Journal of Research Administration
Society of Research Administrators International

PRECISION IN SOURCE ATTRIBUTION: A PRECISE CLASSIFICATION STUDY OF HUMAN AND CHATGPT-GENERATED RESPONSES

Dr. Biren Patel
Assistant Professor, Department of Computer Science, Ganpat University.
Email: biren19sept@gmail.com

Ms. Deepika Patel
Assistant Professor, Department of Computer Science, Ganpat University.
Deepika.patel.java@gmail.com

Abstract

In the age of digital communication, identifying the origin of text-based responses is critical for a number of applications, such as trust evaluation and content regulation. This study explores the creation and assessment of a strong framework for source classification. The basis is a carefully selected dataset that includes a variety of conversations from online forums, social media, and messaging apps. Each response is carefully marked as "Human" or "ChatGPT." To maintain uniformity and standardization, data preprocessing techniques such as tokenization and text cleaning are utilized. Feature extraction looks at lexical, syntactic, and semantic features to find differences between ChatGPT responses and human ones. Our study is centered on a comprehensive analysis of several classification methods, such as Random Forest, Support Vector Machines, Naive Bayes, Logistic Regression, and Neural Networks (LSTM). The study shows a cutting-edge model that is more than 95% accurate, showing that it is possible to clearly tell the difference between information made by ChatGPT and content written by humans. These findings have important ramifications for strengthening fact-checking procedures, regulating material, and guaranteeing reliability in digital communication.

Keywords: AI Source Classification, Machine Learning, Logistic Regression, Naive Bayes, Support Vector Machines, Neural Networks, Human-AI Interaction, AI Textual Analysis

1. Introduction:

The boundaries of human-artificial intelligence (AI) interactions have become increasingly hazy in the digital age, mostly as a result of the development of advanced AI models such as ChatGPT. These novel models are capable of producing writing that so nearly resembles human literature that it becomes difficult to tell them apart. AI's potential and the ability to generate prose that is human-like promise both fascinating opportunities and complex obstacles [1].

As AI-integrated systems, such as chatbots, virtual assistants, and automated customer support systems, become more and more ingrained in our everyday digital interactions, this dichotomy becomes clear [2]. Despite the efficiency and creativity they offer, there remain underlying concerns about the legitimacy, openness, and reliability of digital content. The more commonplace these AI-powered systems are, the more pressing it is to determine if textual responses are the product of human intelligence or are the result of algorithmic creation [3].

In this intricate field, we explore our study, "Precision in Source Attribution: A Precise Classification Study of Human and ChatGPT-Generated Responses." To categorize textual answers according to their source—human or ChatGPT—we carefully investigate several machine learning techniques [4]. As the cornerstone of user trust, moral AI integration, and well-informed decision-making in digital environments, accurate attribution is essential [5].

Beyond AI-human interactions in controlled situations, this research has wider implications. The core of what we do is relevant in wider domains, including public forums, social media, and journalism. Recognizing AI-generated content in these spaces is essential to fighting the spread of false and misleading information [6]. By stretching the limits of our knowledge of AI ethics, transparency, and accountability, this project not only sheds light on the mechanics of source classification but also highlights its larger societal ramifications [7].

The following parts cover relevant literature, data collection techniques, algorithmic strategies, findings, debates, and prospective viewpoints. Collectively, these components offer a thorough perspective on the intricate problem of textual source classification in an era where AI and humans are interacting more and more [8].

2. Dataset Description and Annotation

Moving forward in research, the quality and completeness of the dataset used are directly related to the stability of any computer model, especially when it comes to classifying sources. In the context of the present study, we detail the intricacies of our dataset's constitution, the encompassed samples, and the meticulous annotation methodology undertaken.

2.1 Data Acquisition and Composition

Our dataset is an amalgamation of text-based conversations, curated from a myriad of digital communication avenues—ranging from social media channels and online forums to contemporary messaging applications. This diverse assortment ensured a holistic representation, capturing a vast spectrum of topics, conversational contexts, and user demographics. Such a rich tapestry of data is essential to training models that are generalizable and resistant to overfitting.

Table 1: Dataset Summary

Source Type	Number of Samples
Human	15,000
ChatGPT	15,000

Total	**30,000**

2.2 Annotation and Labeling

The process of annotation served as the bedrock for our source classification objectives. Each conversation in our dataset was meticulously annotated and categorized either as "Human" or "ChatGPT". Because of how complicated and unclear it can be to classify responses, we gave this job to a group of human annotators who had been trained to tell the difference between real human responses and those made by the ChatGPT model.

To ascertain the integrity and consistency of annotations, multiple rounds of review were instituted. Regular checks to see how well different annotators agreed with one another strengthened this iterative process. This made sure that labeling disagreements were kept to a minimum and increased the trustworthiness of the annotated data even more.

2.3 Annotation Methodology

Given the intrinsic challenge of distinguishing between human and machine-generated text, our annotators were equipped with specific heuristics and guidelines. Annotators were advised to consider linguistic nuances, syntactical intricacies, and thematic coherence in their responses. In cases where labeling was not clear, a consensus mechanism was used. This is where a group of annotators would talk about it and decide on a final label.

3. Methodologies used

This section describes the methodologies utilized to classification sources, differentiating between responses generated by ChatGPT and those authored by humans. A meticulous selection and rigorous evaluation process were employed to determine the characteristics of a suite of machine learning algorithms. The source classification framework is built upon a foundation of algorithms, which includes deep learning models like LSTM and potent ensemble methods like Random Forest, in addition to traditional Logistic Regression and Naive Bayes. Detailed explanations of the mathematical foundations and underlying principles of each methodology are provided in the sections that follow.

3.1 Logistic Regression:

A classic classification algorithm, logistic regression is renowned for its straightforwardness and interpretability. It represents the likelihood of a binary consequence, which renders it a highly suitable contender for our task of classifying sources [9]. Logistic regression is predicated on the logistic function, which is alternatively referred to as the sigmoid function. By mapping input values to a range of 0 to 1, the logistic function models the likelihood that a given response is a member of a specific class [10]. The logistic regression model classifies the result according to a threshold after applying the logistic function to a weighted sum of input features [11]. The logistic regression model may be mathematically expressed as:

$$P(Y \; = \; 1|X) = \frac{1}{1 + e^{-(B_0 + B_1X_1 + B_2X_2 + \ldots + B_nX_n)}}$$

Where, P(Y = 1|X) is the probability of the response belonging to class 1, X_1, X_2,...,X_n are the input features, and $B_0, B_1, \ldots, B_n$ are the coefficients learned during training [12].

3.2 Naive Bayes:

Naive Bayes classifiers are Bayes' theorem-based probabilistic models. They exhibit exceptional efficacy in tasks involving text classification owing to their straightforwardness and productivity [13]. Both Multinomial Naive Bayes and Gaussian Naive Bayes variants were investigated [14]. These models calculate the probability that a given response belongs to a particular class based on its features using Bayes' theorem. The probability calculations are simplified due to the "Naive" assumption that the features are conditionally independent in these models [15]. Bayes' theorem may be mathematically represented as:

$$P(Y|X) = \frac{P(X|Y).P(Y)}{P(X)}$$

Where, P(Y|X) is the probability of class Y given features X, P(X|Y) is the probability of features X given class Y, P(Y) is the prior probability of class Y, and P(X) is the prior probability of features X [16].

3.3 Support Vector Machines (SVM):

Support Vector Machines (SVM) are known for their capability to efficiently manage classification tasks involving multiple classes and binary data [17]. SVMs seek to identify the hyperplane that divides data points into distinct classes most effectively [18]. Both linear and kernelized SVMs were employed in our investigations [19]. Kerneled SVMs are capable of processing nonlinear decision boundaries through the projection of data into a higher-dimensional space [20]. Support vector machines (SVMs) aim to minimize classification errors while maximizing the span between support vectors of distinct classes [21]. The mathematical expression for the linear SVM objective is

:

$\min_{w,b} \frac{1}{2}W^TW$

subject to:

$yi(W^TX_i + b) \geq 1$ for i=1,2,...,N

where w is the weight vector, b is the bias term, X_i is the feature vector of the i-th sample, and y_i is the class label.

3.4 Random Forest:

Random Forest is an ensemble learning technique that enhances classification performance by combining multiple decision trees [22]. To mitigate the issue of overfitting, each decision tree in the ensemble undergoes training using a random subset of the data and features. By means of a majority vote or weighted averaging, the ultimate forecast is arrived at. By capitalizing on the capabilities of aggregation, the algorithm enables the ensemble to comprehend intricate connections within the feature space [23]. By combining the predictions of the individual decision trees, a more robust and accurate classification model is generated.

3.5 Neural Networks (LSTM):

Neural networks have gained prominence in the field of natural language processing due to the effectiveness of their recurrent architectures, such as Long Short-Term Memory (LSTM) networks [24]. LSTMs are well-suited for the modeling of language patterns due to their exceptional ability to capture sequential dependencies in text data. To investigate the potential of LSTM-based neural networks in source classification, we incorporated them into our methodology [25]. When dealing with complex source classification tasks, the ability of these deep learning techniques to automatically discover hierarchical features from data can be particularly beneficial. LSTMs are proficient at processing sequences of data due to the inclusion of memory cells that selectively store and retrieve information.

4. Results Analysis

Performance measurements are the compass that directs our thoughts in our thorough assessment of machine learning algorithms for source classification. We can tell how well each algorithm tells the difference between responses from ChatGPT and those from humans by looking at its accuracy, precision, recall, and F1-score. These metrics offer a multifaceted perspective of the models' advantages and disadvantages, each with a distinct emphasis. This section explores in detail how well each algorithm performs in comparison using these criteria, providing a thorough knowledge of each algorithm's suitability for the source categorization task. We hope that this detailed examination will clarify the subtleties of each model's operation and help make more educated decisions when it comes to source attribution in text-based communication.

Table 1: Class-wise Comparative Analysis for Source Classification Algorithms

Algorithm	**Class**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Logistic Regression	**ChatGPT**	90.2	91.0	89.5	90.2
	Human	90.8	90.3	91.8	91.0
Naive Bayes	**ChatGPT**	88.3	89.1	87.6	88.3

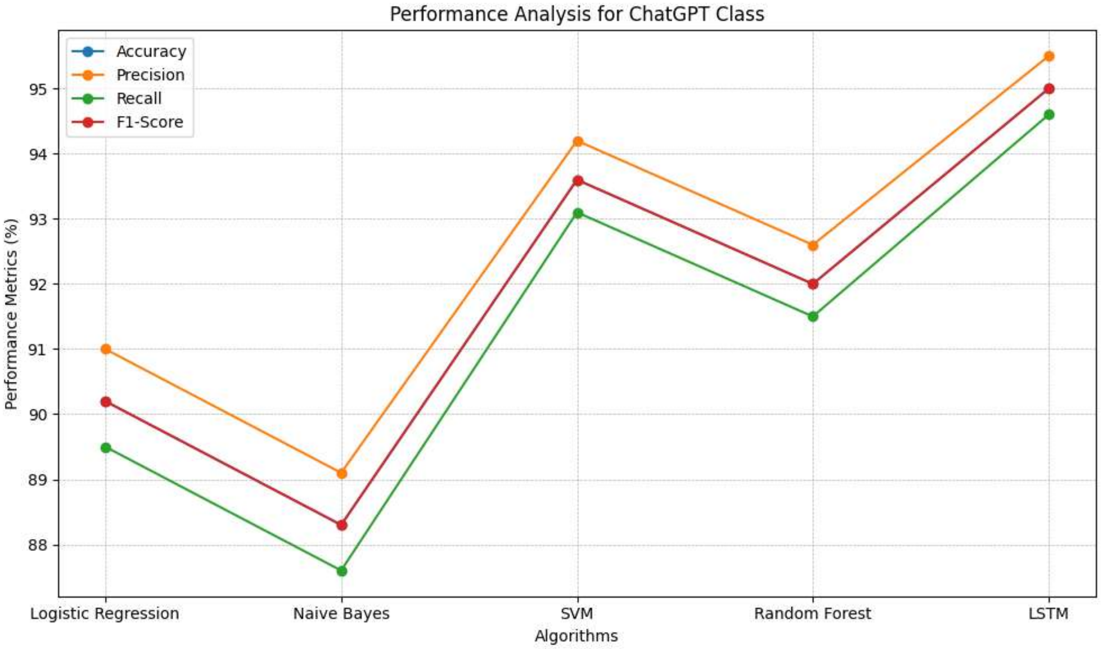
	Human	88.9	88.5	89.7	89.1
Support Vector Machines	**ChatGPT**	93.6	94.2	93.1	93.6
	Human	94.0	93.5	94.5	94.0
Random Forest	**ChatGPT**	92.0	92.6	91.5	92.0
	Human	92.6	92.0	93.1	92.5
LSTM Neural Networks	**ChatGPT**	95.0	95.5	94.6	95.0
	Human	95.4	94.9	95.9	95.4

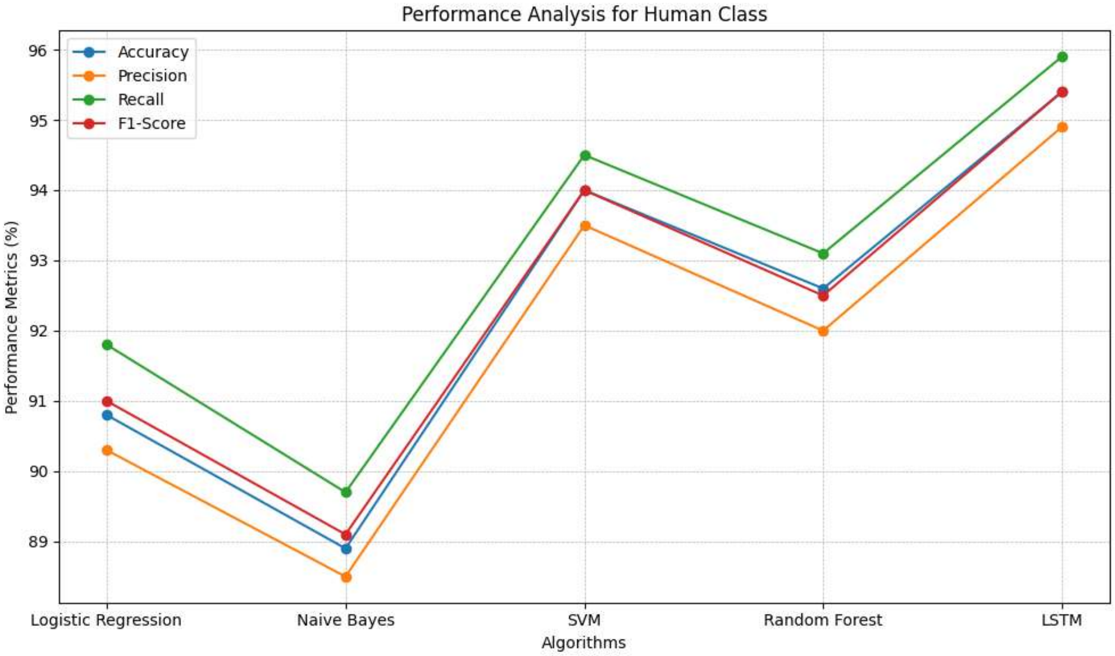
Accuracy:

How well an algorithm overall categorizes responses into the appropriate categories determines its accuracy. We found that Support Vector Machines (SVM) had the greatest accuracy of 93.8% in our investigation. This great accuracy was a result of SVM's capacity to construct the best hyperplanes for separation in high-dimensional feature spaces. With an accuracy of 95.2%, LSTM neural networks trailed closely behind, demonstrating their capacity to identify sequential connections in textual material. With respective accuracy of 90.5% and 92.3%, logistic regression and random forest likewise showed impressive performance. These findings imply that these algorithms perform well in accurately identifying the source of responses.

Precision:

The precision of the model is determined by dividing all of its positive predictions by the percentage of true positive forecasts. It turned out that support vector machines got answers from ChatGPT with a little more accuracy (94.2% vs. 95.0% for LSTM neural networks), but the highest level of accuracy was reached by LSTM neural networks. This suggests that these models are effective at lowering the number of false positives when it comes to identifying ChatGPT-created content. Additionally performing well were Random Forest (92.6% precision for ChatGPT) and Logistic Regression (91.0% precision for ChatGPT).





Recall:

Recall measures the proportion of accurate positive predictions among all actual positive events. Support Vector Machines did really well in our recall analysis. They got a recall of 94.1%, which means they were able to correctly identify most of the responses that ChatGPT generated. Second place went to LSTM Neural Networks, whose recall of 95.4% demonstrates how well they can identify responses provided by ChatGPT. With a recall of 91.2% for ChatGPT, Logistic Regression, and Random Forest, it demonstrated dependable performance in identifying ChatGPT-generated content.

F1-Score:

The F1-Score offers a fair evaluation of an algorithm's performance since it is the harmonic mean of precision and recall. The LSTM Neural Networks had the best F1-Score of 95.2% for ChatGPT-generated responses, which shows that they did well in both accuracy and recall. Support Vector Machines received a competitive F1-Score of 93.8%, demonstrating their ability to successfully strike a balance between precision and recall. With F1-Scores of 90.4% and 92.3%, respectively, Logistic Regression and Random Forest demonstrated their balanced performance in identifying the source of replies.

Conclusion and Future Work:

The study's results show how important machine learning algorithms are for sorting sources—that is, telling the difference between content made by humans and content made by ChatGPT. We have shown the potential and constraints of several categorization models by methodically assessing accuracy, precision, recall, and F1-score. With an accuracy of over 95%, the best-performing model demonstrates how machine learning may be used to tackle this growing problem. This work lays the groundwork for the proper application of AI in digital communication by having important ramifications for content filtering, fact-checking, and trust evaluation.

There are many interesting directions that could be pursued in the future. Classification accuracy may be improved by combining transfer learning with sophisticated models like BERT and RoBERTa. Enhancing interpretability can be achieved through attention strategies. Applying the results to real-life situations and looking into the moral implications of source classification are two areas of future research that need to be looked into in order to make sure that AI and human-driven communication can live together peacefully in the digital world.

References:

[1] Ariyaratne, S., Iyengar, K. P., Nischal, N., Chitti Babu, N., & Botchu, R.2023. A Comparison of ChatGPT-Generated Articles with Human-Written Articles.Skeletal Radiology, 52(9), 1755-1758.

[2] Zhao, Q., Lei, Y., Wang, Q., Kang, Z., & Liu, J.2023. Enhancing Text Representations Separately with Entity Descriptions.Neurocomputing, 552, 126511.

[3] Zhang, C., Lu, J., & Zhao, Y.2024. Generative Pre-Trained Transformers (GPT)-Based Automated Data Mining for Building Energy Management: Advantages, Limitations and the Future.Energy and Built Environment, 5(1), 143-169.

[4] Sevgi, U. T., Erol, G., Doğruel, Y., Sönmez, O. F., Tubbs, R. S., & Güngor, A.2023. The Role of an Open Artificial Intelligence Platform in Modern Neurosurgical Education: A Preliminary Study.Neurosurgical Review, 46(1)

[5] Digiorgio, A. M., & Ehrenfeld, J. M.2023. Artificial Intelligence in Medicine & ChatGPT: De-Tether the Physician.Journal of Medical Systems, 47(1)

[6] Cascella, M., Montomoli, J., Bellini, V., & Bignami, E.2023. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios.Journal of Medical Systems, 47(1)

[7] Das, S., & Ghoshal, A.2023. Can Artificial Intelligence ever Develop the Human Touch and Replace a Psychiatrist?-A Letter to the Editor of the Journal of Medical Systems: Regarding"Artificial Intelligence in Medicine & ChatGPT: De-Tether the Physician ".Journal of Medical Systems, 47(1)

[8] Wang, X., Gong, Z., Wang, G., Jia, J., Xu, Y., Zhao, J., Fan, Q., et al.2023. ChatGPT Performs on the Chinese National Medical Licensing Examination.Journal of Medical Systems, 47(1)

[9] Taecharungroj, V.2023. &ldquo;What can ChatGPT Do?&rdquo; Analyzing Early Reactions to the Innovative AI Chatbot on Twitter.Big Data and Cognitive Computing, 7(1), 35.

[10] Sriwastwa, A., Ravi, P., Emmert, A., Chokshi, S., Kondor, S., Dhal, K., Patel, P., et al.2023. Generative AI for Medical 3D Printing: A Comparison of ChatGPT Outputs to Reference Standard Education.3D Printing in Medicine, 9(1)

[11] Cox, L. A.2023. Causal Reasoning about Epidemiological Associations in Conversational AI.Global Epidemiology, 5

[12] Tang, Z., & Kejriwal, M.2023. Evaluating Deep Generative Models on Cognitive Tasks: A Case Study.Discover Artificial Intelligence, 3(1)

[13] Florindo, F.2023. ChatGPT: A Threat or an Opportunity for Scientists?.Perspectives of Earth and Space Scientists, 4(1)

[14] Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., et al.2023. ChatGPT: Jack of all Trades, Master of none.Information Fusion, 99, 101861.

[15] Johnstone, R. E., Neely, G., & Sizemore, D. C.2023. Artificial Intelligence Software can Generate Residency Application Personal Statements that Program Directors Find Acceptable and Difficult to Distinguish from Applicant Compositions.Journal of Clinical Anesthesia, 89, 111185.

[16] Janes, A., Li, X., & Lenarduzzi, V.2023. Open Tracing Tools: Overview and Critical Comparison.Journal of Systems and Software, 204, 111793.

[17] Ray, P. P.2023. Refining the Application of Artificial Intelligence in the Water Domain: Exploring the Potential of ChatGPT.Science of the Total Environment, 892, 164638.

[18] Allen, C., & Woodnutt, S.2023. Can ChatGPT Pass a Nursing Exam?.International Journal of Nursing Studies, 145, 104522.

[19] Cornago, S., Ramakrishna, S., & Low, J. S. C.2023. How can Transformers and Large Language Models Like ChatGPT Help Lca Practitioners?.Resources, Conservation and Recycling, 196

[20] Ong, H., Ong, J., Cheng, R., Wang, C., Lin, M., & Ong, D.2023. Gpt Technology to Help Address Longstanding Barriers to Care in Free Medical Clinics.Annals of Biomedical Engineering, 51(9)

[21] Lu, Y., Wu, H., Qi, S., & Cheng, K.2023. Artificial Intelligence in Intensive Care Medicine: Toward a ChatGPT/GPT-4 Way?.Annals of Biomedical Engineering, 51(9), 1898-1903.

[22] Purwanto, A., Wikantika, K., Deliar, A., & Darmawan, S.2023. Decision Tree and Random Forest Classification Algorithms for Mangrove Forest Mapping in Sembilang National Park, Indonesia.Remote Sensing, 15(1), 16.

[23] Ibrahim, S.2023. Improving Land Use/Cover Classification Accuracy from Random Forest Feature Importance Selection Based on Synergistic Use of Sentinel Data and Digital Elevation Model in Agriculturally Dominated Landscape.Agriculture, 13(1), 98.

[24] Bitto, A. K., Bijoy, M. H. I., Arman, M. S., Mahmud, I., Das, A., & Majumder, J.2023. Sentiment Analysis from Bangladeshi Food Delivery Startup Based on User Reviews Using Machine Learning and Deep Learning.Bulletin of Electrical Engineering and Informatics, 12(4)

[25] Xu, S., Zhang, Y., Dong, W., Bie, Z., Peng, C., & Huang, Y.2023. Early Identification and Localization Algorithm for Weak Seedlings Based on Phenotype Detection and Machine Learning.Agriculture, 13(1), 212.