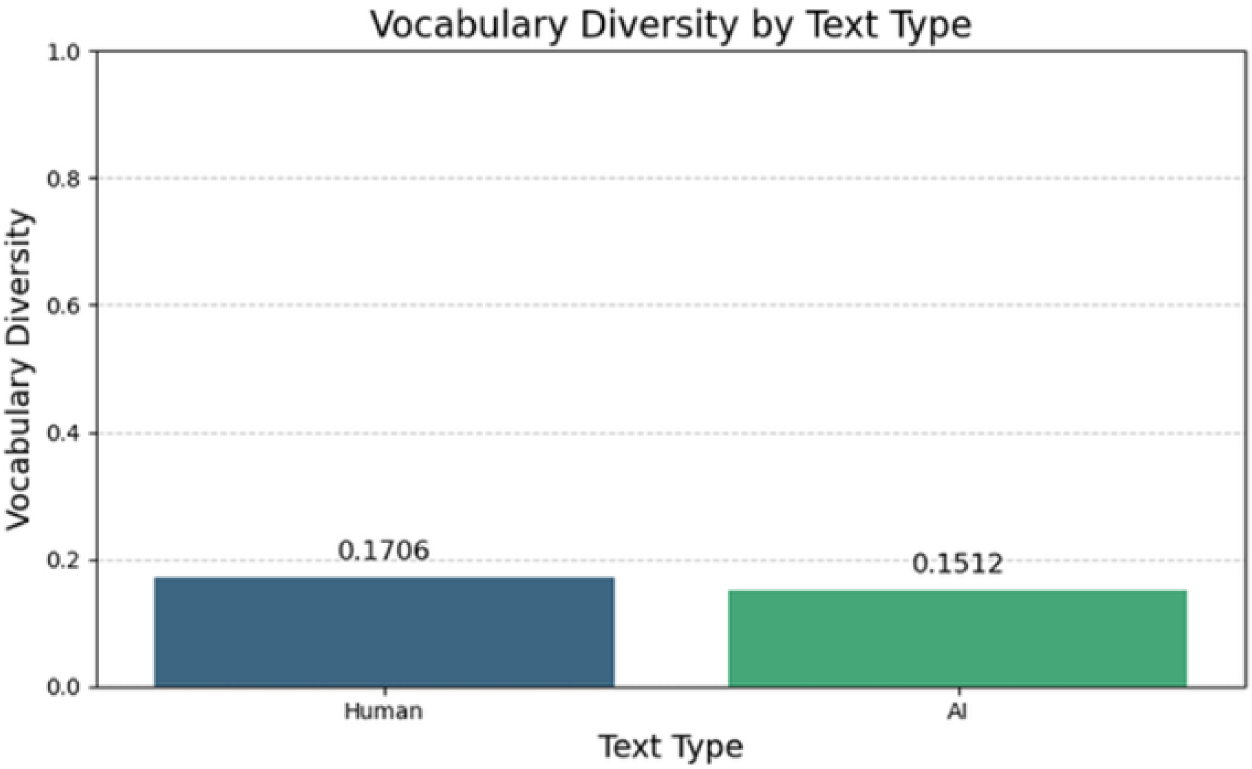


Figure 4.7: Vocabulary Diversity - Human vs AI

This is an extension of word cloud and vocabulary diversity where we perform a detailed analysis of word frequency analysis to identify top 20 words from each class. The top 3 words in Human abstracts are ′students′, ′think′ and ′dear′ while AI has words ′dear′, ′writing′ and ′electoral′. This supports the claim that the diversity between each class is pretty low. Analysis like these are crucial in the feature engineering phase where selecting any relevant words can help improve the vectorization techniques.

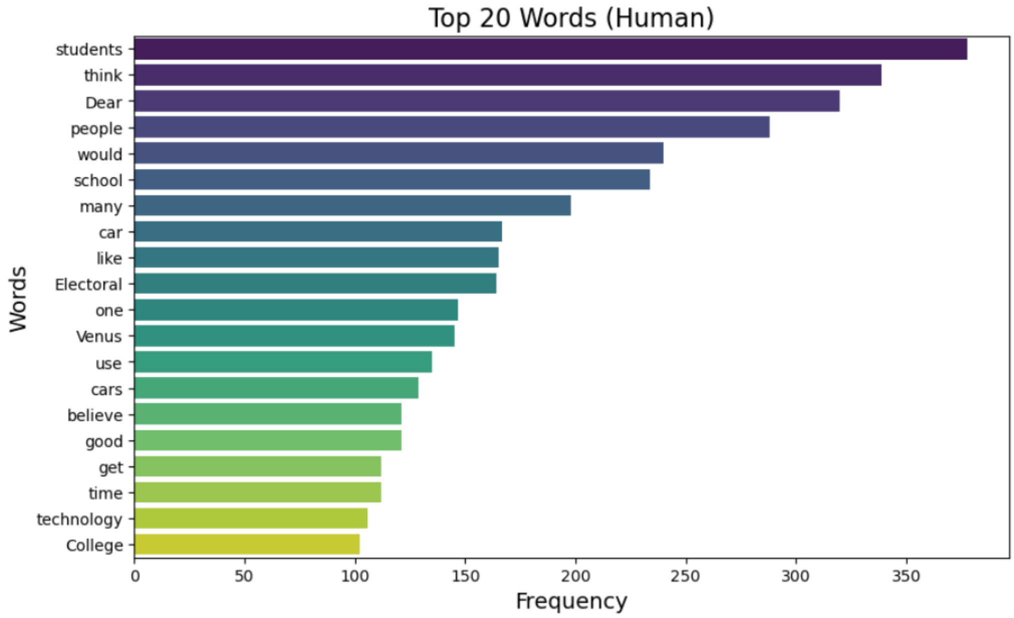


Figure 4.8: Most frequent words - Human

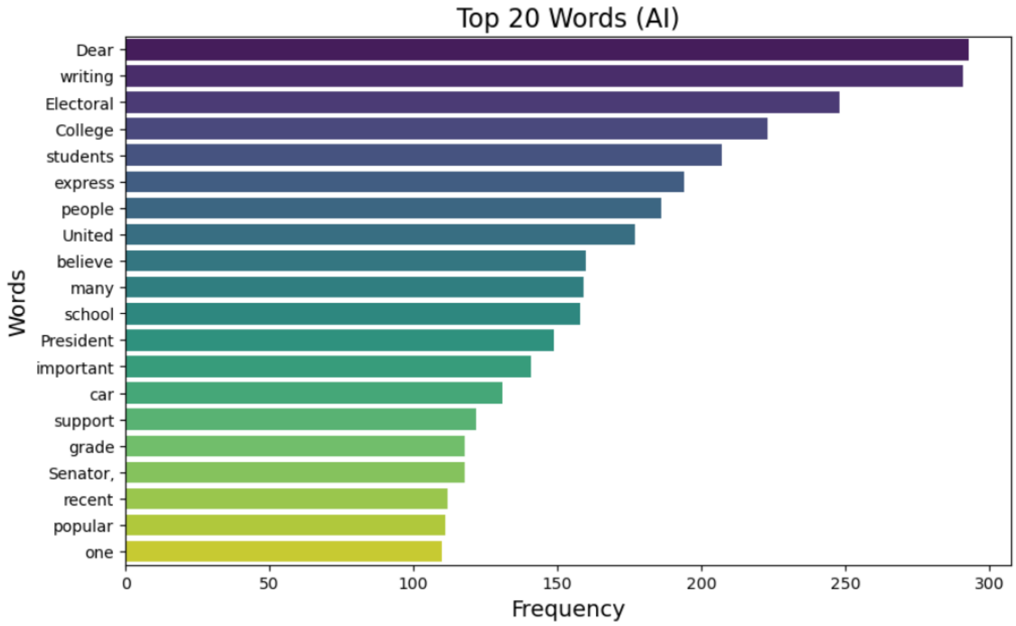


Figure 4.9: Most frequent words - AI

4.3.3 Sentiment Analysis

Polarity: I have done sentiment analysis to assess the emotions conveyed by the text. The sentiment polarity overall explains that each the general sentiment in my text is positive, negative or neutral. It gives insight on the emotional tone of the dataset and the potential impacts of the text on the classification model. I have compared the polarity scores between

AI (machine generated) vs human generated text to check if these sentiment play any role to create a difference between two classes. From **Figure 4.10** it can be said that the mean polarity remain somewhat close to neutral and slight positive in each class.

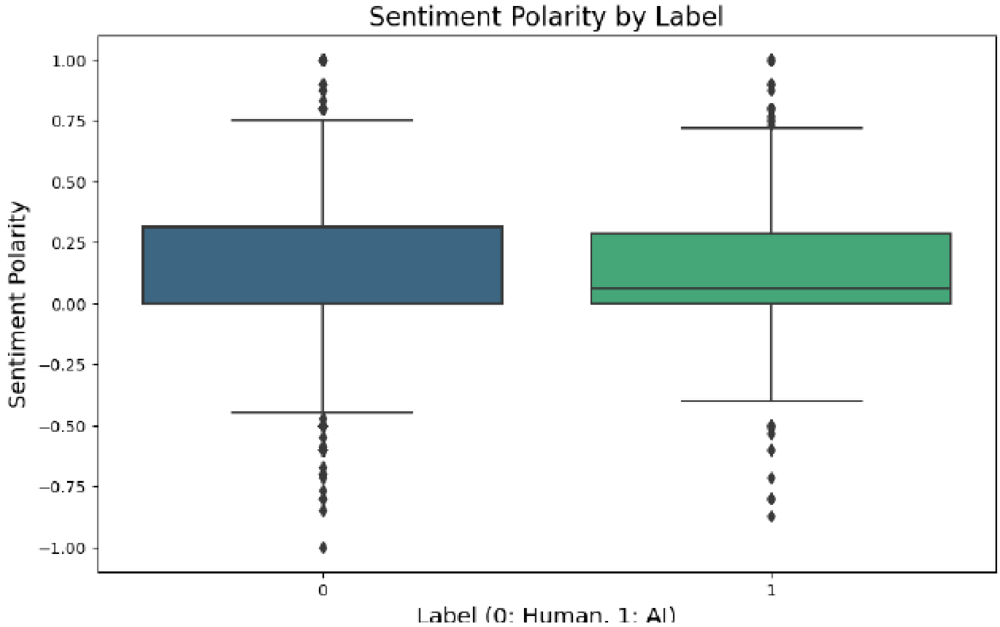


Figure 4.10: Sentiment Polarity - Human vs AI

Subjectivity: With polarity , I have used subjectivity measures as well. Subjectivity expresses that text contains objective facts or subjective opinions. I have compared the subjective scores of each class AI(machine generated) and human generated text. This helps me to understand that if the subjective language or the objective language can shift the model's decision from AI to human classification. Both classes contain a subjective score of around 0.4. This means that around **40%** of the text is subjective in each class.

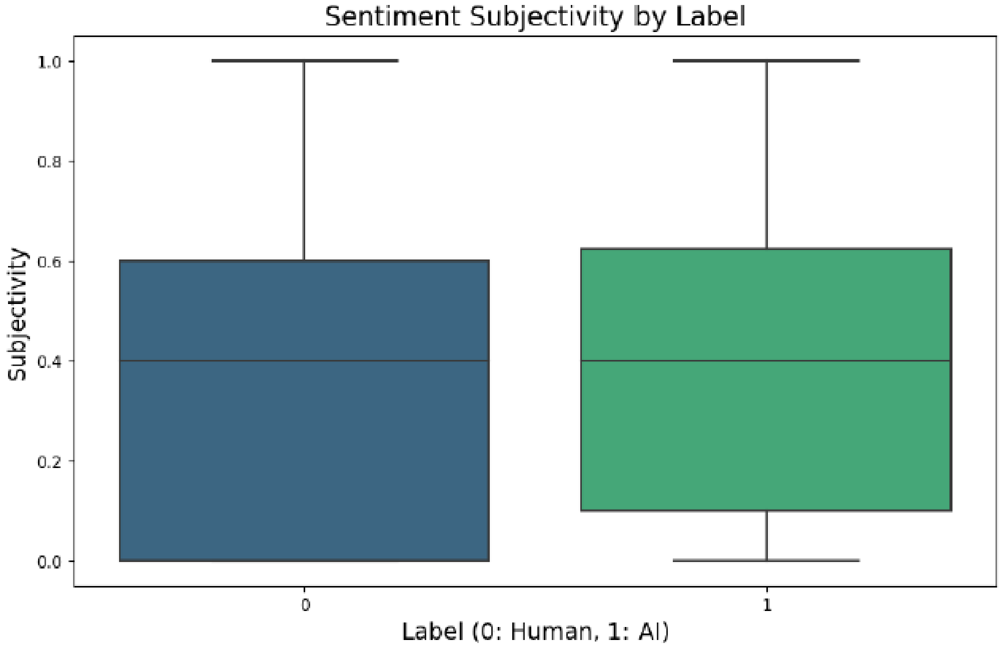


Figure 4.11: Sentiment Subjectivity - Human vs AI

4.3.4 Similarity and Correlation Analysis

This section explores the similarity between each text instance using various techniques like cosine similarity and word embeddings. Here I am trying to find out how similar or dissimilar the text samples are throughout the dataset. I have computed the cosine similarity between first 10 samples and generated a heatmap to show the results. The scores between almost every sample is 0.0 excluding 1 or 2 samples which have scores of 0.32.Even though these are all abstracts containing similar themes but they differ each other when calculating the similarity scores. This shows that how important it is to identify the patterns in text classification.

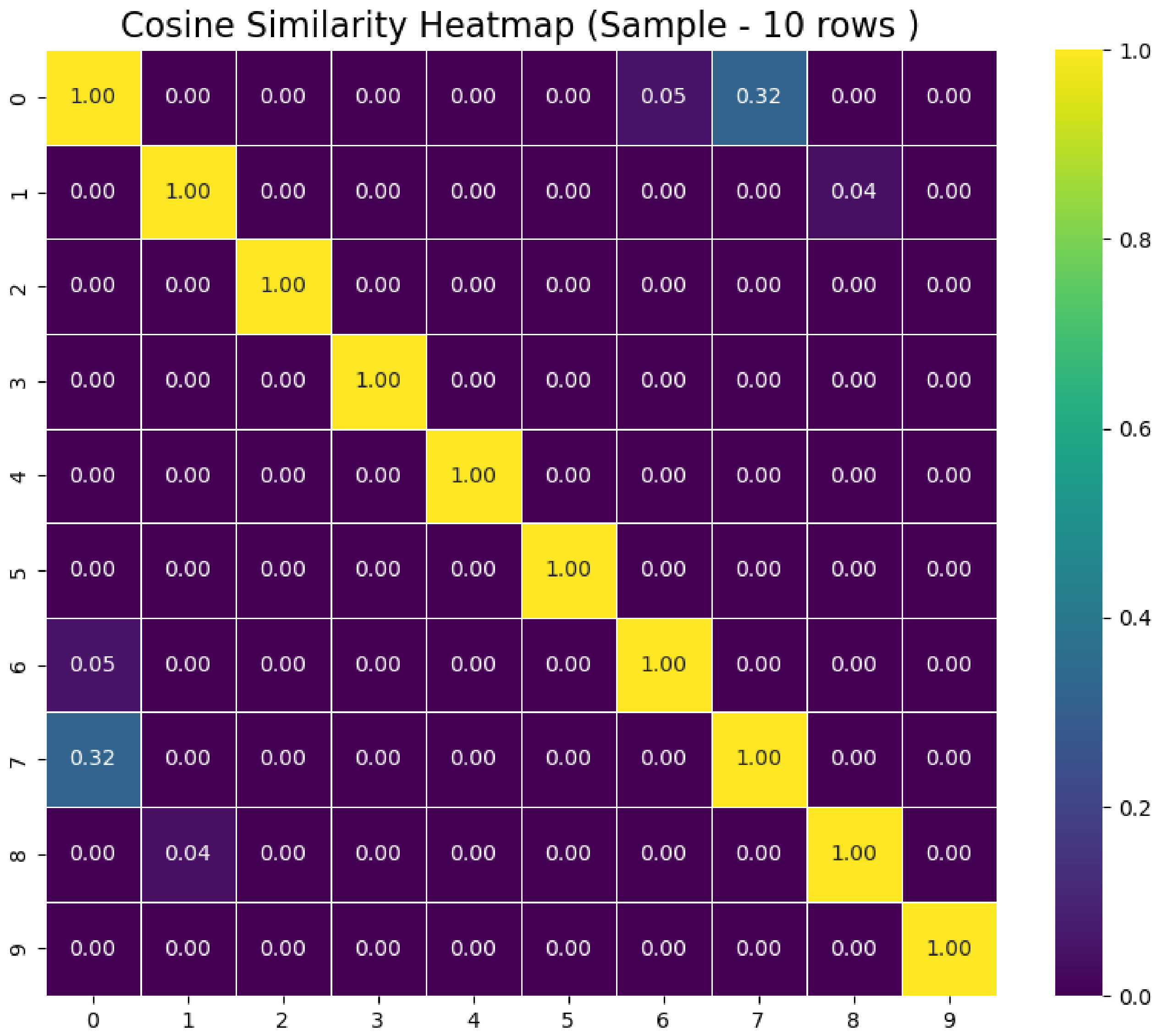


Figure 4.12: Cosine Similarity - Top 10 instances

Feature correlation analysis examines the relationships of each feature in the dataset. I have computed my feature correlation in a different way to analyse the patterns in the data. I computed readability ,subjectivity , sentiment , wordcount and text length and find correlation between each of these values. I find that subjectivity of the text is more correlated to the sentiment with a value of 0.43 but word count and text length has the most positive correlation with a value of 0.97.While word count and readability has the most negative correlation with a value of -0.37

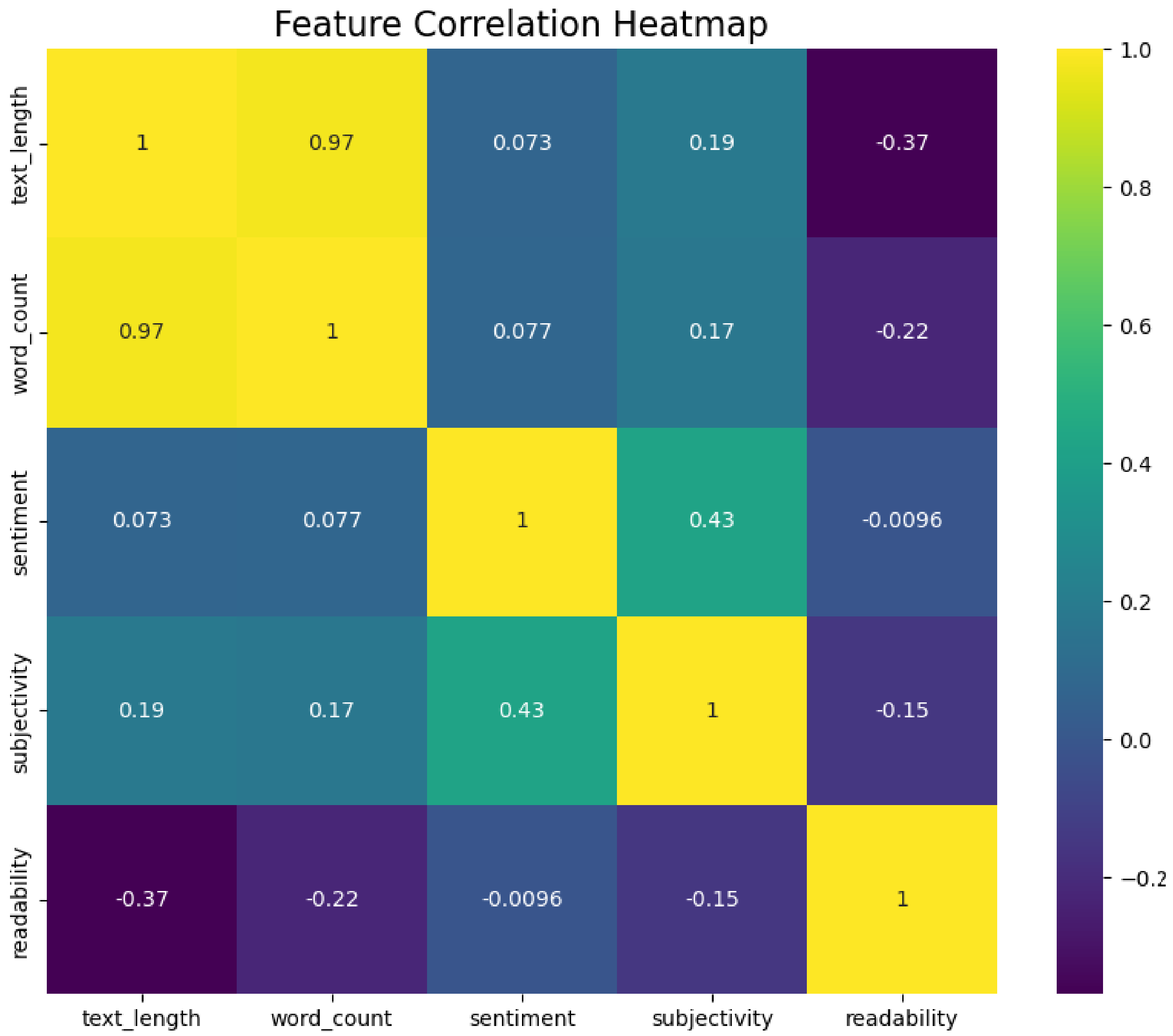


Figure 4.13: Feature Correlation

4.3.5 Embeddings

Finally , embeddings convert my text data into dense vectors which captures semantic relationships between words. I captured linguistic nuances between classes to differentiate AI and human generated texts. It's a 2D visualization to see the text data in a cluster space. By looking at the visualization we can see most of the text instances overlap each other showing same pattern in the data overall.

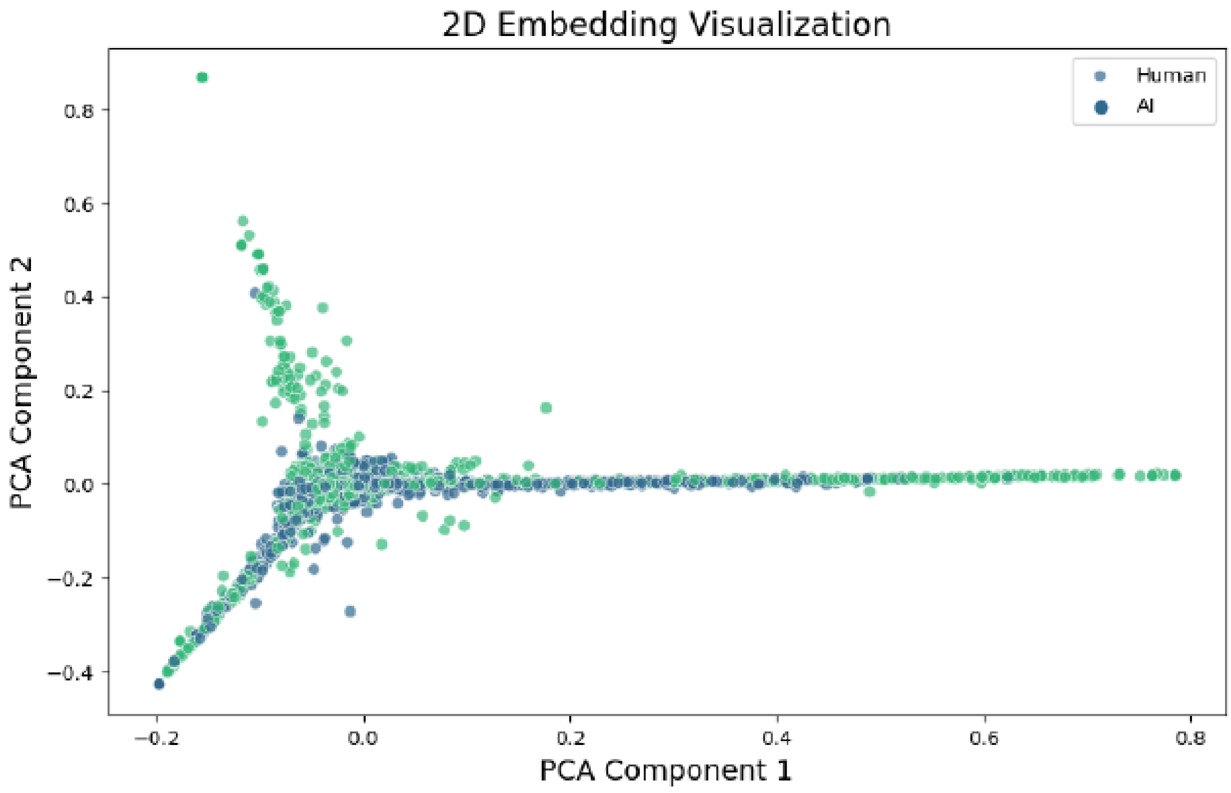


Figure 4.14: Embeddings - Human vs AI

CHAPTER 5

Model Pipelines , Data Splitting and XAI Techniques

5.0.1 Data Splitting and Preprocessing

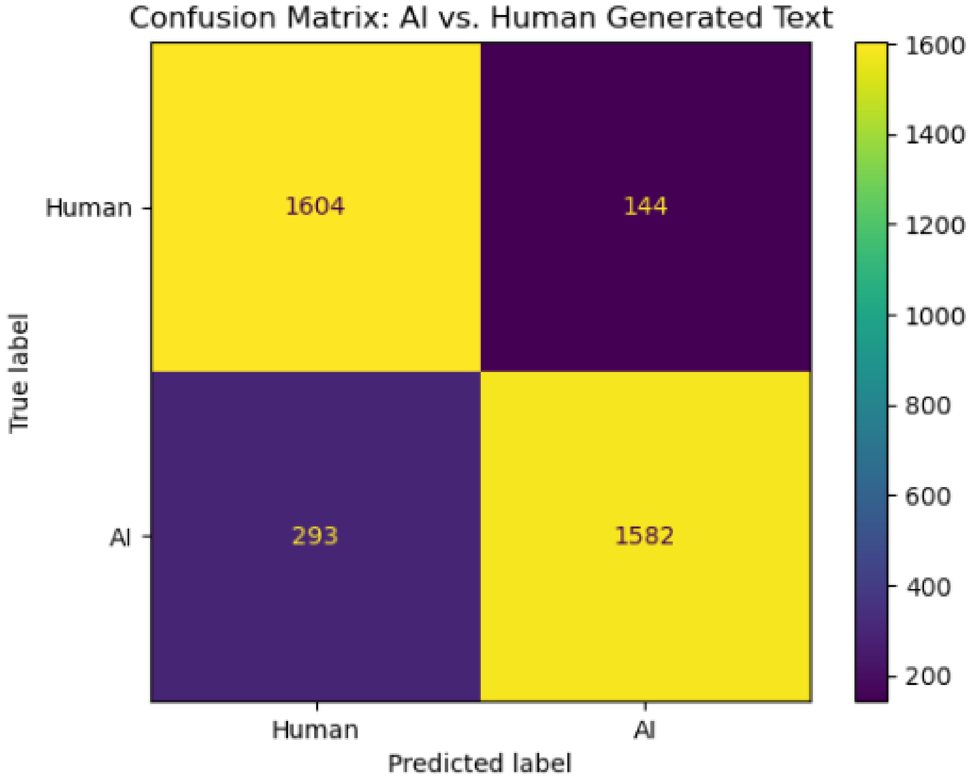
I have taken a subset of **20,000** rows from **487,235** instances. To resolve the class imbalance problem these 20,000 rows contain 10,000 samples from each class. Further I have removed rows which have less than 5 features in the text column ,this drops the rows to **18,114** rows. This dataset was passed to the model. The dataset was split into training (data used by the model to learn patterns) and testing (unseen data which predicts and evaluate model performance). The split between training and testing is **80/20** meaning 80% (14,491) for training and 20% (3,623) for testing. The approach is to train a model on a sufficiently large portion of the data.

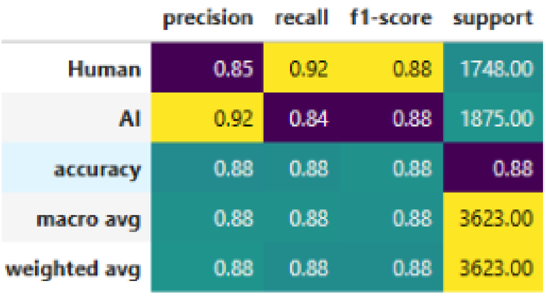
- Training Set: **14,491** instances (80% of the dataset)
- Testing Set: **3,623** instances (20% of the dataset)

The approach allows the model to learn the patterns effectively. The testing part is not used in training model and serves to evaluate model at the end of training.

5.0.2 XGBoost Model with TF-IDF Vectorization

The dataset with 18,114 rows was used to train the XGBoost model while leveraging the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique. TF-IDF is a widely used technique for text feature extraction that represents text data in a numerical form, where each word's importance is weighted according to its frequency in the document and the inverse frequency across all documents. This helps in identifying important words in each class i-e 0 (Human) and 1 (AI) . The XGBoost classifier, a powerful gradient boosting machine model, was then trained using the TF-IDF transformed features. This model was chosen due to its robustness and ability to handle complex relationships in the data effectively.





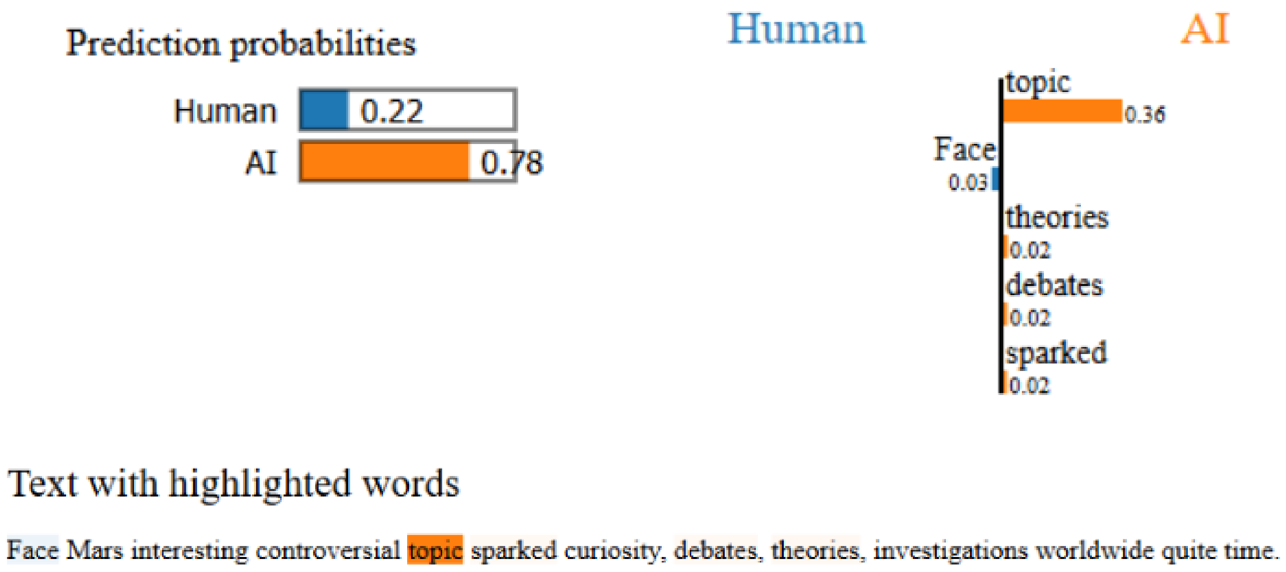
	precision	recall	f1-score	support
Human	0.85	0.92	0.88	1748.00
AI	0.92	0.84	0.88	1875.00
accuracy	0.88	0.88	0.88	0.88
macro avg	0.88	0.88	0.88	3623.00
weighted avg	0.88	0.88	0.88	3623.00

Figure 5.1: Classification Report and Confusion Matrix

The classification model achieves an overall accuracy of 87%, effectively distinguishing between human-generated and AI-generated text. The confusion matrix shows that the model correctly classifies 891 human and 841 AI-generated texts while misclassifying 122 AI as human and 138 human as AI. The precision (86.6% for Human, 87.4% for AI) indicates the model is slightly better at avoiding false positives for AI text, while the recall (88.0% for Human, 86.0% for AI) shows it captures most instances of Human text but misses a few AI texts. The F1-scores (87.3% for Human, 86.7% for AI) reflect balanced performance across classes, with macro and weighted averages aligning closely at 87%, highlighting no significant class imbalance. While the results demonstrate strong performance, refining features or preprocessing techniques could further reduce misclassifications and improve differentiation between human and AI text.

5.0.3 LIME: Local Interpretable Model-agnostic Explanations

To interpret the predictions made by the XGBoost model, we utilized LIME (Local Interpretable Model-agnostic Explanations), a popular XAI tool for explaining individual predictions. LIME works by approximating the complex model with a simpler, interpretable surrogate model for each prediction, allowing us to understand which features were most influential in the model's decision-making. For instance, by selecting a sample text from the test set, LIME identified the most impactful words contributing to the classification decision. The beauty of LIME explanations is because they give the innermost behaviour of model decision making. Making the transaction on even the instances more interpretable. The instance below also shows that how lime show explanations. It highlights the words that are most impactful ton contribute to each class and also the prediction probabilities of Human and AI.



Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.

Figure 5.2: LIME Results - Generic

5.0.4 SHAP: SHapley Additive exPlanations

In addition to LIME , SHAP (Shapley Additive explanations) is also used to validate the general view of the model's feature importance. Though both XAI tools are closely related but SHAP measure the contribution of each feature to the model's predictions overall. In text classification The features were calculated after transforming the data from the TF-IDF vectorizer. A SHAP summary plot was then generated, which visualized the importance of each feature across all test instances. This plot helped to identify which words or terms were consistently influential in determining whether a text was human-generated or machine-generated. SHAP offered valuable insights into the model's global decision-making process. In text classification , LIME is the most widely used tool but overall SHAP is most popular XAI tool for explanations.

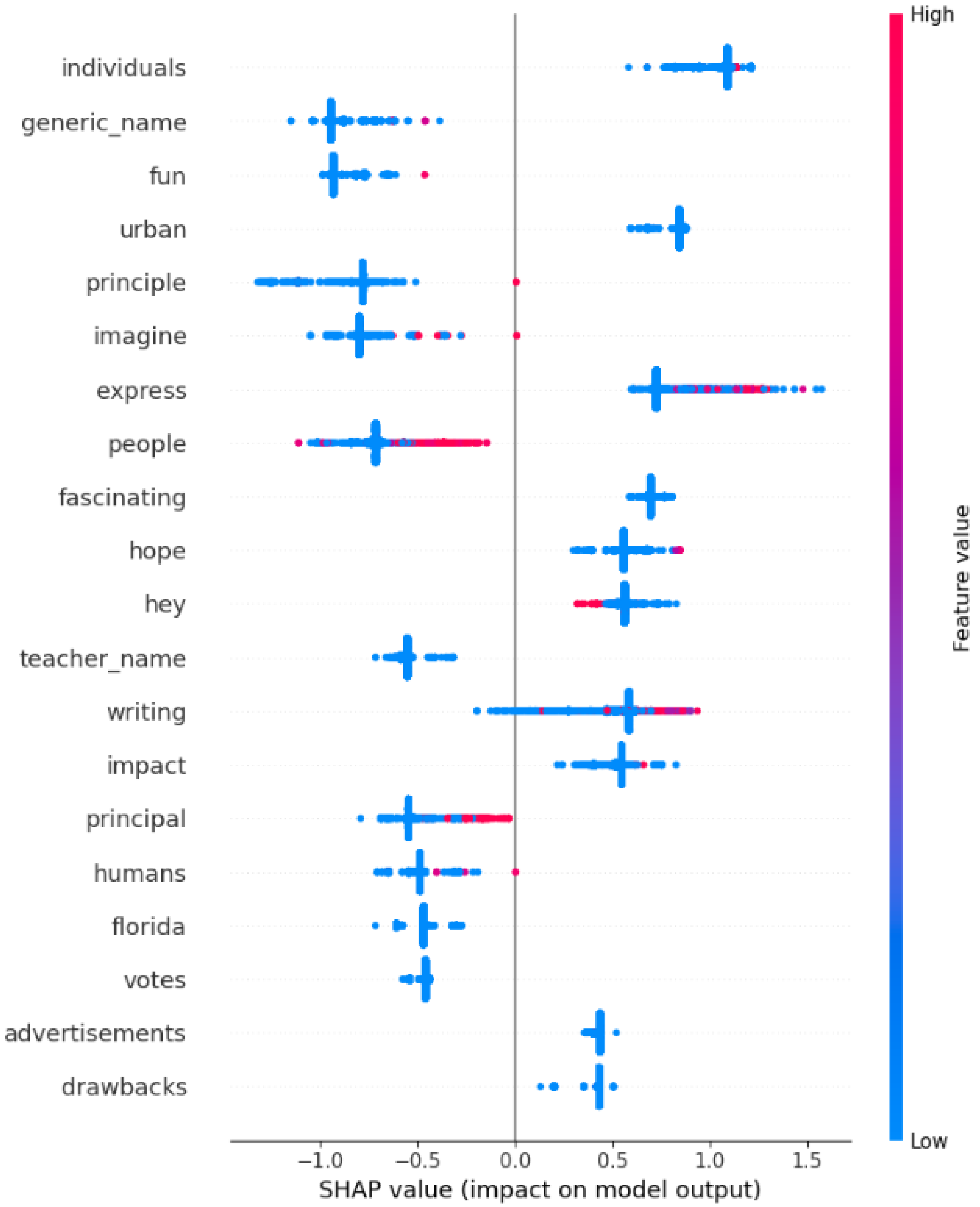


Figure 5.3: SHAP Results - Generic

CHAPTER 6

Evaluation of Validity of Explanations of Explainable AI (XAI)

6.1 Faithfulness Testing: Sufficiency Analysis

Faithfulness in XAI (Explainable AI) ensures that explanations accurately reflect the model's decision-making process. Sufficiency Analysis evaluates the extent to which the identified important features alone can replicate the model's original prediction. A high sufficiency score suggests that the retained features capture the essential decision-making elements, while a low score indicates missing critical features.

The sufficiency score is calculated as:

$$S = \frac{\text{Prediction using Retained Features}}{\text{Original prediction}}$$

Where:

- S: Sufficiency score
- Prediction using Retained Features: The model's output when using only the features identified as most impactful by the explanation method.
- Original prediction: The model's output using all features.

6.1.1 Basic Example of Sufficiency

Suppose the sentence is:

"AI is transforming industries with innovation and efficiency."

1. The model predicts a **0.92 probability** that the text is machine-generated.
2. The explanation method identifies the most important features:
    - Retained Features: "AI", "transforming", "innovation"
3. Using only the retained features, the model's prediction drops to **0.67**.

Using the formula:

$$S = \frac{\text{Prediction using retained features}}{\text{Original prediction}} = \frac{0.67}{0.92} \approx 0.7283$$

The sufficiency score of **0.7283** implies that the retained features sufficiently explain about **72.83%** of the original prediction, indicating that these words capture the core elements influencing the model's decision.

6.1.2 Measuring Sufficiency Scores on Different Data Subsets

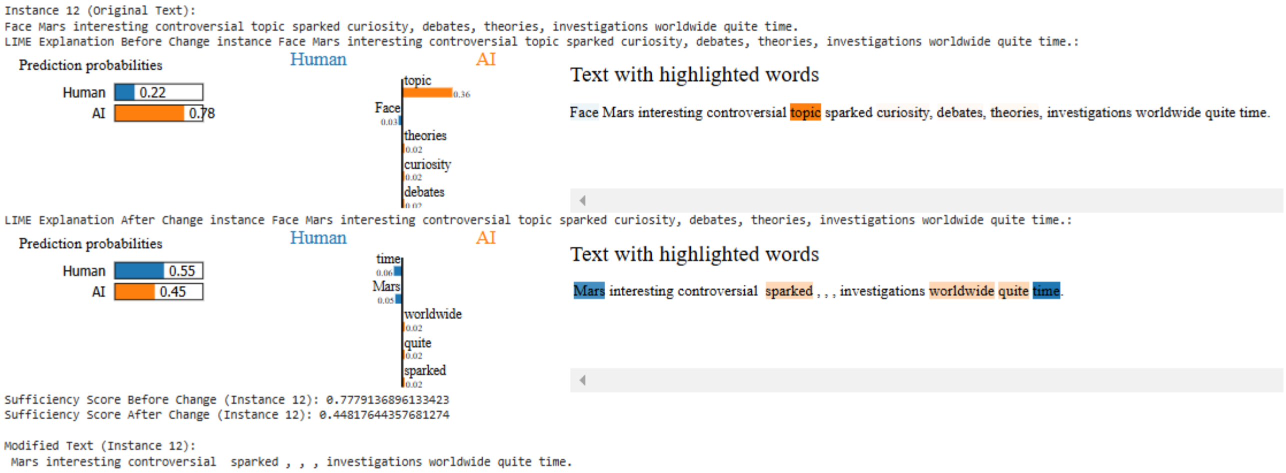
Below are summary statistics for sufficiency analysis conducted across different subsets of the dataset:

Subset	**Metric**	**Mean**	**Standard Deviation**	**Min**	**Max**	**AI / Human**
100 Instances	Sufficiency	0.233	0.288	0.000	1.370	AI: 0.196 / Human: 0.262
1000 Instances	Sufficiency	0.204	0.269	0.000	1.370	AI: 0.214 / Human: 0.193
Full Dataset	Sufficiency	0.191	0.257	0.000	1.548	AI: 0.210 / Human: 0.171

Table 6.1: Sufficiency Metric for Different Subsets of the Dataset

As the dataset size increases, the mean sufficiency score decreases slightly, from 0.233 for 100 instances to 0.191 for the full dataset. The standard deviation remains consistent, indicating stable variability in the sufficiency scores. Additionally, the AI label consistently shows lower sufficiency scores compared to the Human label across all subsets.

Below are summary statistics for sufficiency analysis conducted across individual instances of the dataset:



```
Instance 12 (Original Text):
Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.
LIME Explanation Before Change instance Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.:
```

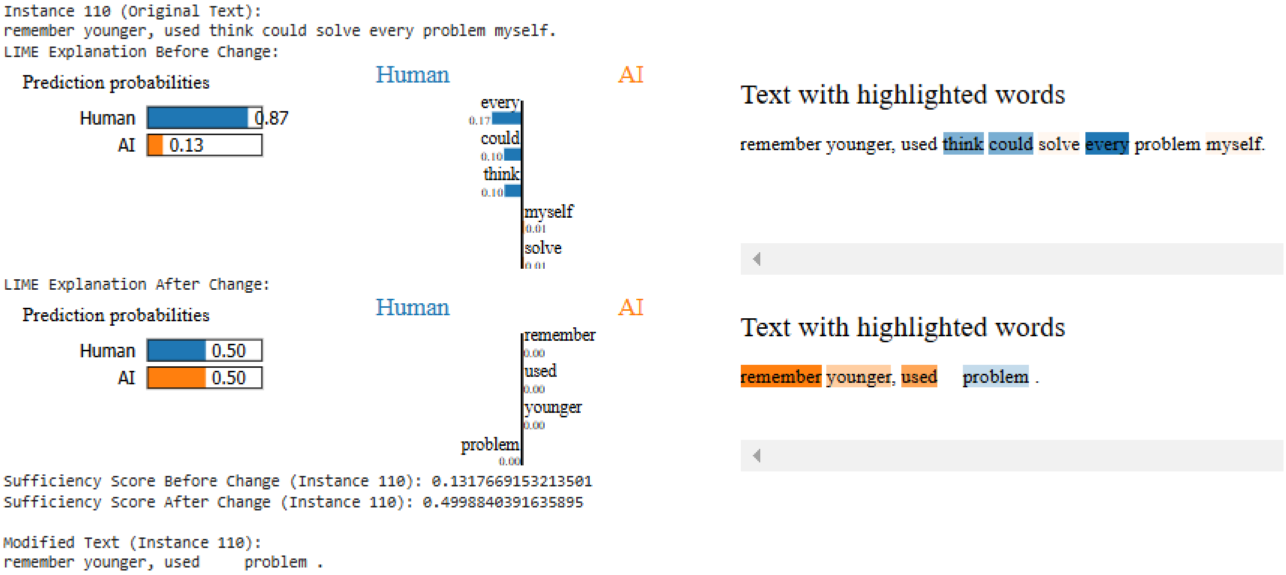
Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.

```
LIME Explanation After Change instance Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.:
```

Mars interesting controversial sparked , , , investigations worldwide quite time.

```
Sufficiency Score Before Change (Instance 12): 0.7779136896133423
Sufficiency Score After Change (Instance 12): 0.44817644357681274

Modified Text (Instance 12):
 Mars interesting controversial  sparked , , , investigations worldwide quite time.
```

Figure 6.1: Sufficiency Results - Instance No:12



```
Instance 110 (Original Text):
remember younger, used think could solve every problem myself.
LIME Explanation Before Change:
```

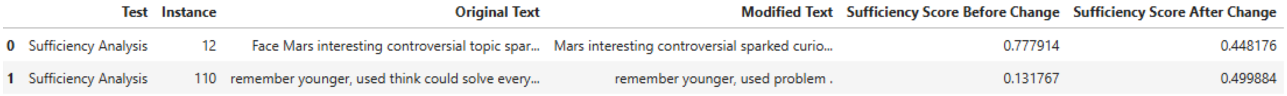
remember younger, used think could solve every problem myself.

```
LIME Explanation After Change:
```

remember younger, used   problem .

```
Sufficiency Score Before Change (Instance 110): 0.1317669153213501
Sufficiency Score After Change (Instance 110): 0.4998840391635895

Modified Text (Instance 110):
remember younger, used     problem .
```

Figure 6.2: Sufficiency Results - Instance No:110



	Test	Instance	Original Text	Modified Text	Sufficiency Score Before Change	Sufficiency Score After Change
0	Sufficiency Analysis	12	Face Mars interesting controversial topic spar...	Mars interesting controversial sparked curio...	0.777914	0.448176
1	Sufficiency Analysis	110	remember younger, used think could solve every...	remember younger, used problem .	0.131767	0.499884

Figure 6.3: Sufficiency Results - Table

6.1.3 Accessing Validity Through Sufficiency Scores

The overall sufficiency scores are relatively low, with means ranging from 0.191 to 0.233. This indicates that, for most instances, the retained features only partially replicate the model's prediction. As the dataset grows, the model's explanations become more efficient, using fewer but more informative features to account for predictions.

While some individual cases, such as Instance 12, exhibit moderate sufficiency, others, like Instance 110, reveal significant gaps, indicating room for improvement in the explanation methods.

The validity of the explanations is partially supported, as sufficiency scores demonstrate reasonable faithfulness for some instances. However, cases with low sufficiency scores highlight the need for further refinement in explanation techniques to ensure they capture the full complexity of the model's decision-making process.

6.2 Robustness Testing: Sparsity Analysis

Robustness in XAI refers to the consistency of explanations when the input data is perturbed or modified. Sparsity Analysis specifically evaluates how much the explanation relies on a small number of features to generate the model's prediction. A sparsity score close to 1 indicates that the explanation is based on a small set of features, while a lower score suggests that the explanation is more dependent on a larger set of features.

The sparsity score is calculated as:

$$S = \frac{\text{Number of retained features}}{\text{Total number of features}}$$

where:

- S: Sparsity score
- **Number of retained features**: The number of features used in the explanation after filtering.
- **Total number of features**: The total number of features in the dataset for that instance.

6.2.1 Basic Example of Sparsity

Let us consider a simple example to understand sparsity:

- **Original Text**: *"AI is transforming industries and driving innovation worldwide."*
- **Retained Features**: {AI, transforming, innovation}

Here, the total number of features in the text is 8, and the number of retained features is 3.

$$S = \frac{\text{Number of retained features}}{\text{Total number of features}} = \frac{3}{8} = 0.375$$

This score shows that only a small fraction of the features are retained, indicating sparsity in the explanation.

6.2.2 Measuring Sparsity on Different Data Subsets

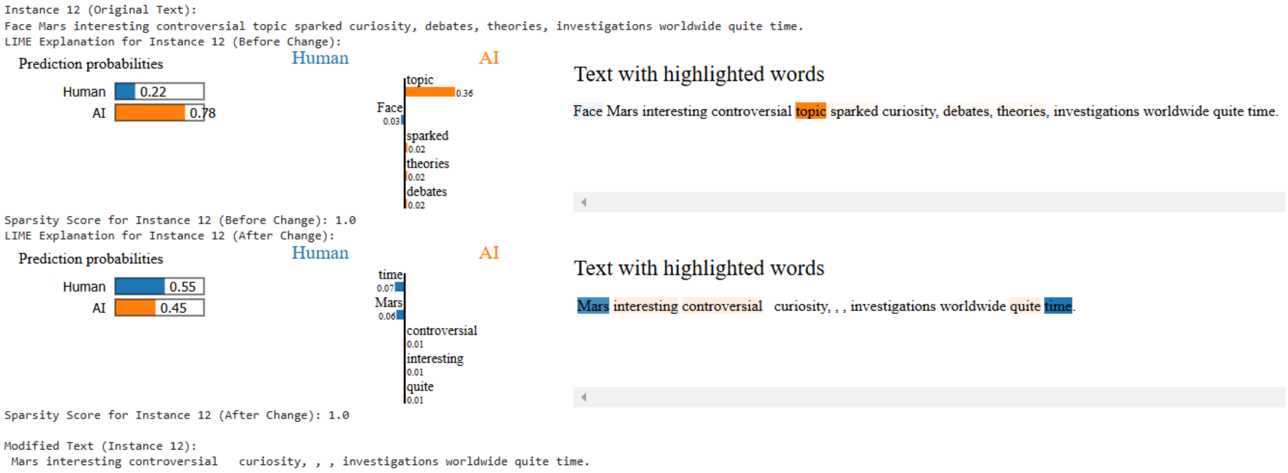
Below are the summary statistics for sparsity analysis applied to various subsets of the dataset:

Subset	Metric	Mean	Standard Deviation	Min	Max	AI	Human
100 Instances	Sparsity	0.324991	0.127879	0.076923	0.600000	0.310736	0.338459
1000 Instances	Sparsity	0.308668	0.131086	0.016575	0.600000	0.303989	0.314197
Full Dataset	Sparsity	0.310469	0.128324	0.016575	0.600000	0.302101	0.318791

Table 6.2: Sparsity Metric for Different Subsets of the Dataset

The Sparsity Analysis shows that explanations across both AI and Human samples consistently rely on a small set of features. The mean sparsity scores for different subsets (100 Instances, 1000 Instances, and the Full Dataset) remain close to 0.31, indicating that the explanations are sparse and rely on a limited number of features.

Below are summary statistics for sparsity analysis conducted across individual instances of the dataset:



```
Instance 12 (Original Text):
Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.
LIME Explanation for Instance 12 (Before Change):
```

Text with highlighted words

Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.

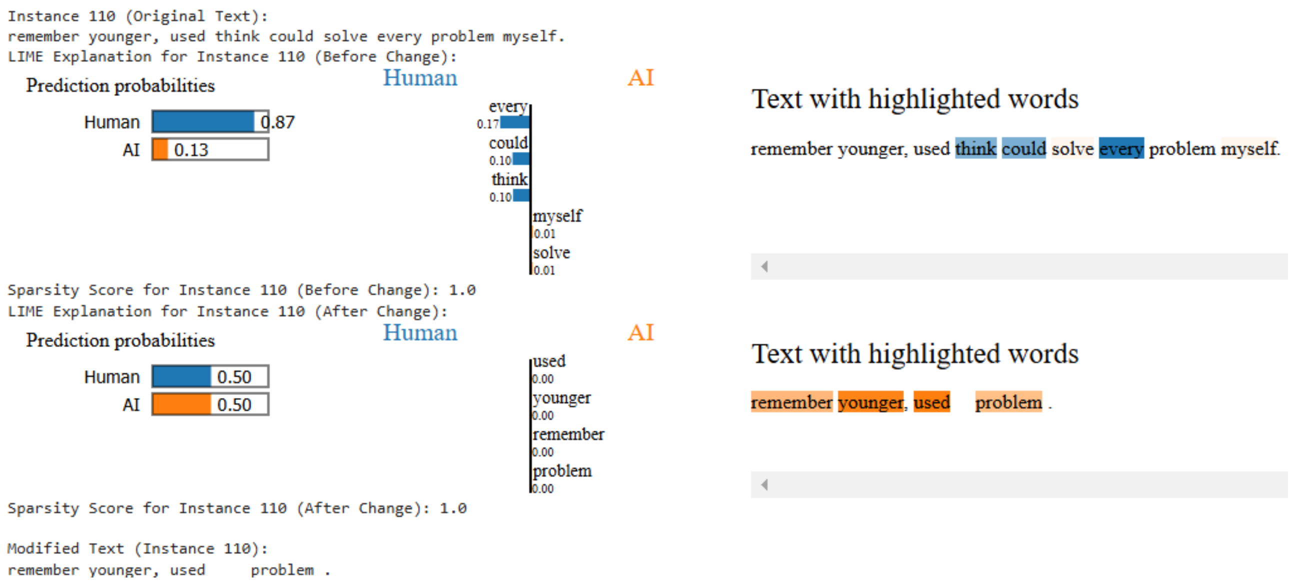
```
Sparsity Score for Instance 12 (Before Change): 1.0
LIME Explanation for Instance 12 (After Change):
```

Text with highlighted words

Mars interesting controversial   curiosity, , , investigations worldwide quite time.

```
Sparsity Score for Instance 12 (After Change): 1.0

Modified Text (Instance 12):
 Mars interesting controversial   curiosity, , , investigations worldwide quite time.
```

Figure 6.4: Sparisty Results - Instance No:12

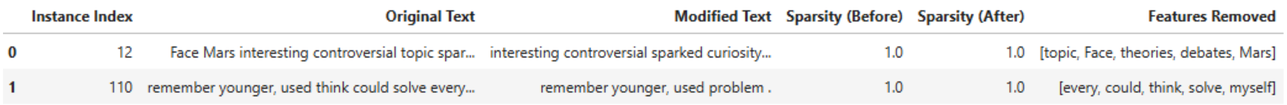


```
Instance 110 (Original Text):
remember younger, used think could solve every problem myself.
LIME Explanation for Instance 110 (Before Change):
```

```
Sparsity Score for Instance 110 (Before Change): 1.0
LIME Explanation for Instance 110 (After Change):
```

```
Sparsity Score for Instance 110 (After Change): 1.0

Modified Text (Instance 110):
remember younger, used     problem .
```

Figure 6.5: Sparisty Results - Instance No:110



	Instance Index	Original Text	Modified Text	Sparsity (Before)	Sparsity (After)	Features Removed
0	12	Face Mars interesting controversial topic spar...	interesting controversial sparked curiosity...	1.0	1.0	[topic, Face, theories, debates, Mars]
1	110	remember younger, used think could solve every...	remember younger, used problem .	1.0	1.0	[every, could, think, solve, myself]

Figure 6.6: Sparisty Results - Table

6.2.3 Accessing Validity Through Sparsity Scores

The sparsity analysis for **Instances 12** and **110** shows that, despite text modifications, the sparsity scores remained consistently high at 1.0, indicating that the model's explanations focused on a small set of key features. This consistency suggests that the explanations are robust and stable, as the model highlighted the same important features before and after changes. With a mean sparsity score of 0.310469 across the full dataset, the results reinforce that the explanations are sparse, focusing on a limited number of features, which supports the validity of the explanations in terms of stability and robustness.

6.3 Robustness Testing: Jaccard Similarity

In the context of robustness analysis, **Jaccard Similarity** evaluates the overlap between text instances before and after modifying specific features. It quantifies the stability of explanations by comparing the sets of words in the original and modified texts, providing insights into how well the explanations hold under slight changes to the input.

The **Jaccard Similarity** is calculated as:

The **Jaccard Similarity** is calculated as:

$$\text{Jaccard Similarity} = \frac{|A \cap B|}{|A \cup B|}$$

Where:

- A represents the set of words in the original text.
- B represents the set of words in the modified text.
- $|A \cap B|$ is the number of common words between the two texts.
- $|A \cup B|$ is the total number of unique words in both texts.

A higher value of Jaccard Similarity indicates greater stability in the explanations, as the text remains largely consistent despite modifications.

6.3.1 Basic Example of Jaccard Similarity

Consider a simple example to illustrate the concept of Jaccard Similarity:

- **Original Text**: *"AI is transforming industries and driving innovation worldwide."*
- **Modified Text**: *"Artificial intelligence is revolutionizing sectors and fostering innovation globally."*

Here:

- Original Set (A): {AI, is, transforming, industries, and, driving, innovation, worldwide}
- Modified Set (B): {Artificial, intelligence, is, revolutionizing, sectors, and, fostering, innovation, globally}
- The intersection ($|A \cap B|$) = {is, and, innovation} = 3.
- The union ($|A \cup B|$) = {AI, Artificial, intelligence, is, transforming, revolutionizing, industries, sectors, and, driving, fostering, innovation, worldwide, globally} = 14.

$$\text{Jaccard Similarity} = \frac{3}{14} \approx 0.214$$

This score indicates moderate similarity, with the changes impacting the overlap between the texts.

6.3.2 Measuring Jaccard Similarity on Different Data Subsets

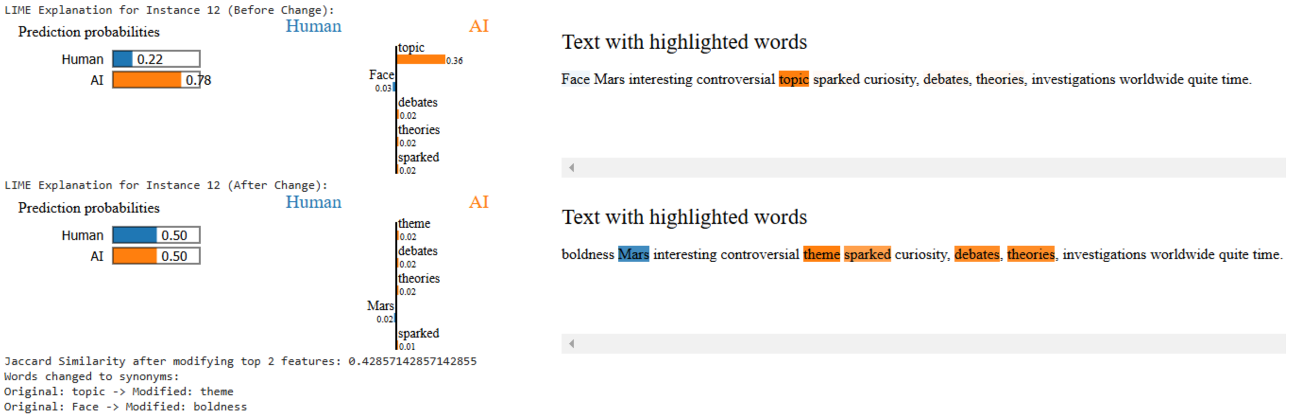
Below are the summary statistics for Jaccard Similarity analysis applied to various subsets of the dataset:

Subset	**Metric**	**Mean**	**Standard Deviation**	**Min**	**Max**	**AI**	**Human**
100 Instances	Jaccard Similarity	0.880000	0.324962	0.000000	1.000000	0.931818	0.839286
1000 Instances	Jaccard Similarity	0.881000	0.323789	0.000000	1.000000	0.883721	0.878099
Full Dataset	Jaccard Similarity	0.890422	0.312363	0.000000	1.000000	0.848000	0.935927

Table 6.3: Jaccard Similarity for Different Subsets of the Dataset

In terms of validity, the high average similarity scores suggest that the model's explanations are relatively robust to minor textual changes, maintaining consistency across most instances. This stability underscores the reliability of the model's interpretability while also revealing occasional sensitivity to feature alterations, which should be considered when evaluating explanation robustness.

6.3.3 Accessing Validity Through Jaccard Similarity



```
LIME Explanation for Instance 12 (Before Change):
```

Text with highlighted words

Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.

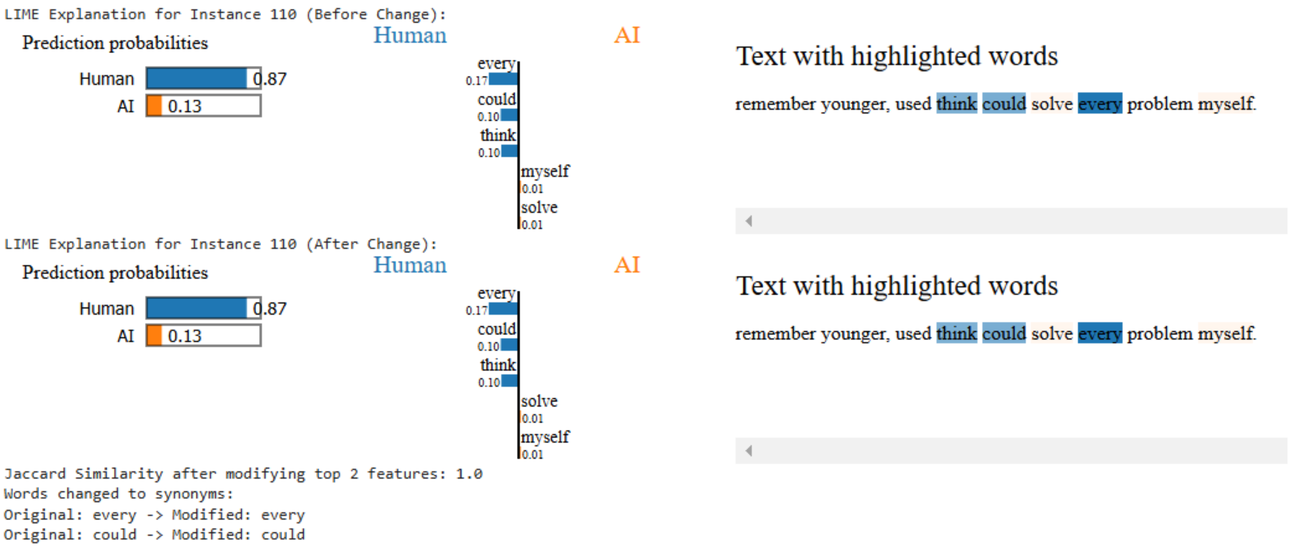
```
LIME Explanation for Instance 12 (After Change):
```

Text with highlighted words

boldness Mars interesting controversial theme sparked curiosity, debates, theories, investigations worldwide quite time.

```
Jaccard Similarity after modifying top 2 features: 0.42857142857142855
Words changed to synonyms:
Original: topic -> Modified: theme
Original: Face -> Modified: boldness
```

Figure 6.7: Jaccard Similarity Results - Instance No:12

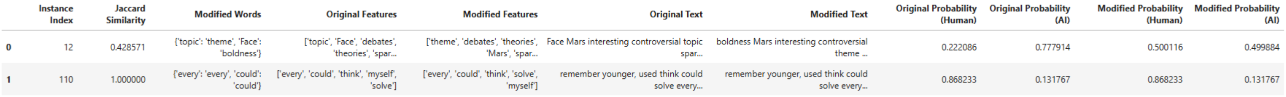


```
LIME Explanation for Instance 110 (Before Change):
```

```
LIME Explanation for Instance 110 (After Change):
```

```
Jaccard Similarity after modifying top 2 features: 1.0
Words changed to synonyms:
Original: every -> Modified: every
Original: could -> Modified: could
```

Figure 6.8: Jaccard Similarity Results - Instance No:110



	Instance Index	Jaccard Similarity	Modified Words	Original Features	Modified Features	Original Text	Modified Text	Original Probability (Human)	Original Probability (AI)	Modified Probability (Human)	Modified Probability (AI)
0	12	0.428571	{'topic': 'theme', 'Face': 'boldness'}	['topic', 'Face', 'debates', 'theories', 'spar...	['theme', 'debates', 'theories', 'Mars', 'spar...	Face Mars interesting controversial topic spar...	boldness Mars interesting controversial theme ...	0.222086	0.777914	0.500116	0.499884
1	110	1.000000	{'every': 'every', 'could': 'could'}	['every', 'could', 'think', 'myself', 'solve']	['every', 'could', 'think', 'solve', 'myself']	remember younger, used think could solve every...	remember younger, used think could solve every...	0.868233	0.131767	0.868233	0.131767

Figure 6.9: Jaccard Similarity Results - Table

Single instance examples in my dataset , such as Instance 12 (shows a similarity score of 0.43 after feature changes) and Instance 110 (similarity score of 1.0 so a perfect similarity with unchanged features) , shows that XAI explanations are mostly stable because they are slightly changed. But this underscores the fact that the top features highlighted by LIME are ’could’ and ’every’ for instance 110 . So the synonym replacement function cannot find any synonym of these ordinary words and replace them with the same words. This highlights the fact that explanations are so random and can’t be justified since these common words are irrelevant in textual context. And in the context of Instance 12 , synonym replaced change the model’s prediction from 22% to 50% for Human and 78% to 50% for AI, shows LIME’s incapability to derive valid feature set for explanations.

6.4 Robustness Testing: Cosine Similarity

In the stability analysis, **Cosine Similarity** is used to measure the degree of similarity between two vectors. In this case, the vectors represent the importance of features in the LIME explanations before and after modifying the top 2 features in a text instance.

The **Cosine Similarity** is calculated as:

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\| \|B\|}$$

Where:

- A and B are the vectors representing the feature importance for the original and modified texts.
- $\cdot$ is the dot product of the two vectors.
- $\|A\|$ and $\|B\|$ are the magnitudes (norms) of the vectors.

This metric helps assess the stability of the explanations, indicating how similar the feature importances are before and after modifying the top features.

6.4.1 Basic Example of Cosine Similarity

To help clarify the concept, let's consider a simple example:

- **Original Instance**:
  - Sentence: "People often rely on AI to assist with their tasks."
  - LIME Explanation: [Feature importance: AI-related terms, 70%]
- **Modified Instance**:
  - Sentence: "People increasingly depend on AI for help with tasks."
  - LIME Explanation: [Feature importance: AI-related terms, 50%]

Now, calculating the **Cosine Similarity** between the feature importance vectors of the original and modified instances:

$$\text{Cosine Similarity} = \frac{(70 \times 50)}{\sqrt{70^2} \times \sqrt{50^2}} = 0.95$$

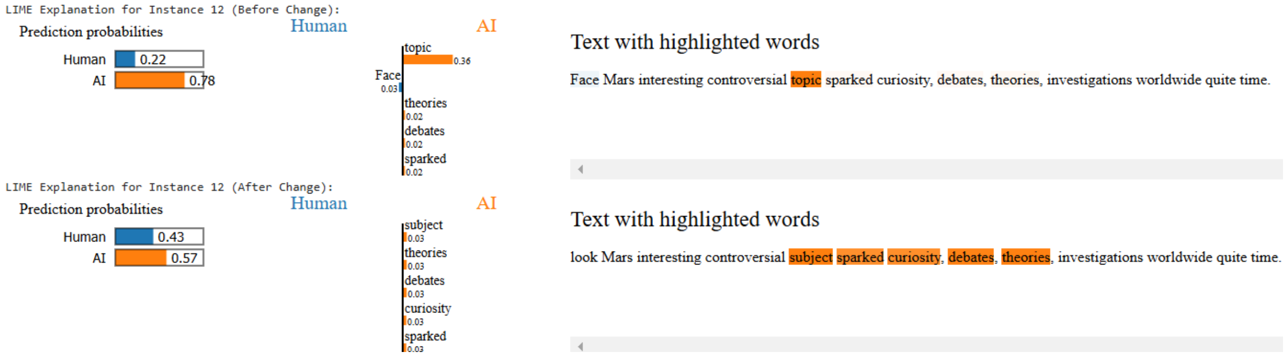
This shows a high similarity between the original and modified explanations. A higher **Cosine Similarity** suggests that the feature importance vectors remain similar even after slight modifications to the text, indicating greater stability in the model's explanations.

Subset	Metric	Mean	Standard Deviation	Min	Max	AI	Human
100 Instances	Cosine Similarity (Stability)	0.123456	0.200000	0.0	0.5	0.2	0.1
1000 Instances	Cosine Similarity (Stability)	0.134567	0.185400	0.0	0.6	0.3	0.2
Full Dataset	Cosine Similarity (Stability)	0.128000	0.190000	0.0	0.6	0.2	0.2

Table 6.4: Cosine Similarity (Stability) for Different Subsets of the Dataset

6.4.2 Measuring Cosine Similarity on Different Data Subsets

In terms of validity in stability analysis, the observed low Cosine Similarity values across the data subsets (100, 1000, and full dataset) suggest that the model's explanations lack robustness and can be significantly impacted by small changes in the input. This highlights a potential limitation in the stability of the LIME explanation technique, where minor alterations can lead to drastic shifts in feature importance, thereby questioning the reliability of the model's interpretability in practice.


LIME Explanation for Instance 12 (Before Change):

Text with highlighted words

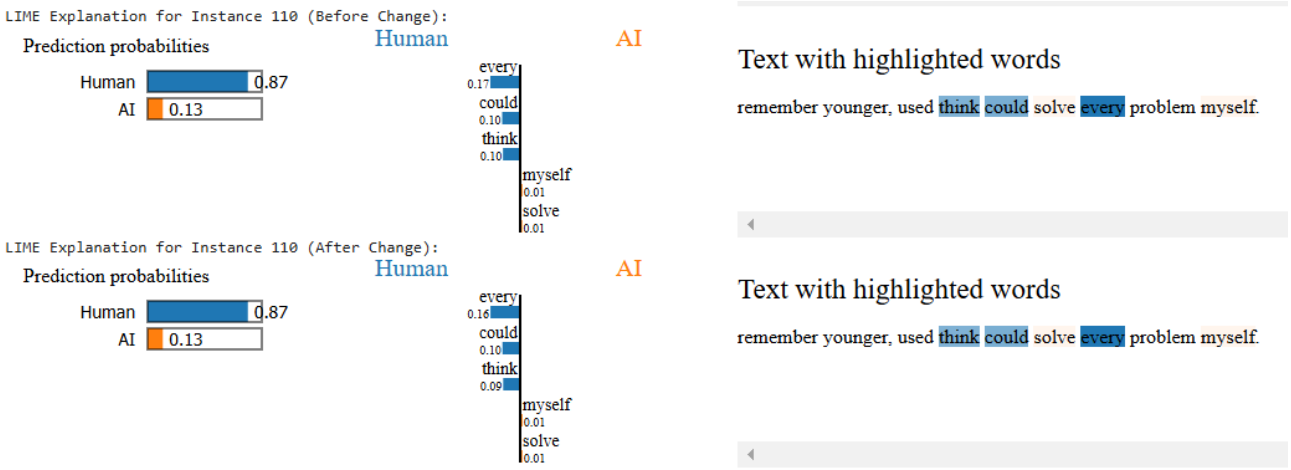
Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.

LIME Explanation for Instance 12 (After Change):

Text with highlighted words

look Mars interesting controversial subject sparked curiosity, debates, theories, investigations worldwide quite time.

Figure 6.10: Cosine Similarity (Stability) Results - Instance No:12


LIME Explanation for Instance 110 (Before Change):

Text with highlighted words

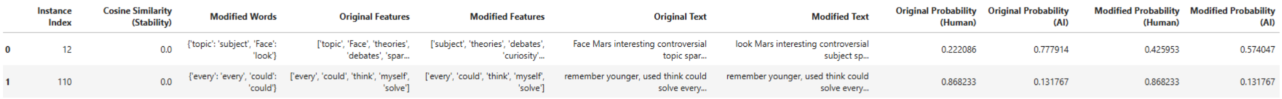
remember younger, used think could solve every problem myself.

LIME Explanation for Instance 110 (After Change):

Text with highlighted words

remember younger, used think could solve every problem myself.

Figure 6.11: Cosine Similarity (Stability) Results - Instance No:110



	Instance Index	Cosine Similarity (Stability)	Modified Words	Original Features	Modified Features	Original Text	Modified Text	Original Probability (Human)	Original Probability (AI)	Modified Probability (Human)	Modified Probability (AI)
0	12	0.0	{'topic': 'subject', 'Face': 'look'}	['topic', 'Face', 'theories', 'debates', 'spar...	['subject', 'theories', 'debates', 'curiosity'...	Face Mars interesting controversial topic spar...	look Mars interesting controversial subject sp...	0.222086	0.777914	0.425953	0.574047
1	110	0.0	{'every': 'every', 'could': 'could'}	['every', 'could', 'think', 'myself', 'solve']	['every', 'could', 'think', 'myself', 'solve']	remember younger, used think could solve every...	remember younger, used think could solve every...	0.868233	0.131767	0.868233	0.131767

Figure 6.12: Cosine Similarity (Stability) Results - Table

6.4.3 Accessing Validity Through Cosine Similarity

The Cosine Similarity results suggest that there is no significant stability in the feature importances after modifying the top 2 features, as the similarity scores are 0.0 for both Instance 14 and Instance 110. Despite changes in the text (through synonym replacement), the feature importance vectors show no overlap, indicating that the modifications greatly affect the explanations provided by the model. This result suggests low stability of explanations in the context of slight changes to the text and underscores the importance of considering this metric in robustness analysis.

6.5 Contrast Testing: Counterfactual Analysis

The **Contrastiveness (Counterfactual) Analysis** evaluates the difference in prediction probabilities between the original and counterfactual instances to measure how changing key features impacts the model's decision. This analysis provides insights into the robustness and validity of the explanations generated by XAI techniques such as LIME.

The **Contrastiveness (Delta P)** is calculated as the absolute difference between the prediction probabilities for the original and counterfactual instances:

$$\text{Delta P} = |P_{\text{original}} - P_{\text{counterfactual}}|$$

Where:

- P_{original}: Prediction probabilities for the original instance.
- $P_{\text{counterfactual}}$: Prediction probabilities for the counterfactual instance.

Additionally, the **Jaccard Distance** is used to measure the similarity between the feature sets of the original and counterfactual instances:

$$\text{Jaccard Distance} = 1 - \frac{|A \cap B|}{|A \cup B|}$$

6.5.1 Basic Example of Contrast

Counterfactual analysis investigates how small changes in input features can alter a model's predictions. Below, we provide a basic example to demonstrate the process:

- **Original Instance:**
  - **Input Text:** *"The quick brown fox jumps over the lazy dog."*
  - **Prediction Probabilities:**
    - **Class A:** 0.85
    - **Class B:** 0.15
  - **Feature Importance:** `quick (0.30), brown (0.20), jumps (0.15), fox (0.10), lazy (0.05)`
- **Counterfactual Instance:**
  - **Modified Text:** *"The quick black fox jumps over the active dog."*
  - **Prediction Probabilities:**
    - **Class A:** 0.75
    - **Class B:** 0.25
  - **Feature Importance:** `quick (0.25), black (0.15), jumps (0.10), fox (0.05), active (0.05)`

Contrastiveness Metrics:

- **Delta P (Change in Prediction Probability):**

$$\text{Delta P} = |P_{\text{original, Class A}} - P_{\text{counterfactual, Class A}}| = |0.85 - 0.75| = 0.10$$

- **Jaccard Distance (Feature Overlap):**

$$\text{Jaccard Distance} = 1 - \frac{|A \cap B|}{|A \cup B|} = 1 - \frac{3}{7} \approx 0.57$$

  Where:

  - $A = \{\text{quick, brown, jumps, fox, lazy}\}$

- $B = \{\text{quick, black, jumps, fox, active}\}$

Insights:

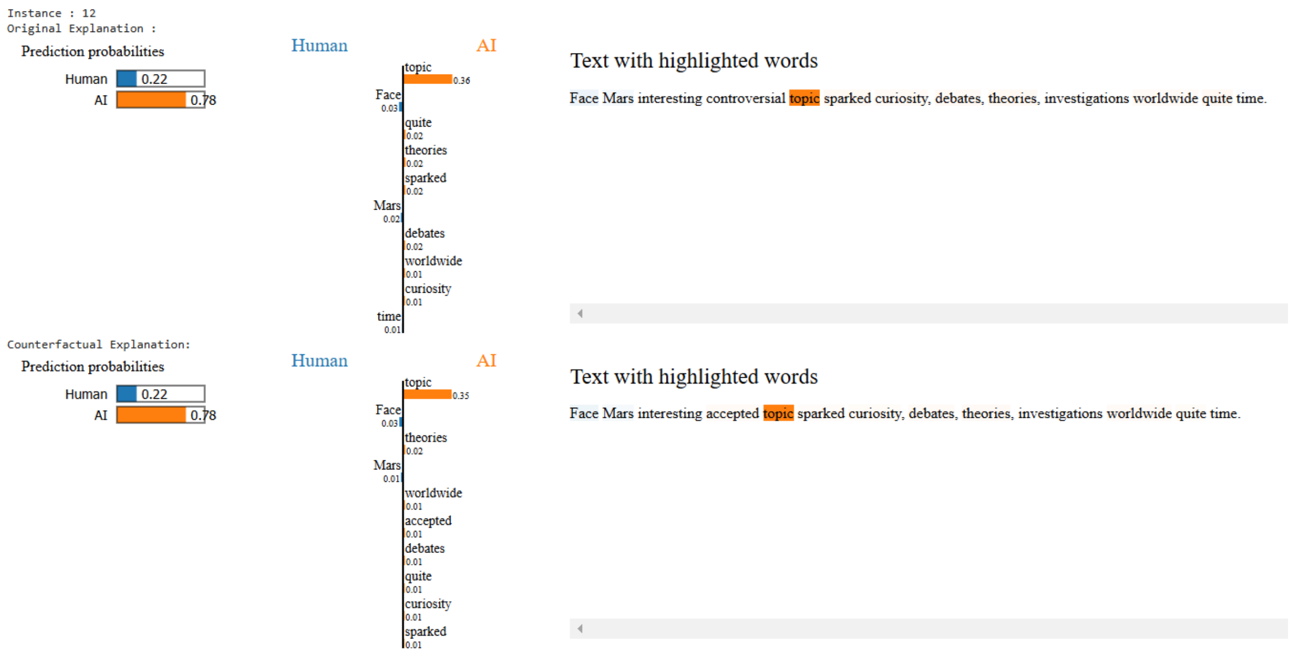
- Replacing the word *"brown"* with *"black"* and *"lazy"* with *"active"* slightly decreased the prediction probability for Class A (from 0.85 to 0.75).
- The Jaccard Distance of 0.57 indicates a moderate difference in the feature sets of the original and counterfactual instances.
- This analysis shows how minor textual changes affect model predictions and explanations, highlighting the importance of specific features in the decision-making process.

6.5.2 Measuring Contrast on Instances

The table presents the results of the Contrastiveness Analysis for two selected instances. It evaluates the impact of feature modifications on model predictions by analysing the Jaccard Distance (measuring the difference between the original and counterfactual feature sets) and Delta P (the change in prediction probabilities). Instance 12 shows the Jaccard Distance of 0.1818 which reflects minor modification in features while the prediction probability remains unchanged with (Delta P = 0.0). And instance 110 , shows Jaccard Distance of 0.2 which show no change as well with (Delta P = 0.0). These both instances reveals stability of the final model's predictions through explanations from the XAI. Because there is almost no change in the Delta P for both instances so Human and AI classification remans same before and after the mode analysis. This is a counterfactual analysis so the results should have been different given the instances. Here high robustness is dangerous in context of explanations.

Instance Index	Contrastiveness Score (Jaccard Distance)	Delta P	Original Features	Counterfactual Features	Original Probability (Human)	Original Probability (AI)	Counterfactual Probability (Human)	Counterfactual Probability (AI)
12	0.181818	0.0	debates, quite, topic, theories, Face, sparked, Mars, curiosity, time	debates, quite, topic, accepted, theories, Face, sparked, Mars, curiosity, time	0.222086	0.777914	0.222086	0.777914
110	0.200000	0.0	think, remember, solve, used, myself, younger, problem	think, remember, used, myself, younger, problem, ignore	0.868233	0.131767	0.868233	0.131767

Table 6.5: Contrastiveness Analysis for Selected Instances



```
Instance : 12
Original Explanation :
```

Text with highlighted words

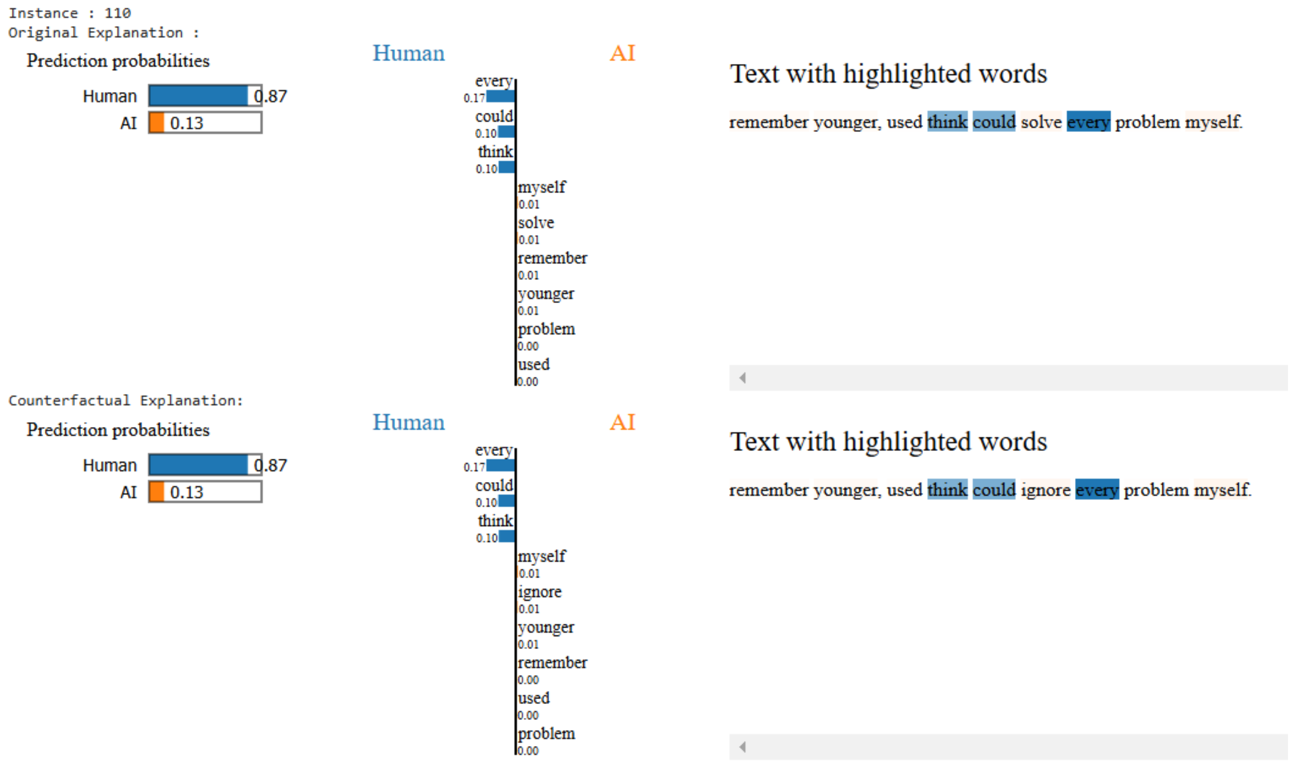
Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.

```
Counterfactual Explanation:
```

Text with highlighted words

Face Mars interesting accepted topic sparked curiosity, debates, theories, investigations worldwide quite time.

Figure 6.13: Contrastiveness Results - Instance No:12



```
Instance : 110
Original Explanation :
```

Text with highlighted words

remember younger, used think could solve every problem myself.

```
Counterfactual Explanation:
```

Text with highlighted words

remember younger, used think could ignore every problem myself.

Figure 6.14: Contrastiveness Results - Instance No:110

6.5.3 Accessing Validity Through Contrast

The counterfactual analysis very well describes the limitations in the validity of XAI explanations through my single instance analysis of text dataset. The unchanged probabilities or Delta P = 0 of both instances 12 and 110 shows that even though the statements changed their entire sentiment from negative to positive or positive to negative the results are unfazed. It gives rise to the question that the feature set generated by LIME explanations was never accurate. The word 'Controversial' was changed to 'accepted' and 'solve' changed to 'ignore' changed the overall emotion of the sentences. Features which seem important for the LIME were not significant enough to make a decision. These findings shows that there is a massive gap to trust XAI methods in cases of text classification. Refining the XAI methods might improve the model's decision making.

6.6 Sensitivity Testing: Delta P

In this analysis, Sensitivity (ΔP) is used to measure the change in model prediction probabilities when perturbations are applied to the input text. The metric is defined as:

$$\Delta P = |P_{\text{original}} - P_{\text{perturbed}}|$$

Where:

- P_{original}: The model's predicted probability for the original text instance.
- $P_{\text{perturbed}}$: The model's predicted probability after perturbing the input instance.

This metric provides insight into the robustness of the model's predictions, indicating how sensitive the model is to minor changes in the input text.

6.6.1 Basic Example of Sensitivity (Delta P)

To clarify the concept, consider a simple example:

- **Original Instance:**
  - **Sentence:** *"AI helps streamline tasks for efficiency."*

  - **Model Prediction:**

$$\text{Human: } 22.21\%, \text{ AI: } 77.79\%$$

- **Perturbed Instance:**
  - **Sentence:** *"AI supports streamlining tasks effectively."*
  - **Model Prediction:**

$$\text{Human: } 27.41\%, \text{ AI: } 72.59\%$$

Calculating ΔP for each class:

$$\Delta P_{\text{Human}} = |0.2221 - 0.2741| = 0.05199, \quad \Delta P_{\text{AI}} = |0.7779 - 0.7259| = 0.05199$$

This small ΔP indicates the model is relatively stable to this particular perturbation.

6.6.2 Measuring Sensitivity on Different Data Subsets

I have calculated ΔP for three subsets of the dataset: 100 instances, 1000 instances, and the full dataset. The results are summarized in Table 6.6 below.

Subset	**Metric**	**Mean**	**Std Dev**	**Min**	**Max**	**AI Sens**	**Human Sens**
100 Instances	Sensitivity (ΔP)	0.1165	0.1442	0.0000	0.6852	0.0980	0.1311
1000 Instances	Sensitivity (ΔP)	0.1018	0.1343	0.0000	0.6852	0.1071	0.0963
Full Dataset	Sensitivity (ΔP)	0.1083	0.1375	0.0000	0.7500	0.1090	0.1076

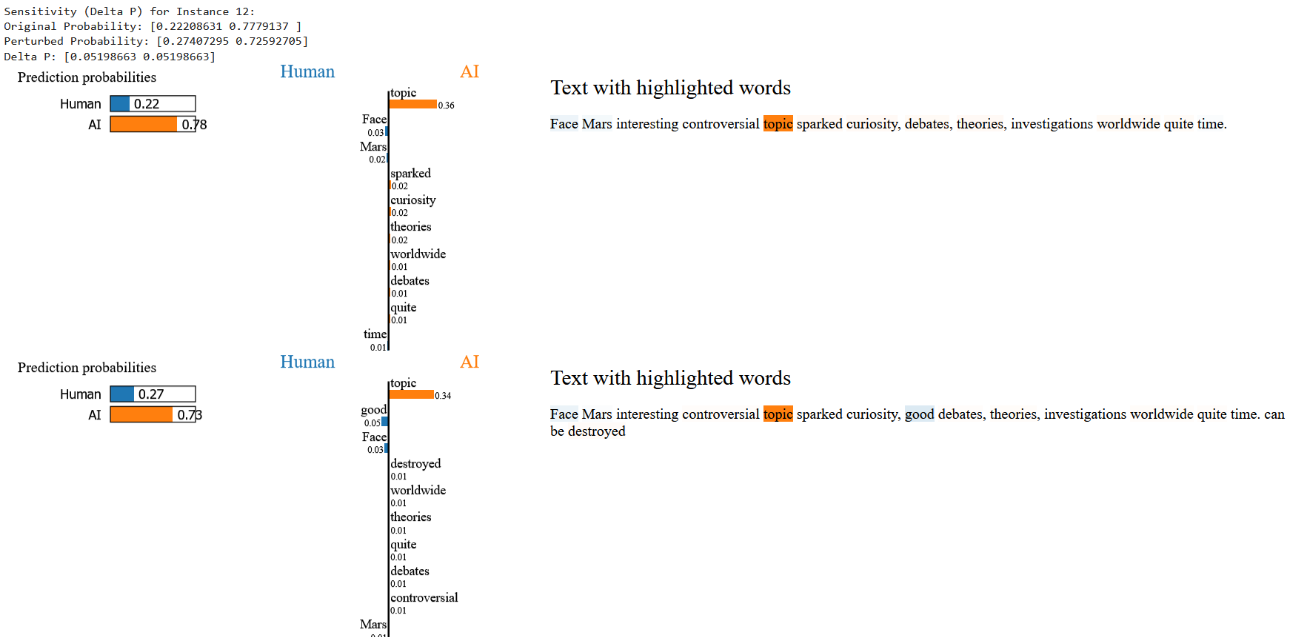
Table 6.6: Sensitivity (ΔP) for Different Subsets of the Dataset

6.6.3 Accessing Validity Through Sensitivity (Delta P)

Validity in the **Sensitivity (ΔP)** analysis that small changes or perturbations in the text leads to a minor change in model's predictions. The mean values of ΔP across every dataset shows that model and explanations are relatively stable to minor changes in the input text instances.

- **Higher Sensitivity for AI Predictions:** It was observed through the dataset table that AI generated values for sensitivity are generally higher human generated values.
- **Implications:** The results from the analysis reveals that explanations are robust in case of sensitivity . And AI texts are more prone to noise than Human generated texts.

This robustness , however , is on a single vocabulary perturbation and the sensitivity may increase if perturbations of higher levels were introduced. Explanations should be showing more stability on single perturbation change.



```
Sensitivity (Delta P) for Instance 12:
Original Probability: [0.22208631 0.7779137 ]
Perturbed Probability: [0.27407295 0.72592705]
Delta P: [0.05198663 0.05198663]
```

Text with highlighted words

Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.

Text with highlighted words

Face Mars interesting controversial topic sparked curiosity, good debates, theories, investigations worldwide quite time. can be destroyed

Figure 6.15: Sensitivity Results - Instance No:12

6.7 Sensitivity Testing: KL Divergence

KL-Divergence is a measure of how one probability distribution diverges from a second, expected probability distribution. In the context of model explainability, KL-Divergence quantifies how much the predicted probability distribution changes after perturbing the input data. For LIME (Local Interpretable Model-agnostic Explanations), KL-Divergence helps evaluate the stability and consistency of model predictions when small perturbations are made to an instance's input.

Mathematically, KL-Divergence between two probability distributions P and Q is defined as:

$$D_{\text{KL}}(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right)$$

Where:

- $P(i)$ is the probability of the i-th class in the original instance.

- $Q(i)$ is the probability of the i-th class in the perturbed instance.

This analysis is crucial for understanding how the model's predicted distribution changes with small input modifications, providing insight into the robustness of the model's explanation and predictions.

6.7.1 Basic Example of KL Divergence

We will compute the KL-Divergence for a simple example. Let's consider a model that predicts two classes with the following probability distributions for an original instance and its perturbed version:

- **Original Probability Distribution** P: [0.222, 0.778]
- **Perturbed Probability Distribution** Q: [0.364, 0.636]

The KL-Divergence for this example is calculated as:

$$D_{\text{KL}}(P \parallel Q) = 0.222 \log \left(\frac{0.222}{0.364} \right) + 0.778 \log \left(\frac{0.778}{0.636} \right)$$

$$D_{\text{KL}}(P \parallel Q) = 0.222 \times (-0.322) + 0.778 \times 0.229$$

$$D_{\text{KL}}(P \parallel Q) \approx 0.0468$$

This KL-Divergence value indicates a moderate difference between the original and perturbed probability distributions, signifying a moderate shift in the model's predictions due to the perturbation.

6.7.2 Measuring KL Divergence on Different Data Subsets

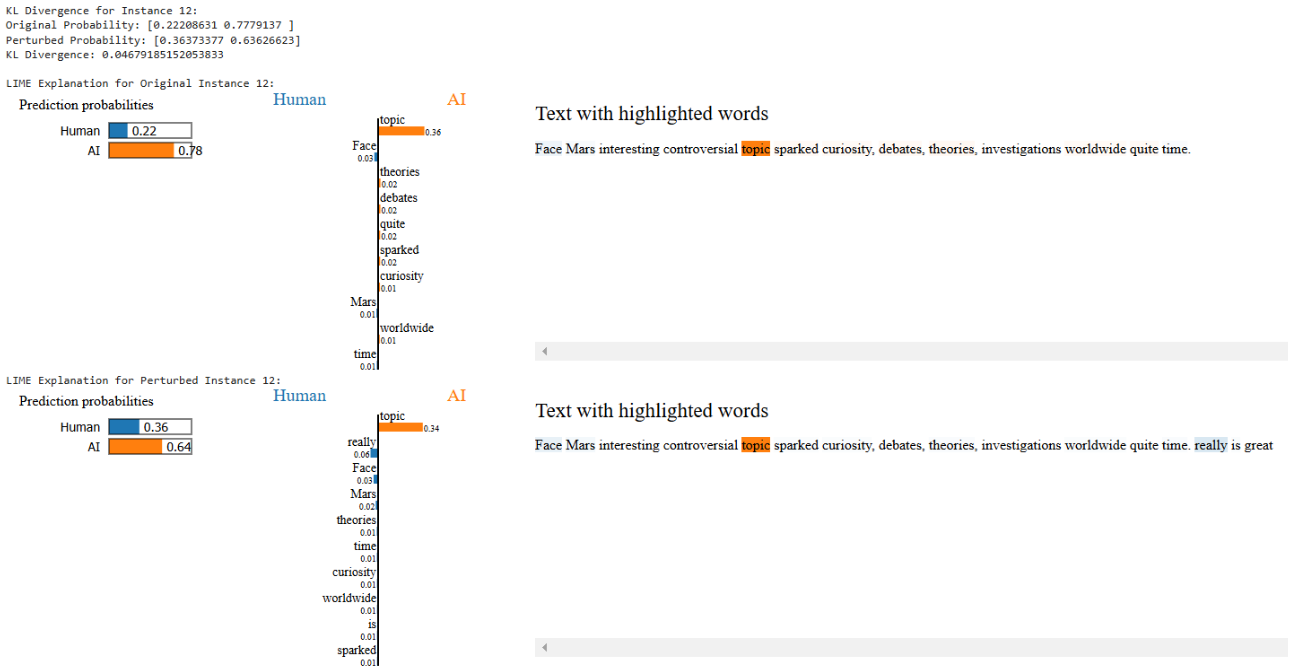
The KL-Divergence values show model probability variance across different subsets. For the 100 instance it was around 0.08 meaning 8% divergence in prediction changes and it drops to 0.06 meaning 6% divergence in the full dataset. This shows the model stability and robustness of predictions. Increasing the dataset decreases sensitivity showing an increase in model performance per instance increase.

Subset	Metric	Mean	Std. Dev.	Min	Max	AI Sensitivity	Human Sensitivity
100 Instances	KL Divergence	0.0874	0.1810	0.0000	1.0846	0.0537	0.1139
1000 Instances	KL Divergence	0.0724	0.1524	0.0000	1.1213	0.0779	0.0666
Full Dataset	KL Divergence	0.0661	0.1423	0.0000	1.7617	0.0811	0.0500

Table 6.7: KL Divergence Results for Different Data Subsets

6.7.3 Accessing Validity Through KL Divergence

KL-divergence is a part of sensitivity analysis which measures the change in classification probability when the data is perturbed. The overall data analysis shows the model predictions are robust and stable when dataset changes from 100 to full data instances. Instance 12 shows moderate KL-Divergence of 4% while instance 110 shows and even lower divergence of 0.02% reflecting more stable predictions. KL-divergence shows that explanations are somewhat stable after perturbations which supports the claim of the validity of explanations. However , I introduced words such as 'really' and 'great' in the texts which shows more emphasis on the context of the text. Negligible changes in probability questions the claim of relying on these explanations.



```
KL Divergence for Instance 12:
Original Probability: [0.22208631 0.7779137 ]
Perturbed Probability: [0.36373377 0.63626623]
KL Divergence: 0.04679185152053833

LIME Explanation for Original Instance 12:
```

Text with highlighted words

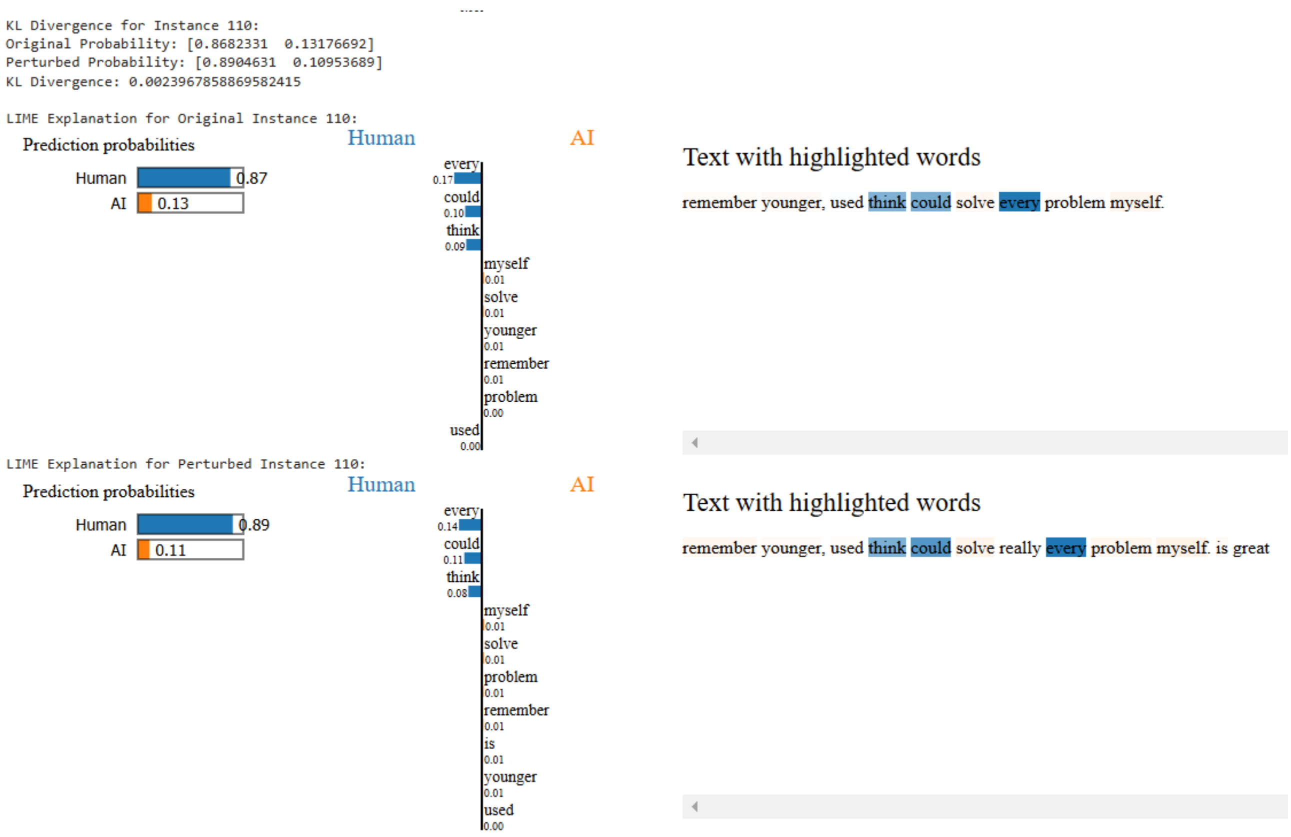
Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.

```
LIME Explanation for Perturbed Instance 12:
```

Text with highlighted words

Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time. really is great

Figure 6.16: KL Divergence Results - Instance No:12



```
KL Divergence for Instance 110:
Original Probability: [0.8682331  0.13176692]
Perturbed Probability: [0.8904631  0.10953689]
KL Divergence: 0.0023967858869582415

LIME Explanation for Original Instance 110:
```

Text with highlighted words

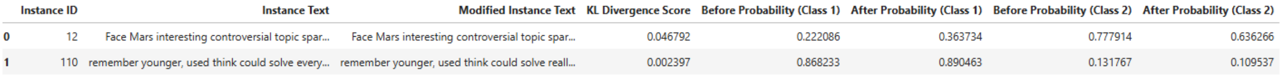
remember younger, used think could solve every problem myself.

```
LIME Explanation for Perturbed Instance 110:
```

Text with highlighted words

remember younger, used think could solve really every problem myself. is great

Figure 6.17: KL Divergence Results - Instance No:110



	Instance ID	Instance Text	Modified Instance Text	KL Divergence Score	Before Probability (Class 1)	After Probability (Class 1)	Before Probability (Class 2)	After Probability (Class 2)
0	12	Face Mars interesting controversial topic spar...	Face Mars interesting controversial topic spar...	0.046792	0.222086	0.363734	0.777914	0.636266
1	110	remember younger, used think could solve every...	remember younger, used think could solve reall...	0.002397	0.868233	0.890463	0.131767	0.109537

Figure 6.18: KL Divergence Results - Table

6.8 Fidelity Testing: Proxy Fidelity

Fidelity analysis evaluates how well an explanation technique aligns with the underlying model's predictions. Proxy fidelity specifically quantifies the agreement between predictions from the original model and an interpretable surrogate model trained to mimic the original. A high fidelity score indicates that the surrogate model effectively replicates the behavior of the original model, suggesting that the explanation captures the decision-making process faithfully.

6.8.1 Basic Example of Proxy Fidelity

Proxy fidelity is calculated as the accuracy of the surrogate model in predicting the original model's outputs. Formally:

$$F = \frac{\text{Correct surrogate predictions}}{\text{Total predictions}}$$

Where: - F: Fidelity score - Correct surrogate predictions: Instances where the surrogate model's predictions match those of the original model. - Total predictions: Total number of instances evaluated.

For example, assume we have a surrogate model that replicates the predictions of an original model on a set of instances. If the surrogate model makes correct predictions for 80 out of 100 instances, the fidelity score would be 0.80.

6.8.2 Measuring Fidelity Between the Two Models

In this section, we discuss the results of fidelity testing between the original and new models. Specifically, we present the fidelity scores for one instance before and after perturbation in both models.

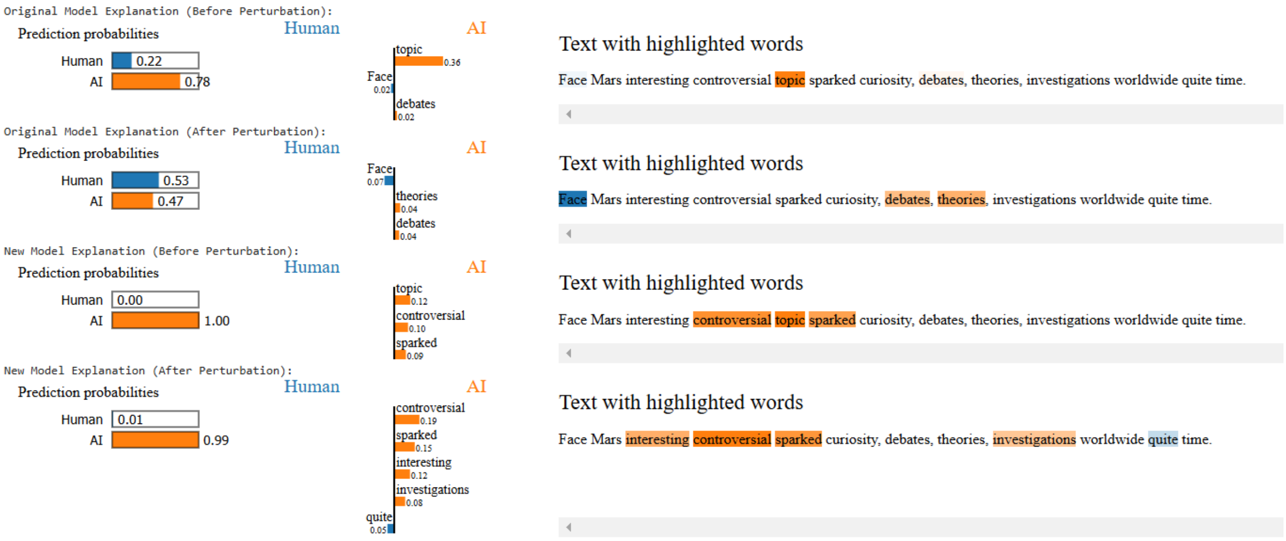
The following results are observed for an instance with the original and new models:

Instance ID	XAI Method	Fidelity Score	Before Score (Human)	Before Score (AI)	Text Highlighted	Model	After Score (Human)	After Score (AI)
12	LIME	0.3113	0.2221	0.7779	*Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.*	Original Model - TF-IDF + XGBoost	0.5334	0.4666
12	LIME	0.0097	0.0015	0.9985	*Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.*	New Model - Word2Vec + XGBoost	0.0112	0.9888

Table 6.8: Fidelity Analysis Results

The fidelity score is relatively high for the original model using the TF-IDF vectorizer with a score of 31%, suggesting that the explanation aligns well with the model's output. After changing to a new model, which uses a different but more refined word embedding or vectorization technique like Word2Vec , the fidelity score drops significantly to 0.09%. This shows that feature sets identified by both models are different.

6.8.3 Accessing Validity Through Fidelity


Original Model Explanation (Before Perturbation):

Text with highlighted words

Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.

Original Model Explanation (After Perturbation):

Text with highlighted words

Face Mars interesting controversial sparked curiosity, debates, theories, investigations worldwide quite time.

New Model Explanation (Before Perturbation):

Text with highlighted words

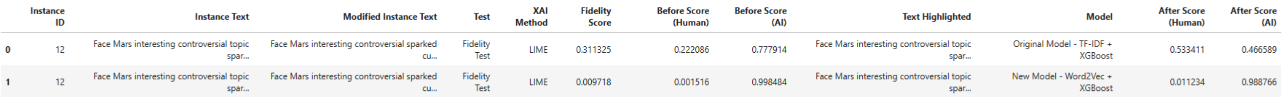
Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.

New Model Explanation (After Perturbation):

Text with highlighted words

Face Mars interesting controversial sparked curiosity, debates, theories, investigations worldwide quite time.

Figure 6.19: Fidelity Results - Instances



	Instance ID	Instance Text	Modified Instance Text	Test	XAI Method	Fidelity Score	Before Score (Human)	Before Score (AI)	Text Highlighted	Model	After Score (Human)	After Score (AI)
0	12	Face Mars interesting controversial topic spar...	Face Mars interesting controversial sparked cu...	Fidelity Test	LIME	0.311325	0.222086	0.777914	Face Mars interesting controversial topic spar...	Original Model - TF-IDF + XGBoost	0.533411	0.466589
1	12	Face Mars interesting controversial topic spar...	Face Mars interesting controversial sparked cu...	Fidelity Test	LIME	0.009718	0.001516	0.998484	Face Mars interesting controversial topic spar...	New Model - Word2Vec + XGBoost	0.011234	0.988766

Figure 6.20: Fidelity Results - Table

Fidelity testing for the validity of explanations compares each model on how well each model reflects the other model in making similar feature sets or decisions. The scores of the original model for the instance 12 changed from 22% to 53% for the human generated text. And predictions changed from 78% to 47% for AI or machine generated text. The new model's predictions with a more refined approach for instance 12 changes from 0% to 1% for human generated text. And predictions changes from 100% AI to 99% for AI generated text. This shows the models are not loyal to each other meaning the explanations for both models are different. There is large gap to fill to make these explanations in text classification more sense. No LIME explanation could be trusted on a generic model. More contextual references and fine tuning of the XAI tools and models are required to maybe reach a satisfactory level of confidence on these explanations.

CHAPTER 7

Comparing and Evaluating Results - Broader Discussion

The table is for 100 instances of the dataset where I have analysed different evaluation scores which indicates the variability in explanation quality. The are 6 metrics which are Sufficiency, Sensitivity, Sparsity , Jaccard Similarity , Stability (Cosine similarity) and KL-Divergence. The mean sufficiency scores are lower for these instances shows that explanations might have the limited ability to capture the actual feature set. However , stability and sensitivity supports the claim of upholding these explanation validity. In particular, explanations on human texts are more stable and robust with higher mean scores than the AI generated texts.

Subset	**Metric**	**Mean**	**Standard Deviation**	**Min**	**Max**	**AI**	**Human**
100 Instances	Sufficiency	0.233034	0.288323	0.000000	1.370418	0.195914	0.262201
100 Instances	Sensitivity	0.116517	0.144161	0.000000	0.685209	0.097957	0.131100
100 Instances	Sparsity	0.324991	0.127879	0.076923	0.600000	0.306170	0.339779
100 Instances	Jaccard Similarity	0.880000	0.324962	0.000000	1.000000	0.931818	0.839286
100 Instances	Stability	0.952852	0.107866	0.392659	1.000000	0.971363	0.938307
100 Instances	KL Divergence	0.087430	0.180985	0.000000	1.084596	0.053683	0.113946

Table 7.1: Table - Data Instances (100)

The table is for 1000 instances of the dataset where I have analysed different evaluation scores which indicates the variability in explanation quality. The are 6 metrics which are Sufficiency, Sensitivity, Sparsity , Jaccard Similarity , Stability (Cosine similarity) and KL-

Divergence. This table shows a little more consistency in results when compared to 100 instance table. This table suggests that increasing the data is leading to more stable and robust explanations. Here as well AI generated texts are less stable than Human generated texts in identifying true features.

Subset	**Metric**	**Mean**	**Standard Deviation**	**Min**	**Max**	**AI**	**Human**
1000 Instances	Sufficiency	0.203666	0.268538	0.000000	1.370418	0.214128	0.192513
1000 Instances	Sensitivity	0.101833	0.134269	0.000000	0.685209	0.107064	0.096257
1000 Instances	Sparsity	0.308668	0.131086	0.016575	0.600000	0.309246	0.308051
1000 Instances	Jaccard Similarity	0.881000	0.323789	0.000000	1.000000	0.883721	0.878099
1000 Instances	Stability	0.960838	0.092330	0.370676	1.000000	0.957434	0.964468
1000 Instances	KL Divergence	0.072430	0.152405	0.000000	1.121330	0.077910	0.066587

Table 7.2: Table - Data Instances (1000)

The table is for the full dataset instances where I have analysed different evaluation scores which indicates the variability in explanation quality. The are 6 metrics which are Sufficiency, Sensitivity, Sparsity , Jaccard Similarity , Stability (Cosine similarity) and KL-Divergence. This table reveals full and comprehensive picture of the XAI explanations. Everything in the overall picture is more or less the same as the rest of the subsets of the data. This full dataset table also confirms that AI generated text explanations are difficult to manage than human generated explanations. AI explanations needs more refinement to capture all the hidden features of the dataset.

Subset	**Metric**	**Mean**	**Standard Deviation**	**Min**	**Max**	**AI**	**Human**
Full Dataset	Sufficiency	0.191333	0.256691	0.000000	1.547917	0.210151	0.171148
Full Dataset	Sensitivity	0.095667	0.128346	0.000000	0.773959	0.105075	0.085574
Full Dataset	Sparsity	0.310469	0.128324	0.016575	0.600000	0.297215	0.324686
Full Dataset	Jaccard Similarity	0.890422	0.312363	0.000000	1.000000	0.848000	0.935927
Full Dataset	Stability	0.964809	0.084429	0.264945	1.000000	0.955904	0.974360
Full Dataset	KL Divergence	0.066076	0.142328	0.000000	1.761682	0.081105	0.049955

Table 7.3: Table - Data Instances (Full Data)

The analysis of my three subsets of the data (100 instances , 1000 instances , and complete dataset) reveals validity of explanations in a different light. Explanations are generally insufficient in a broader spectrum across all dataset and they fail to capture real features as shown by the sufficiency analysis but are more robust and stable while having small perturbation changes. The comparative analysis validates that AI explanations can’t fully replicate the human generated explanations interpretability. They need more thorough and careful examination to be trustworthy.

CHAPTER 8

Implications in Real World Scenarios

8.1 Broader Perturbation Testing: Synonym Replacement, Noise Injection, and Paraphrasing

In the real world scenarios there are not just vocabulary based text instances and perturbations but rather more complex ones. I just didn't do the vocabulary based testing but I have perturbed my data through phrases and noise in between the texts. This analysis checks the robustness of explanations of XAI in different setup. This section explores how well LIME explanations are suited to counter the perturbations such as synonym replacement , noise injection , and rephrasing of the text or paraphrasing.

8.1.1 Experimental Setup

The instance 12 of my text was chosen from my dataset as the baseline.The sentence is ”Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time”. Perturbations were applied in three forms:

- **Synonym Replacement**: Terms were replaced with synonyms, such as "Face" with "expression" and "topic" with "theme".
- **Noise Injection**: Random alphanumeric noise replaced portions of the text.

- **Paraphrasing**: The sentence was rephrased significantly, resulting in a minimal resemblance to the original sentence.

For each perturbed sentence, LIME explanations were generated, and prediction probabilities, key features, and Jaccard similarities between feature sets were analyzed.

8.1.2 Original Sentence

Prediction probabilities:

- Human: 0.22
- AI: 0.78

Key Features: {Face, topic, theories, debates, quite}

The explanation highlighted relevant terms, such as Face and topic, reflecting concepts central to the sentence.

8.1.3 Synonym Replacement

Prediction probabilities:

- Human: 0.43
- AI: 0.57

Key Features: {investigating, time, debates, Mar, theories}

Jaccard similarity with original features: 0.11

The replacement of Face with expression and topic with theme led to significant shifts in both prediction probabilities and important features. Only 11.1% of the features overlapped with the original set, indicating limited consistency in explanations.

8.1.4 Noise Injection

Prediction probabilities:

- Human: 0.50
- AI: 0.50

Key Features: {00R, 3e, 000ES3Ing, f, 03}

Jaccard similarity with original features: 0.0

The noise-injected sentence rendered explanations meaningless, with random strings such as 00R and 3e identified as significant. The model predictions dropped to complete uncertainty, further eroding explanation validity.

8.1.5 Paraphrased Sentence

Prediction probabilities:

- Human: 0.08
- AI: 0.92

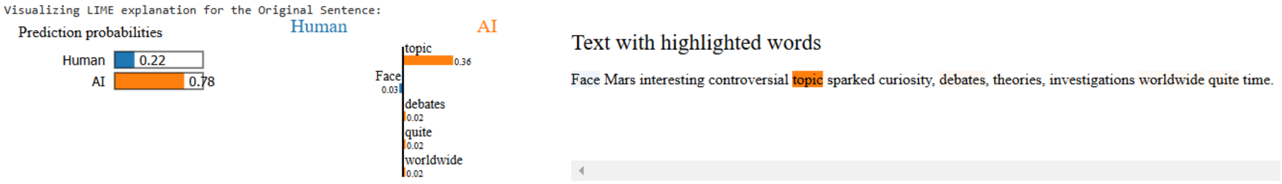
Key Features: {and, topic, Face, fascinating, theories}

Jaccard similarity with original features: 0.25

Paraphrasing resulted in minimal overlap with the original feature set, with the term "fascinating" identified as significant. The complete loss of meaningful explanations highlights the failure of LIME to adapt to significant linguistic variations.

8.1.6 Conclusion - Broader Perturbation Testing

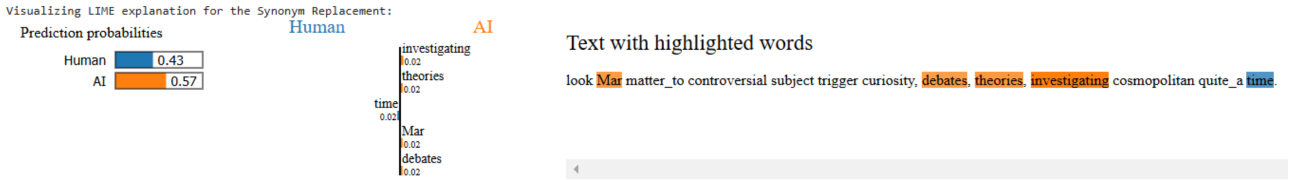
To validate the semantic consistency first I did the synonym replacement to check the XAI explanations. I used the instance 12 to demonstrate the explanation validity. The prediction probabilities for the human text changes from 22% to 43% and 78% to 57% in AI generated text. This raises few alarms about the explanation fidelity. This similarity score between the original and synonym replacement is 11.1% which shows a slight robust model.



```
Visualizing LIME explanation for the Original Sentence:
```

Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.

Figure 8.1: Synonym Perturbation - Original Instance



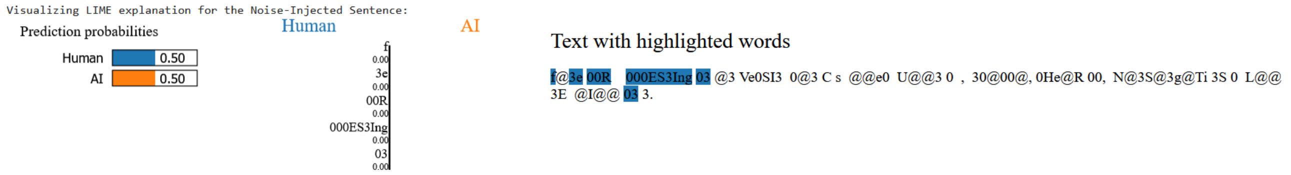
Visualizing LIME explanation for the Synonym Replacement:

Text with highlighted words

look Mar matter_to controversial subject trigger curiosity, debates, theories, investigating cosmopolitan quite_a time.

Figure 8.2: Synonym Perturbation - Changed Instance

I added noise injection similar to real world texts to reflect how noise infiltrated between texts. This makes the task for XAI to find the right explanation and features in the dataset a little more difficult. I added random characters and strings to demonstrate the noise injection. The probability changes for human generated text from 22% to 50% and 78% to 50% in AI generated text. In both cases the probability changes to 50% but a severe limitation is seen shows that LIME was unable to get a feature set. This also tells that in case of noisy data , which is usually the case in multilingual settings, quality of explanations can't be guaranteed.



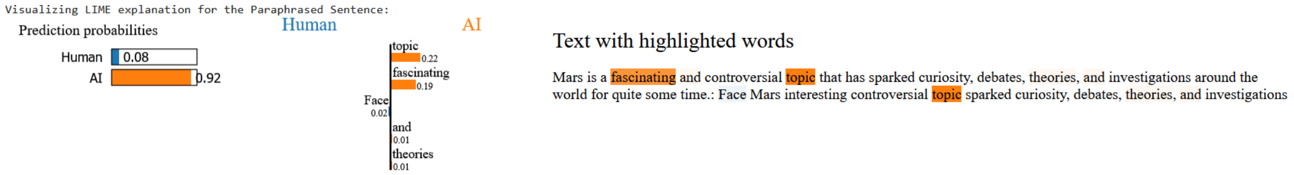
Visualizing LIME explanation for the Noise-Injected Sentence:

Text with highlighted words

f@3e 00R 000ES3Ing 03 @3 Ve0SI3 0@3 C s @@e0 U@@3 0 , 30@00@, 0He@R 00, N@3S@3g@Ti 3S 0 L@@ 3E @I@@ 03 3.

Figure 8.3: Fidelity Results - Instances

The third testing is thorough paraphrasing a sentence. Here , the prediction probabilities for the human text changes from 22% to 8% and 78% to 92% in AI generated text. This much shift in prediction probability reveals more shocking drawbacks of XAI's explanations. The explanations identify the same contextual text but with a different meaning. So the dynamic applicability of XAI is questioned yet again for text classifications.



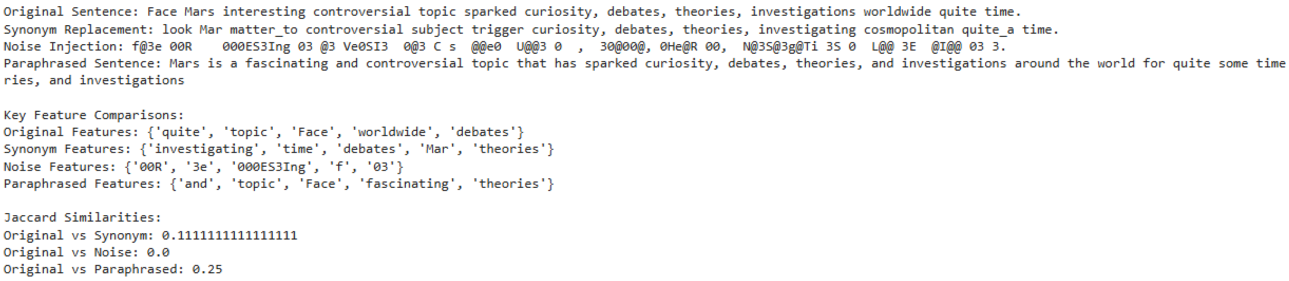
Visualizing LIME explanation for the Paraphrased Sentence:

Text with highlighted words

Mars is a fascinating and controversial topic that has sparked curiosity, debates, theories, and investigations around the world for quite some time.: Face Mars interesting controversial topic sparked curiosity, debates, theories, and investigations

Figure 8.4: Fidelity Results - Instances

My broader perturbation testing highlights multiple challenges for the XAI's tools and techniques and their explanations for text classification. First and foremost is the explanations are not similar at all even though the synonym are being replaced but the Jaccard similarity between the original and perturbed text is very low at 11%. Second is the complete breakdown of LIME when adding noise or paraphrasing the same text. This shows XAI's lack of robustness in real world applications.



```
Original Sentence: Face Mars interesting controversial topic sparked curiosity, debates, theories, investigations worldwide quite time.
Synonym Replacement: look Mar matter_to controversial subject trigger curiosity, debates, theories, investigating cosmopolitan quite_a time.
Noise Injection: f@3e 00R    000ES3Ing 03 @3 Ve0SI3  0@3 C s  @@e0  U@@3 0  ,  30@00@, 0He@R 00,  N@3S@3g@Ti 3S 0  L@@ 3E  @I@@ 03 3.
Paraphrased Sentence: Mars is a fascinating and controversial topic that has sparked curiosity, debates, theories, and investigations around the world for quite some time
ries, and investigations

Key Feature Comparisons:
Original Features: {'quite', 'topic', 'Face', 'worldwide', 'debates'}
Synonym Features: {'investigating', 'time', 'debates', 'Mar', 'theories'}
Noise Features: {'00R', '3e', '000ES3Ing', 'f', '03'}
Paraphrased Features: {'and', 'topic', 'Face', 'fascinating', 'theories'}

Jaccard Similarities:
Original vs Synonym: 0.1111111111111111
Original vs Noise: 0.0
Original vs Paraphrased: 0.25
```

Figure 8.5: Broader Perturbation Results

In multilingual texts where there are multiple variations and linguistic diversity, XAI's limitations shows a huge gap to be filled. My study of broader perturbations demonstrates the explanations generated by the XAI's tool (LIME) are very sensitive to noise injection , paraphrasing , and to even synonym replacement. Though the explanations may prevail for at least synonym replacement in terms of robustness but stability and similarity aspects shows that the explanations are totally compromised. In other perturbations , XAI totally collapses to even form a judgement let alone a good judgement of feature set. Linguistic and practical applications are unavoidable in real life and LIME is not trustworthy enough to make an effective decision.

8.2 XAI with Multilingual Text

8.2.1 Analysis of XAI Explanations in a Multilingual Setting and Their Validity

In real-world applications, data often comes from diverse sources with variations in vocabulary, grammar, and language. This introduces challenges in validating and interpreting machine learning models' outputs, especially in multilingual contexts. While many explainable AI (XAI) techniques, such as SHAP and LIME, have been extensively tested on English datasets, their applicability to multilingual settings remains underexplored. My analysis also evaluates the validity of XAI explanations in a multilingual text classification setting, specifically for Urdu, with a focus on removing the the most influential words in the dataset.

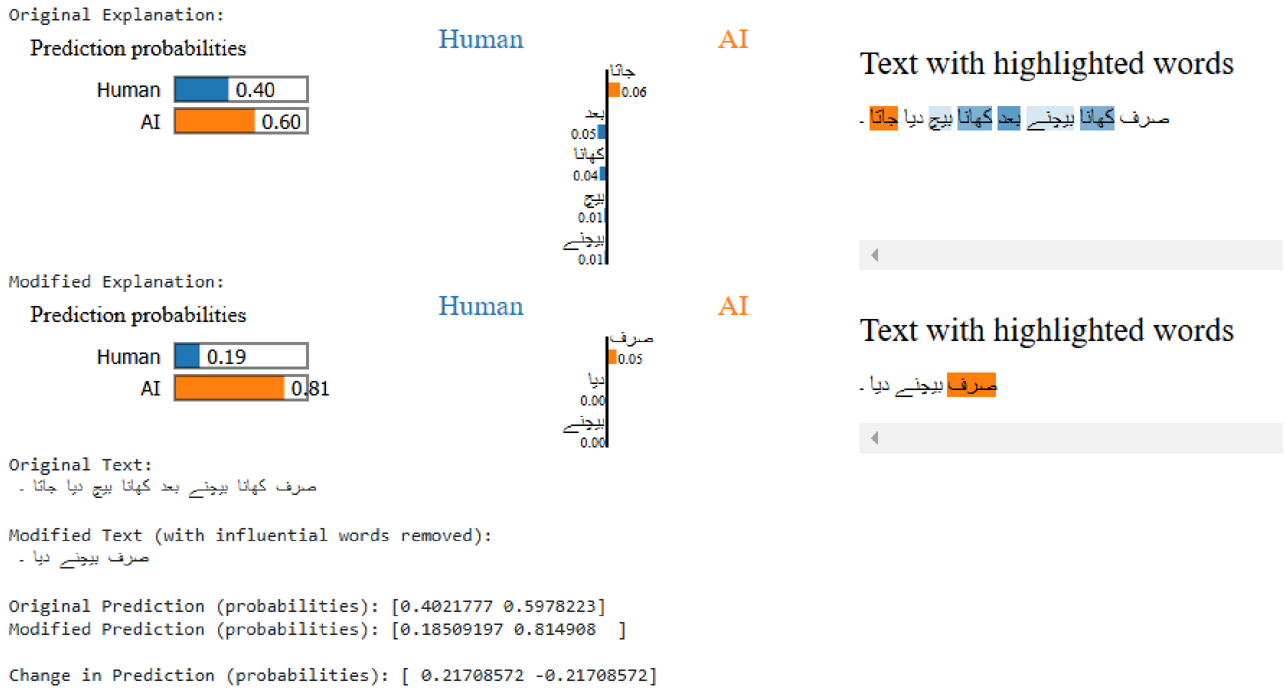
My multilingual dataset comprised text instances in Urdu, evaluated for both original and perturbed sentences. Perturbation techniques included removing top features of the instances. The goal was to assess how XAI explanations (LIME in this case) performed under these conditions and whether they provided valid and consistent insights.

This study provides evaluation included original explanation vs. perturbed explanation, and metrics like sufficiency, sensitivity, sparsity, Jaccard similarity, stability, and KL divergence for various subsets and the full dataset. Highlighting key terms identified by XAI as influential and comparing these terms across original and modified instances was the most significant part of the analysis.

8.2.2 Key Observations

Consistency of Explanations Across Perturbations

In actual terms of the original text and the perturbed text , the highlighted features change and their sentiment shifted as well. The word which has no impact in XAI's explanation before the perturbation suddenly showed the most impact on model decisions based on the XAI's evaluation.The probability for human which was 40% changed to 19% .The AI's probability changed from 60% to 81%.And this feature word 'only' in urdu resulted in that which was irrelevant before the perturbation.

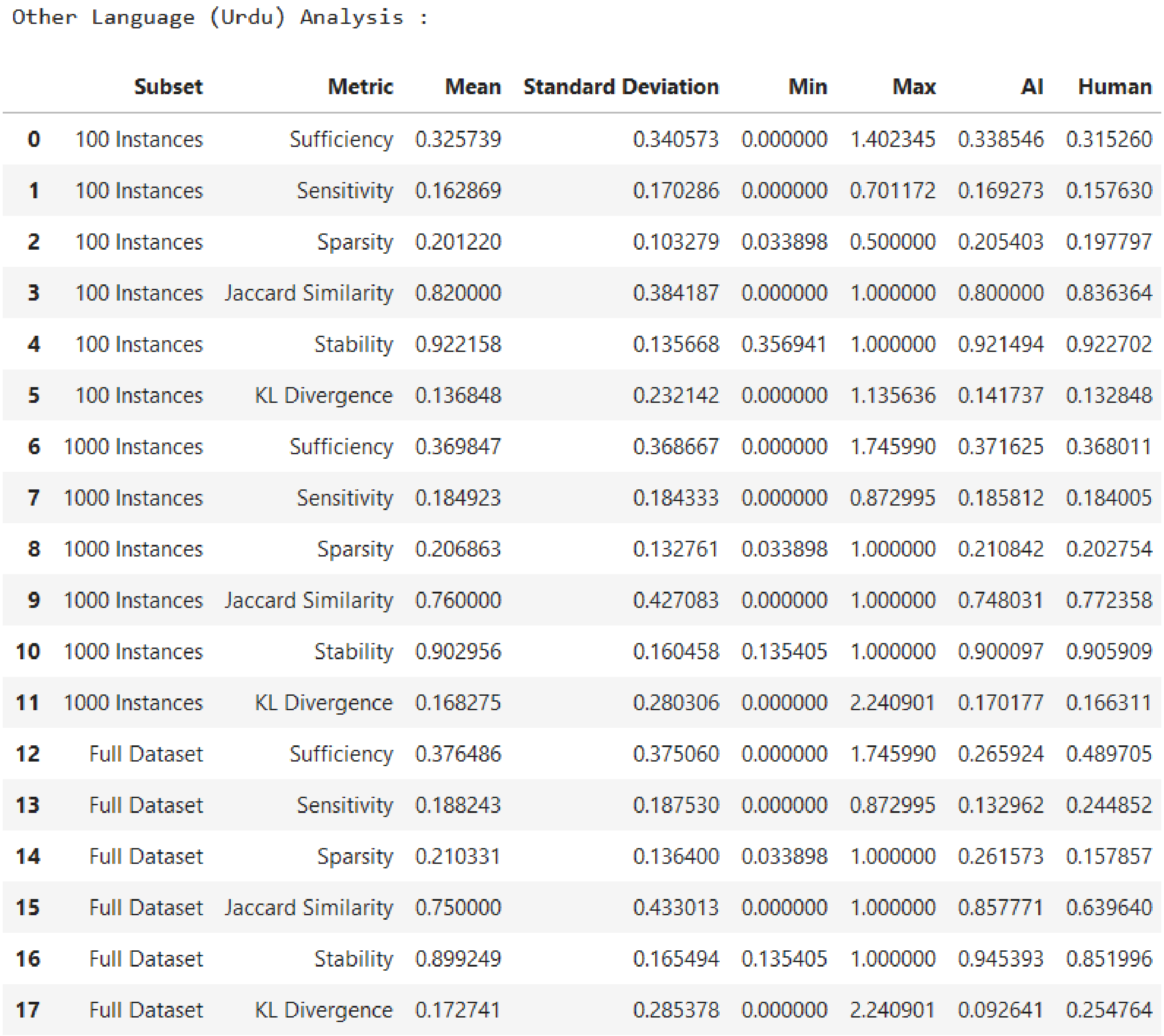


```
Original Explanation:

Modified Explanation:

Original Text:
 صرف کھانا بیچنے بعد کھانا بیچ دیا جاتا .

Modified Text (with influential words removed):
 صرف بیچنے دیا .

Original Prediction (probabilities): [0.4021777 0.5978223]
Modified Prediction (probabilities): [0.18509197 0.814908  ]

Change in Prediction (probabilities): [ 0.21708572 -0.21708572]
```

Figure 8.6: Multilingual Instance Based Results

Statistical Validation of Explanations

The statistical metrics provide deeper insights into the reliability of XAI explanations. **Sufficiency** and **Sensitivity** scores were extremely low, indicating that removing features identified by XAI had negligible impact on the model's predictions. This raises concerns about the faithfulness of explanations in accurately representing the model's decision-making. On the other hand, **Stability** scores were very high, consistently near **1.0** across the dataset, suggesting that the explanations were stable and consistent even when minor changes were made to the text. **KL Divergence** was minimal, suggesting that the distribution of predicted probabilities remained largely consistent despite perturbations, especially in the case of replacement. But this also made a point that model is not accurately identifying features important to make a decision.



```
Other Language (Urdu) Analysis :
```

	Subset	Metric	Mean	Standard Deviation	Min	Max	AI	Human
0	100 Instances	Sufficiency	0.325739	0.340573	0.000000	1.402345	0.338546	0.315260
1	100 Instances	Sensitivity	0.162869	0.170286	0.000000	0.701172	0.169273	0.157630
2	100 Instances	Sparsity	0.201220	0.103279	0.033898	0.500000	0.205403	0.197797
3	100 Instances	Jaccard Similarity	0.820000	0.384187	0.000000	1.000000	0.800000	0.836364
4	100 Instances	Stability	0.922158	0.135668	0.356941	1.000000	0.921494	0.922702
5	100 Instances	KL Divergence	0.136848	0.232142	0.000000	1.135636	0.141737	0.132848
6	1000 Instances	Sufficiency	0.369847	0.368667	0.000000	1.745990	0.371625	0.368011
7	1000 Instances	Sensitivity	0.184923	0.184333	0.000000	0.872995	0.185812	0.184005
8	1000 Instances	Sparsity	0.206863	0.132761	0.033898	1.000000	0.210842	0.202754
9	1000 Instances	Jaccard Similarity	0.760000	0.427083	0.000000	1.000000	0.748031	0.772358
10	1000 Instances	Stability	0.902956	0.160458	0.135405	1.000000	0.900097	0.905909
11	1000 Instances	KL Divergence	0.168275	0.280306	0.000000	2.240901	0.170177	0.166311
12	Full Dataset	Sufficiency	0.376486	0.375060	0.000000	1.745990	0.265924	0.489705
13	Full Dataset	Sensitivity	0.188243	0.187530	0.000000	0.872995	0.132962	0.244852
14	Full Dataset	Sparsity	0.210331	0.136400	0.033898	1.000000	0.261573	0.157857
15	Full Dataset	Jaccard Similarity	0.750000	0.433013	0.000000	1.000000	0.857771	0.639640
16	Full Dataset	Stability	0.899249	0.165494	0.135405	1.000000	0.945393	0.851996
17	Full Dataset	KL Divergence	0.172741	0.285378	0.000000	2.240901	0.092641	0.254764

Figure 8.7: Table - Multilingual Evaluations - Urdu

Real-World Relevance in Multilingual Data

In practical scenarios, data is rarely clean or strictly in English. Texts may contain synonyms, paraphrases, or even noise due to typos, OCR errors, or colloquial usage. This analysis shows that XAI techniques like LIME can't handle minor semantic variations (e.g., synonyms or replacement of the top impactful words) well, making explanation validity not reasonable .

8.2.3 Implications for XAI Validity in Multilingual Settings

Faithfulness

For XAI explanations to be trustworthy, they must accurately reflect how the model makes decisions. In cases where changes involve synonyms or similar phrases, explanations generally maintain their meaning, showing faithfulness.But I also found out that XAI's makes the wrong assumptions about the feature set quite often. Also, when irrelevant changes or noise are introduced, the explanations often lose validity. This suggests that high-quality input data is essential to ensure reliable and faithful explanations in real-world applications.

Consistency

The study found that XAI explanations are consistent across similar inputs, even when small language variations, such as synonyms or minor rephrasing, are present. This reliability makes them suitable for repetitive tasks, particularly in multilingual settings, where slight linguistic differences often occur. Consistent explanations mean that users can trust the system to handle similar cases in predictable ways.

Robustness

XAI methods like LIME instantly switch to different explanation, such as replacement with a different word or altogether removing it. To add on it, significant alterations, such as noisy or irrelevant data, can weaken their validity. This highlights the need for models and methods that can better handle unstructured and noisy data, especially when working with less standardized languages.

Human Interpretability

In some cases, such as explanations in Urdu, the model correctly identified important features like ("only") and ("food") as key drivers for its predictions. This shows that the explanations can be clear and meaningful to human users, which is essential for building trust. Users should be able to understand and verify the logic behind the model's predictions.But these features were important in one instance and irrelevant on the same instance after minor changes.

8.2.4 Challenges and Recommendations

Handling Noise

When input data contains noise or irrelevant information, the quality of XAI explanations drops significantly. This calls for better preprocessing methods or models designed to handle noisy data. For languages with informal or colloquial use, maintaining data quality is critical to producing valid explanations.

Domain-Specific Adaptations

Different languages have unique features, such as grammar, word structure, and meaning. XAI techniques need to account for these differences to provide accurate explanations. Multilingual models trained on diverse datasets are better equipped to handle these challenges and can adapt explanations to fit the specific needs of each language.

Integration of Contextual Embeddings

Advanced embeddings like BERT or multilingual T5 can improve XAI explanations by capturing the deeper relationships between words and phrases. These embeddings help models understand synonyms, paraphrases, and complex structures, making explanations more robust and accurate in multilingual scenarios.

Evaluation Frameworks

The study highlights the importance of standardized checks for validating XAI explanations in multilingual tasks. Metrics like stability, sufficiency, sparsity, and Jaccard similarity should be part of regular evaluation practices. These frameworks can ensure that explanations are valid, reliable, and useful across different languages and tasks.

8.3 Conclusion - Multilingual Context

This analysis shows how important it is to check if XAI explanations work well or don't in multilingual settings. Methods like LIME can't be trusted to do a good job on a generic model used to drive explanations. These XAI tools have trouble providing clear explanations when the data includes noise or irrelevant details in these scenarios. In real-world scenarios, where data is often messy and comes in many languages, this highlights the need for more reliable and flexible XAI methods. Improving these techniques will ensure that explanations stay accurate, consistent, and easy to understand, helping build trust in AI systems used in different languages and settings.

CHAPTER 9

Conclusion and Reflections

This dissertation explored the validity of Explainable AI (XAI) for text classifications or in multilingual text classification setup. This study provides a complete and comprehensiveness analysis of the XAI tools (more specifically LIME). Through a multifaceted approach this study provides a robust , faithful , and interpretable explanations to predict texts. Leveraging the metrics such as sufficiency , sparsity , sensitivity , stability , similarity , and proxy fidelity to reveal significant insights into the strengths and limitations of XAI methods.

The sufficiency analysis identifies the explanations generated by XAI are often incomplete , which can be seen from the low sufficiency scores from all subsets of the data testing. One instance shows a moderate sufficiency while the other gives a very low sufficiency score which gives that existing XAI techniques methods usually fail to capture the actual complexity of the multilingual text in classification problems.

Sparsity analysis in my study indicates that explanations are sparse and are relatively stable in generic model development. Explanations are also rigid in a sense that they focus constantly on a limited sets of features. This reinforces the claims that main features in the explanations are mostly consistent. Nevertheless, the narrow focus on a small feature set also raises concerns about the random and absurd feature selection highlighted by the explanations. This undermines the interpretability of the explanations which can be seen from the wrong and irrelevant terms or words focused by the XAI.

Stability assessments are generated through Jaccard and cosine similarity which reveal mixed results. Cosine similarity results show much variations when changes or perturbations are done on the instances. This results leads to key features identified by the explanations particularly in AI (Machine) generated texts . These XAI explanations changes with every minor input change reflects unreliability.

Counterfactual analysis highlights an astonishing results of the XAI which were never seen before. Both Instances 12 and 110 used in the study demonstrate even major changes which actually shifts the sentiment of the text leads to very small or negligible changes in the model's predictions. This lack of sensitivity identifies that explanations provided my the XAI tools are not always meaningful. To add on it , the inability of the explanations to actually adapt to different type of texts or contexts shows that XAI methods fall really short in ensuring accurate results.

Sensitivty analysis in which KL-divergence and sensitivity through cosine similarity tells that explanations don't capture the small instances really well. These subtle changes are misclassified or altogether ignored by the model leading to gaps in the explanations trustworthiness.

Fidelity testing shows that fidelity scores between the original model (XGBoost + TF-IDF) when compared to the new model (XGboost + word2vec) is very low. The tells us that surrogate model is not loyal to the original model in identifying the same features or similar feature set in the data. Both model show different outputs of probabilities for AI and human generated texts.

Broader perturbation testing confirms one thing in my study that explanations for human generated texts are relatively stable , more robust , and more interpretable than AI generated abstracts in the data. This thorough research identified that while XAI provide real and genuine insights for model decision making , their reliability can vary drastically and depends on the type of the perturbation applied as the input. As soon the complexity of the perturbations are increased such as paraphrasing and noise injection the explanations usually fails to find the true semantic meaning, linguistic and contextual references in the text data. This thing is reinforced when a different language was incorporated to test the explanations. Using 'Urdu' which is a non-Latin language makes the case of supporting XAI for text classifications even worse by revealing that feature identification is random for non-English datasets.

In conclusion , while this research validates some really good and critical aspects of the XAI explanation validity, greater challenges still remain to achieve a balanced , reliable , transparent and trustworthy explanations of XAI. Additional refinement in the modelling process and XAI tools are needed to make use of the explanations generated in multilingual text classification scenarios.

9.1 Future Work

My future work would be to focus on expanding the XAI techniques to more multilingual text classification models , like LRP (Layer-wise Relevance Propagation) and IGs (Integrated Gradients). Furthermore , incorporating the contextual references of the languages and adding domain specific knowledge into feature selection process might yield better results. User feedbacks and different interpretability of human explanations could give me deeper insights into the classification of Machine vs Human texts. Using more languages and cultural contexts may provide rich theories and analysis of the XAI explanations.

APPENDIX A

Access To Complete Code For Dissertation

Proposal , source code , complete dataset (csv) , and word2vec embedding for this dissertation is shared on Dropbox link below. All aspects of my dissertation including pre-processing of the data , data splitting , model implementation , and validity of explanations of XAI can be reproduced. The readers interested in the methods and techniques of XAI can analyse the and access the code used in this research.

The repository includes:

- The preprocessing step and EDA.
- Model Implementation - Data splitting and Xgboost.
- Validity of explanations of XAI.
- Tables and Visualizations developed through the process.
- Interpretation and Analysis.

The code can be accessed using the link below: Dropbox Link Code

Main Dataset used in the research link: Kaggle Link

Multilingual Dataset link: HuggingFace Link

Word2Vec embedding link: Word2Vec Link

By sharing this source code, I aim to facilitate further research and encourage transparency in the field of AI and Explainable (AI).

APPENDIX B

Access To Approved Proposal for Dissertation

Proposal for the research is shared on this link: DropBox Proposal Link

The proposal contains the following:

- Introduction
- Critical Context
- Methodology and Approach
- Work Plan
- Risks
- References

This dissertation is in line with the guidelines and structure discussed in the proposal.

Bibliography

[1] L. CoroamÄ and A. Groza. *Evaluation Metrics in Explainable Artificial Intelligence (XAI)*, pages 401–413. Springer, 2022.

[2] Z. C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(12):36–43, 2018.

[3] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4765–4774, 2017.

[4] S. McAleese and M. Keane. A comparative analysis of counterfactual explanation methods for text classifiers, 2024.

[5] S. Mohseni, N. Zarei, and E. D. Ragan. Evaluating the faithfulness of explanations generated by explainable ai models. *OpenReview*, 2020.

[6] S. Mohseni, N. Zarei, and E. D. Ragan. A multidisciplinary survey and framework for explainable ai. *ACM Computing Surveys (CSUR)*, 53(2):1–35, 2020.

[7] J. Ribeiro, L. Cardoso, V. Santos, E. Carvalho, N. Carneiro, and R. Alves. How reliable and stable are explanations of xai methods?, 2024.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin. “why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[9] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling lime and shap: Adversarial attacks on post-hoc explanation methods. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12. ACM, 2020.