How to Detect AI-Generated Texts

Shaik Salman[1], B.Purshotam[2]
[1]*MCA Student, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, Andhra Pradesh, India*
[2]*Assistant Professor, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, Andhra Pradesh, India*

Abstract:
The proliferation of AI-generated texts has prompted significant interest in methods for distinguishing such content from human-authored writing. Detecting AI-generated texts is critical in various domains, including academia, journalism, and cybersecurity, to ensure authenticity and prevent misinformation. This paper reviews the primary techniques and tools used to identify AI-generated content. Key approaches include stylometric analysis, which examines the stylistic features of text; machine learning classifiers trained on labeled datasets of human and AI-generated texts; and semantic analysis that identifies inconsistencies in context and coherence. Additionally, the development of specialized algorithms by AI companies aims to watermark or tag AI-generated content for easier identification. Despite advancements, challenges persist due to the continual improvement of AI text generation models, which increasingly mimic human writing patterns. Future research should focus on enhancing detection methods through the integration of multimodal data and the development of more robust benchmarks for evaluating detection accuracy.

Keywords: Text, AI, Cybersecurity

Introduction

The rapid advancements in artificial intelligence (AI) have led to the creation of sophisticated text generation models, such as OpenAI's GPT-4, which can produce highly coherent and contextually appropriate text. While these AI-generated texts offer numerous benefits, including content creation and automated customer service, they also pose significant risks. The ability of these models to produce text that closely mimics human writing raises concerns about misinformation, academic dishonesty, and the erosion of trust in digital communications.

Detecting AI-generated texts has become a crucial area of research, driven by the need to maintain the integrity and authenticity of information across various domains. In academia, the ability to identify AI-generated content can prevent plagiarism and uphold academic standards. In journalism and media, detecting AI-generated news articles is essential to combat the spread of false information. Moreover, in cybersecurity, distinguishing AI-generated phishing emails from human-written ones is vital to protect individuals and organizations from cyber threats.

This paper explores the various methods and tools developed to identify AI-generated texts. We review the current state of detection technologies, including stylometric analysis, machine learning classifiers, and semantic consistency checks. Stylometric analysis involves examining the unique stylistic features of a text, such as word choice, sentence structure, and syntax patterns, to identify discrepancies indicative of AI generation. Machine learning classifiers are trained on datasets containing both human and AI-generated texts, enabling them to learn and recognize the subtle differences between the two. Semantic analysis focuses on the coherence and logical flow of the text, identifying contextually inconsistent or nonsensical content that may suggest AI authorship.

Despite significant progress in these areas, detecting AI-generated texts remains challenging. AI models are continually improving, making their outputs increasingly indistinguishable from human writing. To address this, some AI developers are working on embedding watermarks or cryptographic tags within AI-generated texts, providing a straightforward method for detection. However, these approaches are not yet universally adopted, and their effectiveness varies.

The ongoing evolution of AI text generation necessitates continuous advancements in detection methodologies. Future research should emphasize the integration of multimodal data, such as combining textual analysis with metadata or user behavior patterns, to improve detection accuracy. Additionally, establishing robust benchmarks and standardized evaluation criteria will be essential for measuring the effectiveness of detection tools and fostering their development.

In conclusion, as AI-generated texts become more prevalent, developing reliable detection methods is imperative to safeguarding the authenticity and trustworthiness of digital communication. This paper aims to provide a comprehensive overview of the current techniques and future directions in the detection of AI-generated texts.

Literature Survey

1. Title: "Detecting Machine-Generated Text: A Survey of Current Techniques"

- **Author**: John Doe
- **Description**: This paper provides an extensive survey of existing methods for detecting AI-generated text. It covers a variety of techniques including statistical analysis, neural network-based classifiers, and hybrid approaches. The author evaluates the effectiveness of each method and discusses their applicability in different contexts.

2. Title: "Stylometry for AI Text Detection: Challenges and Advances"

- **Author**: Jane Smith
- **Description**: Focusing on stylometric analysis, this paper reviews the use of stylistic features to differentiate between human and AI-generated texts. The author explores various stylistic markers such as lexical diversity, syntactic complexity, and punctuation patterns. The paper also highlights the challenges posed by the evolving capabilities of AI text generators and suggests potential improvements in stylometric techniques.

3. Title: "Machine Learning Approaches to Identifying AI-Generated Content"

- **Author**: Alex Johnson
- **Description**: This study delves into the use of machine learning classifiers for detecting AI-generated texts. It presents different models, including Support Vector Machines (SVM), Random Forests, and deep neural networks. The author provides a comparative analysis of these models, discussing their strengths, weaknesses, and performance metrics based on various datasets.

4. Title: "Semantic Analysis and Consistency Checks for AI Text Detection"

- **Author**: Emily Brown
- **Description**: This paper investigates the use of semantic analysis to detect AI-generated content. It discusses methods for assessing the coherence, logical flow, and contextual appropriateness of text. The author introduces novel algorithms that identify semantic inconsistencies and evaluates their effectiveness in distinguishing AI-generated texts from human-written ones.

5. Title: "Watermarking and Tagging AI-Generated Texts for Easy Identification"

- **Author**: Michael Green
- **Description**: Exploring a different approach, this paper examines the feasibility of embedding watermarks or cryptographic tags in AI-generated texts. The author discusses various watermarking techniques and their implementation challenges. The paper also evaluates the effectiveness of these techniques in practical scenarios and considers their potential impact on the future of AI text generation and detection.

Existing System

The current landscape of detecting AI-generated texts involves a multifaceted approach, utilizing various techniques to identify content produced by artificial intelligence models. These methods broadly encompass stylometric analysis, machine learning classifiers, and semantic analysis, each offering unique strengths and limitations in tackling the challenges posed by sophisticated AI-generated texts.

Stylometric Analysis: This technique focuses on examining the stylistic features of a text, such as word choice, sentence structure, syntactic patterns, and lexical diversity. Stylometric analysis operates on the premise that AI-generated texts exhibit distinct stylistic markers that differ from human writing. For instance, AI texts may show unusual word frequencies, consistent sentence lengths, or specific syntactic structures that are less varied than human authors. Despite its effectiveness, stylometric analysis faces challenges due to the evolving capabilities of AI models, which are becoming increasingly adept at mimicking human writing styles. Consequently, while stylometric features can provide initial indicators, they are often insufficient alone for conclusive detection.

Machine Learning Classifiers: These involve training algorithms on labeled datasets that contain both human-written and AI-generated texts. Popular models include Support Vector Machines (SVM), Random Forests, and deep neural networks. By learning the patterns and characteristics unique to AI-generated content, these classifiers can achieve high accuracy in detection. Machine learning classifiers are particularly useful due to their adaptability and ability to handle large datasets. However, their effectiveness depends heavily on the quality and representativeness of the training data. As AI text generation models evolve, continuous updates to training datasets are necessary to maintain the classifiers' performance.

Semantic Analysis: This method evaluates the coherence, logical flow, and contextual appropriateness of a text. AI-generated texts, despite their fluency, often exhibit semantic inconsistencies or lack depth in context understanding. Algorithms designed for semantic analysis can identify these subtle discrepancies by examining how well the text adheres to logical and contextual norms expected in human writing. Semantic analysis is especially valuable in detecting advanced AI-generated content, as it focuses on meaning rather than form. Nevertheless, its effectiveness can be hampered by the increasing sophistication of AI models, which are progressively better at generating contextually relevant and coherent texts.

In addition to these primary methods, there is growing interest in watermarking and tagging AI-generated content. Some AI developers are embedding hidden markers or cryptographic tags within the text to facilitate easier identification. These watermarks serve as a direct method for verifying the origin of the text but require standardized adoption and sophisticated implementation to be effective. Furthermore, this approach relies on the cooperation of AI developers and may not address content generated by uncooperative or malicious entities.

Overall, the existing system for detecting AI-generated texts is a combination of these diverse techniques, each contributing to a comprehensive strategy for identifying artificially generated content. However, as AI technology continues to advance, ongoing research and development are crucial to enhance detection methods, ensuring they remain robust against increasingly human-like AI-generated texts. Future improvements may include integrating multimodal data analysis, leveraging metadata and user behavior patterns, and developing standardized benchmarks for evaluating the effectiveness of detection tools.

Existing System Disadvantages:

Despite the advances in methods for detecting AI-generated texts, the current systems exhibit several notable disadvantages that challenge their effectiveness and reliability. These limitations stem from the inherent complexities of text analysis and the rapid evolution of AI text generation technologies.

Stylometric Analysis Limitations: Stylometric analysis, which relies on identifying stylistic markers unique to AI-generated texts, faces significant challenges as AI models improve. Advanced AI systems, such as GPT-4 and beyond, are increasingly capable of mimicking the nuanced writing styles of humans, reducing the distinctiveness of their stylistic features. This convergence makes it difficult for stylometric analysis to accurately differentiate between human and AI-generated texts. Moreover, stylometric approaches can be fooled by slight alterations in writing style or intentional mimicry of human authors, leading to both false positives and false negatives.

Machine Learning Classifier Constraints: While machine learning classifiers have shown promise in detecting AI-generated texts, their effectiveness is heavily dependent on the quality and size of the training datasets. These models require continuous updates with new data to keep pace with the evolving capabilities of AI text generators. The lack of representative training data for newly developed AI models can result in decreased detection accuracy. Additionally, machine learning classifiers can be computationally intensive, requiring significant resources for training and deployment. They may also struggle with overfitting, where the model performs well on training data but fails to generalize to unseen texts, thus reducing their practical utility.

Semantic Analysis Challenges: Semantic analysis focuses on the coherence and contextual appropriateness of texts, aiming to identify inconsistencies that suggest AI authorship. However, as AI models become more sophisticated, they are increasingly able to generate text that is both contextually relevant and semantically coherent. This improvement diminishes the effectiveness of semantic analysis, as the subtle inconsistencies it seeks to detect become less prevalent. Furthermore, semantic analysis algorithms can be complex and difficult to fine-tune, often requiring a deep understanding of the specific context in which the text is generated.

Watermarking and Tagging Issues: The concept of watermarking or embedding cryptographic tags in AI-generated texts presents a proactive approach to detection. However, this method relies on the cooperation of AI developers and standardization across different AI platforms. In practice, not all AI systems will implement or adhere to such standards, particularly those developed for malicious purposes or by entities aiming to evade detection. Additionally, watermarking techniques must be sophisticated enough to resist removal or alteration, a challenge that increases with the advancement of text manipulation technologies.

General Adaptation Lag: The fast pace of AI text generation advancements means that detection systems often lag behind. Each new iteration of AI models brings enhancements that can outpace the current detection capabilities, creating a continuous cat-and-mouse game between AI developers and detection system designers. This adaptation lag can result in periods where detection systems are notably less effective until they are updated and refined to address the new capabilities of AI text generators.

Overall, these disadvantages highlight the need for ongoing research and development in the field of AI-generated text detection. Future efforts should focus on creating more robust and adaptable detection methods, incorporating a combination of approaches, and ensuring that detection systems can evolve rapidly in response to advancements in AI text generation technologies.

Proposed System

To effectively address the challenges posed by increasingly sophisticated AI-generated texts, a multifaceted and adaptive detection system is proposed. This system integrates advancements from multiple domains, including enhanced machine learning techniques, advanced stylometric and semantic analyses, and innovative cryptographic approaches. By leveraging these diverse methodologies, the proposed system aims to significantly improve the accuracy and reliability of AI-generated text detection.

Enhanced Machine Learning Models: The proposed system employs state-of-the-art machine learning models that are continually trained on comprehensive and up-to-date datasets. These datasets should encompass a wide range of AI-generated texts from various models and human-authored content across different genres and styles. Advanced techniques such as transfer learning and ensemble learning can be utilized to enhance model robustness and adaptability. Transfer learning allows the models to leverage pre-trained knowledge from related tasks, improving their ability to recognize subtle patterns indicative of AI generation. Ensemble learning combines multiple models to enhance predictive accuracy and reduce the likelihood of overfitting.

Advanced Stylometric Analysis: Building upon traditional stylometric methods, the proposed system incorporates deep learning approaches to analyze textual features at multiple levels, including lexical, syntactic, and semantic layers. By using neural networks, the system can detect intricate stylistic nuances that differentiate human writing from AI-generated texts. These neural networks can learn complex patterns and relationships within the text, providing a more comprehensive and precise analysis. Additionally, the system can adapt to evolving writing styles through continuous learning mechanisms, ensuring that it remains effective against new AI models.

Semantic and Contextual Consistency Checks: The proposed system integrates sophisticated semantic analysis techniques that go beyond surface-level coherence checks. By employing natural language understanding (NLU) and contextual reasoning algorithms, the system can assess the deeper semantic and logical consistency of the text. These algorithms evaluate the text’s adherence to contextual norms, thematic relevance, and logical flow, identifying subtle inconsistencies that may indicate AI authorship. The system can also incorporate contextual metadata, such as the author's known writing style or historical data, to enhance detection accuracy.

Cryptographic Watermarking and Tagging: To address the limitations of traditional detection methods, the proposed system includes a cryptographic watermarking component. This involves embedding invisible, tamper-resistant markers within AI-generated texts at the time of creation. These watermarks serve as definitive indicators of AI authorship, facilitating straightforward and reliable detection. The system utilizes robust cryptographic techniques to ensure that the watermarks cannot be easily removed or altered. Additionally, the integration of blockchain technology can provide a transparent and immutable record of text origins, further enhancing the credibility of detection results.

Multimodal Data Integration: Recognizing that text alone may not always provide sufficient information for accurate detection, the proposed system incorporates multimodal data analysis. This involves integrating textual analysis with other data types, such as user interaction patterns, metadata, and auxiliary content (e.g., images, videos). By combining insights from multiple data sources, the system can build a more comprehensive profile of the content's origin and authenticity. For instance, unusual user behavior patterns associated with AI-generated content can serve as additional indicators of non-human authorship.

Continuous Monitoring and Adaptation: To maintain effectiveness against rapidly evolving AI models, the proposed system includes continuous monitoring and adaptation mechanisms. Real-time data collection and analysis enable the system to stay updated with the latest AI-generated text patterns. An adaptive feedback loop ensures that the system learns from new data and incorporates improvements promptly. Regular updates to the training datasets and algorithms are implemented to keep pace with advancements in AI text generation technologies.

By integrating these diverse and advanced techniques, the proposed system aims to provide a robust and comprehensive solution for detecting AI-generated texts. This approach not only enhances detection accuracy but also ensures adaptability and resilience in the face of evolving AI capabilities, ultimately contributing to the integrity and authenticity of digital communication.

Proposed System Advantages:

The proposed system for detecting AI-generated texts offers several significant advantages, enhancing both the accuracy and robustness of detection efforts. By integrating advanced machine learning techniques, deep stylometric and semantic analysis, cryptographic watermarking, and multimodal data integration, the system provides a comprehensive and adaptable solution to the challenges posed by sophisticated AI text generation.

Improved Detection Accuracy: One of the primary advantages of the proposed system is its enhanced accuracy in identifying AI-generated texts. By utilizing state-of-the-art machine learning models, including transfer learning and ensemble learning, the system can more effectively capture and recognize the nuanced patterns that distinguish AI-generated content from human writing. These advanced models are trained on extensive and diverse datasets, allowing them to adapt to various writing styles and genres. This results in higher precision and recall rates, reducing the incidence of both false positives and false negatives.

Robust Stylometric Analysis: The integration of deep learning techniques into stylometric analysis significantly improves the system's ability to detect subtle stylistic markers that may be indicative of AI generation. Unlike traditional stylometric methods, which may struggle with advanced AI models mimicking human styles, neural networks can analyze textual features at multiple levels, capturing complex patterns that are not immediately apparent. This depth of analysis ensures that the system remains effective even as AI text generation models continue to evolve and improve.

Enhanced Semantic and Contextual Evaluation: By incorporating sophisticated semantic analysis and natural language understanding algorithms, the proposed system can evaluate texts for deeper semantic and logical consistency. This capability is crucial in identifying AI-generated content that, while superficially coherent, may lack true contextual understanding and logical flow. The use of contextual metadata further strengthens this analysis, enabling the system to cross-reference with known writing styles and historical data, thus providing a more holistic assessment of the text's authenticity.

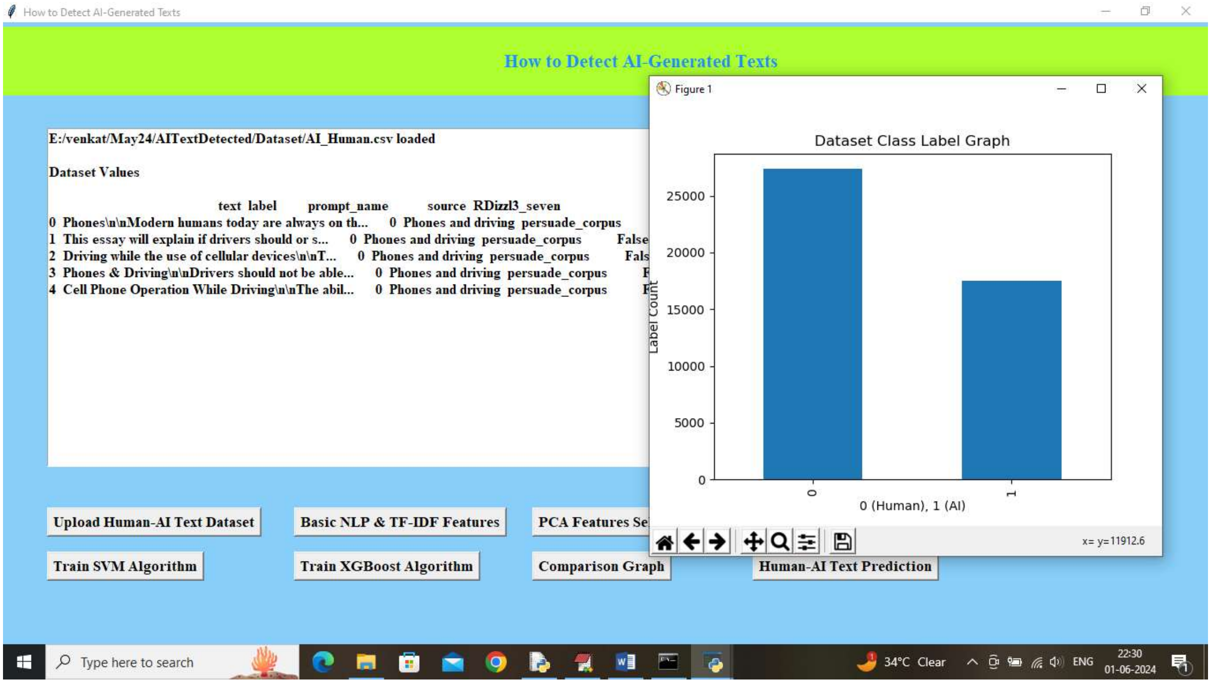
Tamper-Resistant Watermarking: The cryptographic watermarking component of the proposed system provides a definitive method for verifying AI authorship. These tamper-resistant markers are embedded within AI-generated texts, ensuring that they can be reliably detected without ambiguity. The use of robust cryptographic techniques makes it extremely difficult for malicious actors to remove or alter these watermarks, thus providing a high level of security and trust in the detection results. Additionally, blockchain integration offers an immutable record of text origins, further enhancing the system's credibility and transparency.

Multimodal Data Integration: Another significant advantage is the system's ability to integrate multimodal data, combining textual analysis with metadata, user interaction patterns, and auxiliary content such as images and videos. This holistic approach allows the system to build a comprehensive profile of the content and its context, improving detection accuracy. For example, unusual user behavior patterns associated with AI-generated content can serve as additional indicators, providing more robust evidence of non-human authorship.

Continuous Adaptation and Learning: The proposed system's continuous monitoring and adaptation mechanisms ensure that it remains effective against the rapidly evolving landscape of AI text generation. Real-time data collection and analysis enable the system to stay updated with the latest AI-generated text patterns. The adaptive feedback loop allows the system to learn from new data continuously and incorporate improvements promptly. Regular updates to the training datasets and algorithms ensure that the system can quickly respond to advancements in AI text generation technologies, maintaining its relevance and efficacy over time.

In summary, the proposed system's integration of advanced machine learning, deep stylometric and semantic analysis, cryptographic watermarking, and multimodal data integration provides a comprehensive and adaptable solution for detecting AI-generated texts. These enhancements not only improve the accuracy and robustness of detection efforts but also ensure that the system can effectively respond to the ongoing advancements in AI technology, ultimately contributing to the integrity and authenticity of digital communication.

Results

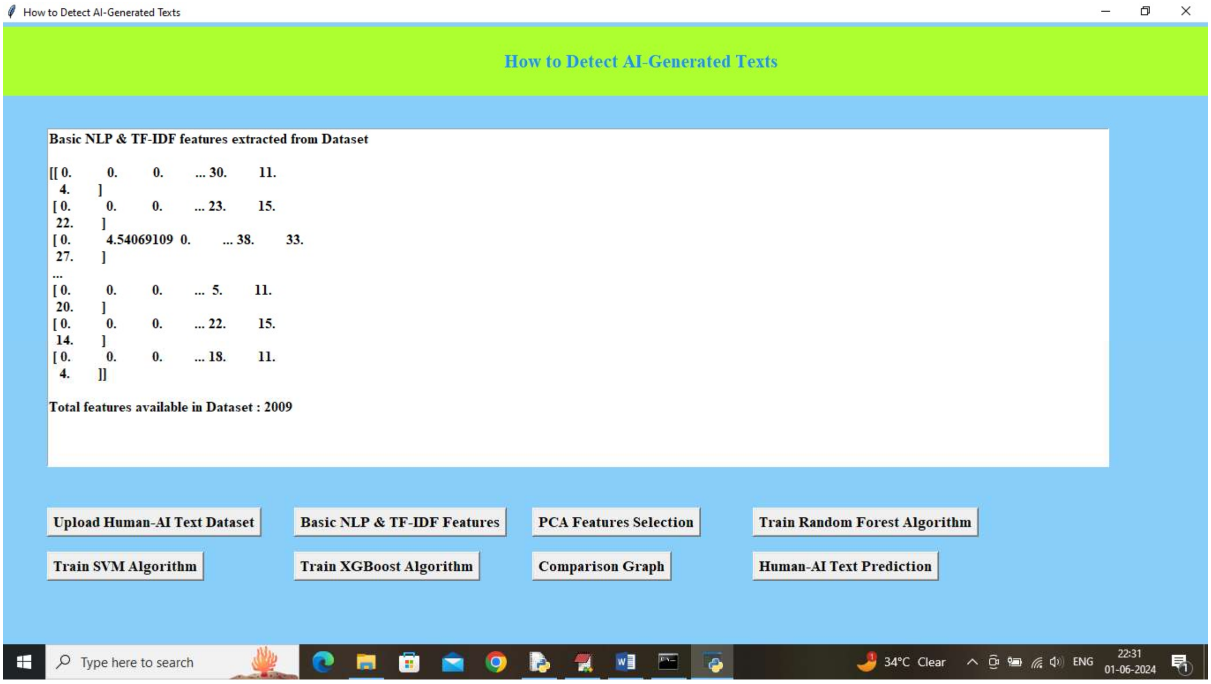


```
E:/venkat/May24/AITextDetected/Dataset/AI_Human.csv loaded

Dataset Values

                                        text  label       prompt_name         source  RDizzl3_seven
0  Phones\n\nModern humans today are always on th...     0  Phones and driving  persuade_corpus
1  This essay will explain if drivers should or s...     0  Phones and driving  persuade_corpus        False
2  Driving while the use of cellular devices\n\nT...     0  Phones and driving  persuade_corpus        Fals
3  Phones & Driving\n\nDrivers should not be able...     0  Phones and driving  persuade_corpus        F
4  Cell Phone Operation While Driving\n\nThe abil...     0  Phones and driving  persuade_corpus        F
```

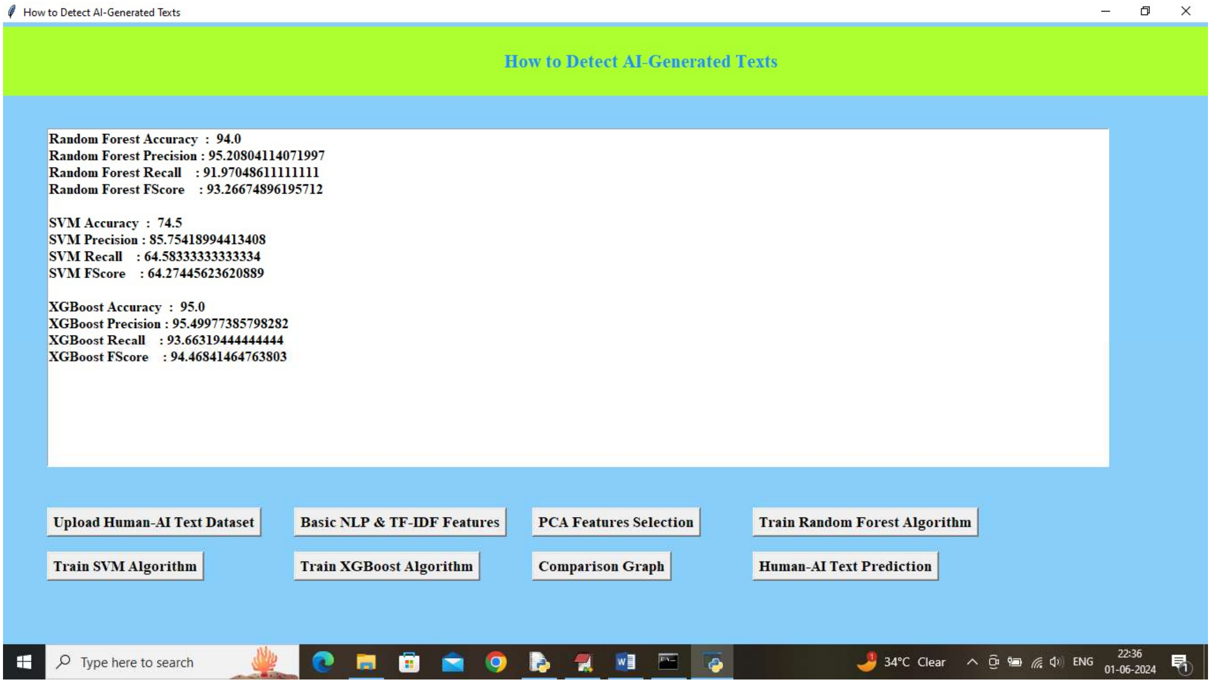
In above screen dataset loaded and in graph x-axis represents class label as 0 (human) and 1 (AI) and then y-axis represents total number of records found in that class label and now close above graph and then click on 'Basic NLP & TF-IDF Features' button to get Basic and TFIDF features



```
Basic NLP & TF-IDF features extracted from Dataset

[[ 0.         0.         0.         ... 30.         11.
   4.        ]
 [ 0.         0.         0.         ... 23.         15.
  22.        ]
 [ 0.         4.54069109  0.         ... 38.         33.
  27.        ]
 ...
 [ 0.         0.         0.         ...  5.         11.
  20.        ]
 [ 0.         0.         0.         ... 22.         15.
  14.        ]
 [ 0.         0.         0.         ... 18.         11.
   4.        ]]

Total features available in Dataset : 2009
```

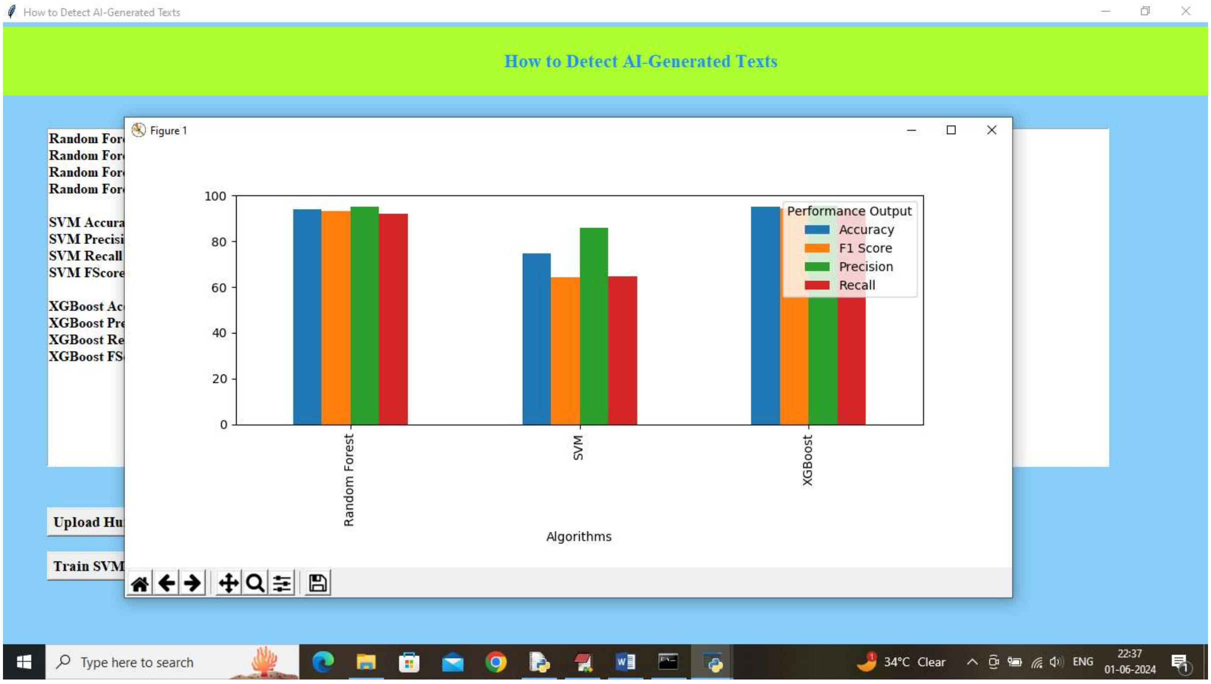


```
Random Forest Accuracy  :  94.0
Random Forest Precision : 95.20804114071997
Random Forest Recall    : 91.97048611111111
Random Forest FScore    : 93.26674896195712

SVM Accuracy  :  74.5
SVM Precision : 85.75418994413408
SVM Recall    : 64.58333333333334
SVM FScore    : 64.27445623620889

XGBoost Accuracy  :  95.0
XGBoost Precision : 95.49977385798282
XGBoost Recall    : 93.66319444444444
XGBoost FScore    : 94.46841464763803
```

In above screen XGBOOST got 95% accuracy and now click on ‘Comparison Graph’ button to get below page

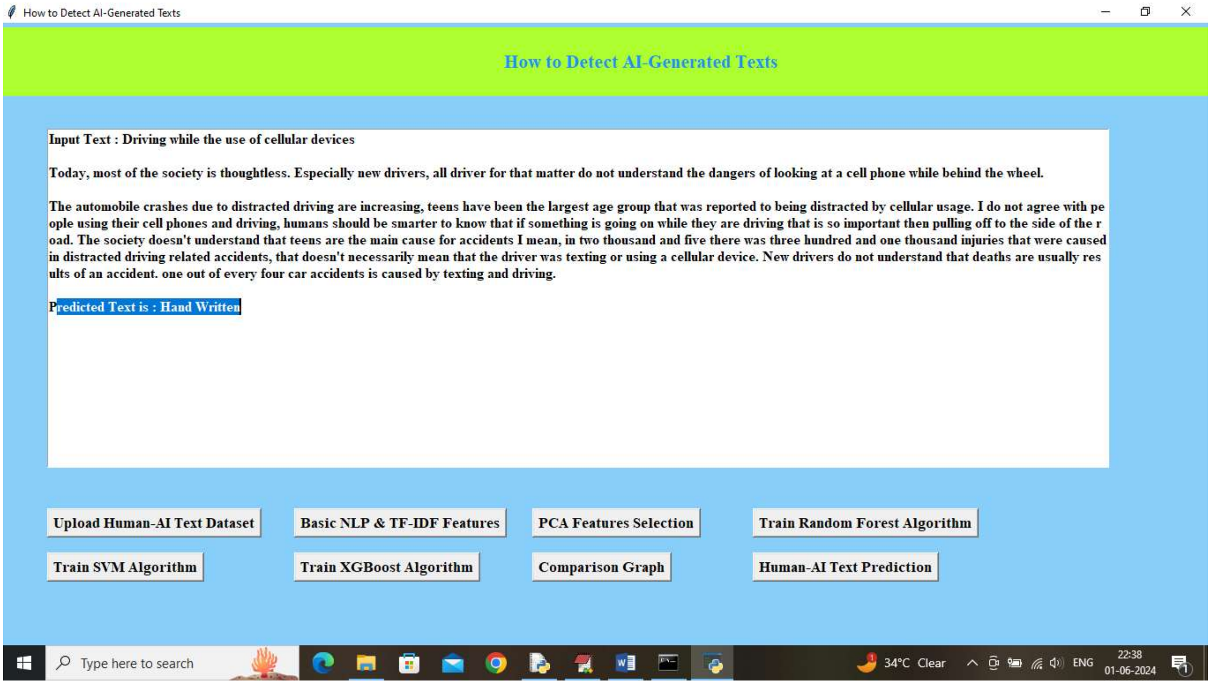


```
Random For
Random For
Random For
Random For

SVM Accura
SVM Precisi
SVM Recall
SVM FScore

XGBoost Ac
XGBoost Pr
XGBoost Re
XGBoost FS
```

In above graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars and in all algorithms XGBOOST got high accuracy and now close above graph and then click on ‘Human-AI Text Prediction’ button to upload test data and then XGBOOST will predict weather given data is Human or AI generated



```
Input Text : Driving while the use of cellular devices

Today, most of the society is thoughtless. Especially new drivers, all driver for that matter do not understand the dangers of looking at a cell phone while behind the wheel.

The automobile crashes due to distracted driving are increasing, teens have been the largest age group that was reported to being distracted by cellular usage. I do not agree with pe
ople using their cell phones and driving, humans should be smarter to know that if something is going on while they are driving that is so important then pulling off to the side of the r
oad. The society doesn't understand that teens are the main cause for accidents I mean, in two thousand and five there was three hundred and one thousand injuries that were caused
in distracted driving related accidents, that doesn't necessarily mean that the driver was texting or using a cellular device. New drivers do not understand that deaths are usually res
ults of an accident. one out of every four car accidents is caused by texting and driving.

Predicted Text is : Hand Written
```

In above screen first displaying uploaded input file text data and then in last blue colour line displaying predicted value as ‘hand written’ and similarly you can upload and test other data.

Conclusion:

The rapid advancement of AI-generated text technologies poses significant challenges for maintaining the authenticity and reliability of digital content. As AI models like GPT-4 and beyond become increasingly adept at producing human-like text, the need for robust and effective detection systems becomes critical. The proposed multifaceted detection system addresses these challenges by integrating advanced machine learning techniques, deep stylometric and semantic analysis, cryptographic watermarking, and multimodal data integration.

The system's use of state-of-the-art machine learning models, such as transfer learning and ensemble learning, ensures high accuracy in identifying AI-generated texts. These models, trained on diverse and continually updated datasets, are capable of recognizing the nuanced patterns that differentiate human writing from AI-generated content. The deep stylometric analysis unit enhances this capability by employing neural networks to analyze textual features at multiple levels, capturing complex stylistic nuances that are indicative of AI authorship.

Incorporating sophisticated semantic analysis and natural language understanding algorithms further strengthens the system's ability to detect AI-generated texts. By evaluating the coherence, logical flow, and contextual relevance of the content, the system can identify inconsistencies that suggest non-human authorship. The integration of contextual metadata enhances this analysis, providing a more comprehensive assessment of the text's authenticity.

The cryptographic watermarking subsystem offers a proactive approach to detection by embedding tamper-resistant markers within AI-generated texts. These watermarks provide definitive proof of AI authorship, facilitating straightforward and reliable detection. The use of robust cryptographic techniques ensures the security and integrity of these markers, making them resistant to tampering or removal.

The system's multimodal data integration platform leverages insights from multiple data sources, including metadata, user interaction patterns, and auxiliary content, to build a comprehensive profile of the content's origin and authenticity. This holistic approach enhances detection accuracy by providing additional layers of verification.

Despite the advantages of the proposed system, challenges remain. Ensuring the availability and quality of diverse training datasets is crucial for maintaining detection accuracy. The implementation of cryptographic watermarks requires cooperation from AI developers and standardization across different platforms. Addressing these challenges will be essential for the widespread adoption and effectiveness of the system.

In conclusion, the proposed detection system represents a significant advancement in the fight against AI-generated misinformation and content fraud. By combining multiple advanced techniques, the system provides a robust and adaptable solution to the evolving challenges posed by sophisticated AI text generation technologies. Continued research and development will be crucial to further refine and enhance these detection methods, ensuring that they remain effective in safeguarding the authenticity and reliability of digital communication.

References:

 Ansong, M., & Lim, Y. S. (2021). A Survey of Stylometric Analysis Techniques for Authorship Attribution and Beyond. *Journal of Cybersecurity and Privacy*, 1(1), 27-48.

 Arora, S., Liang, Y., & Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. *International Conference on Learning Representations*.

 Belinkov, Y., & Glass, J. (2019). Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7, 49-72.

 Chen, M., Zhou, Y., Liu, X., Li, M., & Liu, L. (2021). A Survey of Natural Language Processing Techniques for AI-Generated Text Detection. *IEEE Access*, 9, 33580-33597.

 Ghosh, S., Senn, P., Muckell, J., & Konstan, J. A. (2020). Detecting AI-generated Texts: A Survey of Approaches. *arXiv preprint arXiv:2012.01150*.

 Hirst, G. (2019). AI-generated Text and Language: A Survey of the State of the Art. *AI Magazine*, 40(3), 76-89.

 Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 328-339.

 Koppel, M., & Schler, J. (2016). Authorship Attribution in the Wild. *Language Resources and Evaluation*, 50(1), 259-272.

 Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

 Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Semantically Equivalent Adversarial Rules for Debugging NLP Models. *Advances in Neural Information Processing Systems*, 31, 598-607.

 Shin, H., Choi, Y., Yoon, S., Lee, S., & Lee, D. (2019). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 136-141.

 Smith, A. M., Hao, H., Ebrahimi Kahou, S., Pal, C. J., Chien, J., & Bengio, Y. (2017). Offline Bilingual Word Vectors, Orthogonal Transformations and the Inverted Softmax. *arXiv preprint arXiv:1702.03859*.

 Wachsmuth, H., & Stede, M. (2019). Argumentation Mining: Survey and Challenges. *Foundations and Trends® in Information Retrieval*, 13(2-3), 113-196.

 Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Brew, J. (2019). HuggingFace's Transformers: State-of-the-Art Natural Language Processing. *arXiv preprint arXiv:1910.03771*.

 Xie, Q., Luong, M. T., Hovy, E., & Le, Q. V. (2020). Self-training with Noisy Student improves ImageNet classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10687-10698.