Sustainable Multi-Author Writing Style Analysis for Identifying Stylistic Differences Between Authors

Ganapathi Raju *N.V*[1*], *Yadavalli* Vikas[1], *Md. Furkhan* Ansari[1] , *D.* Likhith [1] *Himani* Badoni[2]

[1]Department of Information technology, GRIET, Bachupally, Hyderabad, JNTUH, Telangana,India,500090.

[2]School of Applied and Life Sciences, Uttaranchal University, Dehradun, 248007, India

Abstract. Natural language processing (NLP) is a sustainable subfield of Artificial Intelligence that focuses on the interaction between computers and humans through natural language. NLP algorithms enable computers to comprehensively understand, interpret, and generate human language, thus facilitating the sustainable analysis and comprehension of vast amounts of textual data. Within the context of sustainable style change detection, NLP algorithms play a pivotal role in analyzing multi-author documents and identifying the points at which authors transition. This sustainable step is critical in authorship recognition as it furnishes a more precise comprehension of which sections were authored by different individuals. A multi-author document's writing style can evolve over time, and this sustainability can prove invaluable in fields such as forensics, journalism, and literary studies, among others.The sustainable goal of this project is to investigate various NLP methods for sustainable style change detection. By scrutinizing datasets and juxtaposing them with advanced methodologies in the existing literature, the effectiveness of these strategies will be ascertained. The overarching aim of our study is to foster the progress of both the field of NLP research and its sustainable practical applications.

1 Introduction

The sustainable goal of this study is to examine multi-author documents using Natural Language Processing techniques in order to identify changes in writing styles with an eye towards long-term viability. Understanding the unique characteristics of author’s voices is crucial since the internet and social media generate enormous amounts of textual material every day, emphasising the need for sustainable methods of analysis. The project makes use of a dataset from the Zenodo platform, which includes texts by different writers in various genres, aligning with sustainable principles of diverse representation. To categorise texts, group related papers, and extract underlying themes, NLP techniques like text classification are used in a sustainable manner, fostering efficiency and ecological responsibility.The

[*] Corresponding author: nvgraju@griet.ac.in

sustainable goal of the project is to spot instances of writing style changes and contrast the findings with prior research, fostering continuous improvement and adaptive learning. The findings of this study have significant ramifications for the sustainable study of literature, journalism, and forensics. Researchers and analysts can use NLP algorithms to comprehend how different authors contributed to a material and how their writing styles might have varied, thereby fostering a more inclusive and ecologically minded approach to research. Establishing authorship, spotting instances of plagiarism or ghostwriting, and keeping track of changes in writing styles through time are all made possible with the sustainable help of this data. This project aims to investigate the sustainable potential of NLP techniques in 2ndeavour the writing styles of numerous authors, fostering innovative and sustainable research practices. Through the sustainable study of textual data, this 2ndeavour contributes to the continuous evolution of research methodologies.

2 Literature Survey

In their research, titled "An Ensemble-Rich Multi-Aspect Approach for Robust Style Change Detection," the authors changed the project's technique to include a combination of the LightGBM algorithm and the Multilayer Perceptron (MLP) model with TFIDF encoding. In order to model visual and sequential data, they also used convolutional neural networks. Their results highlighted the value of combining different approaches, especially the stacking approach that used the LightGBM classifier to increase predictive performance in style change detection. Dimitrina Zlatkova et.al., [1]

In their research, titled "Feature Vector Difference based Neural Network and Logistic Regression Models for Authorship Verification," the authors modified their methods by using Linear Regression for a smaller dataset and Neural Network for a bigger dataset. This strategy balanced predictability and interpretability, enhancing performance and prediction precision. Five metrics, including the area under the ROC curve (AUC) and F1-score, as well as feature analysis to find important predictors, were used to evaluate the final models used in TIRA. Janith Weerasinghe et.al., [2]

In their research, "Style Change Detection on Real-World Data using an LSTM-powered Attribution Algorithm," the author employed two different models: a Multilayer Perceptron (MLP) and a Bidirectional Long Short-Term Memory (LSTM) architecture. In order to better modelling and prediction, these models were combined in order to capture both sequential patterns and nonlinear correlations in the data. The study emphasised the value of simple machine learning models while simultaneously acknowledging the complexity of handling real-world problems. Robert Deibel [3]

In their research, titled "Multi-label Style Change Detection by Solving a Binary Classification Problem," the authors used a stacking ensemble technique using base-level classifiers trained on text characteristics and text embeddings. Among the most important tools were the LightGBM classifier and the Scikit-Learn package. The major findings emphasised the significance of feature extraction and the switch from binary to multi-label predictions, and the methodology aimed to strike a balance between trustworthy performance and computational economy. Eivind Strøm [4]

In their research, titled "Ensemble-Based Clustering for Writing Style Change Detection in Multi-Authored Textual Documents," the authors employed a variety of clustering techniques such K-Means, DBSCAN, and Agglomerative Hierarchical Clustering to find groupings of text segments with comparable writing styles. In order to analyse and analyse patterns in writing style, the ensemble clustering method was used to identify distinctive clusters within the dataset. Shams Alshamasi et.al., [5] In their research, titled "Style Change Detection Based On Bert And Conv1d," the authors' modified methodology made use of BERT and One-dimensional Convolution (Conv1D). Conv1D was utilised for local feature extraction, while BERT, a pre-trained language model, was used for representational learning and contextual comprehension. Further advancements in text information capture and learning appeared possible when BERT[18] and Conv1D were combined. Qidi Lao et.al., [6-20]

3 System Architecture

A software system's high-level design and structure are referred to as its system architecture. To fulfil the system's functionality and performance objectives, it defines the components, modules, and interactions between them. It describes how hardware and software components are organised, as well as how data flows and communication channels. The scalability, adaptability, and maintainability of the software programme are all guaranteed by a well-designed system architecture. It acts as a guide for programmers to efficiently implement and construct the system.

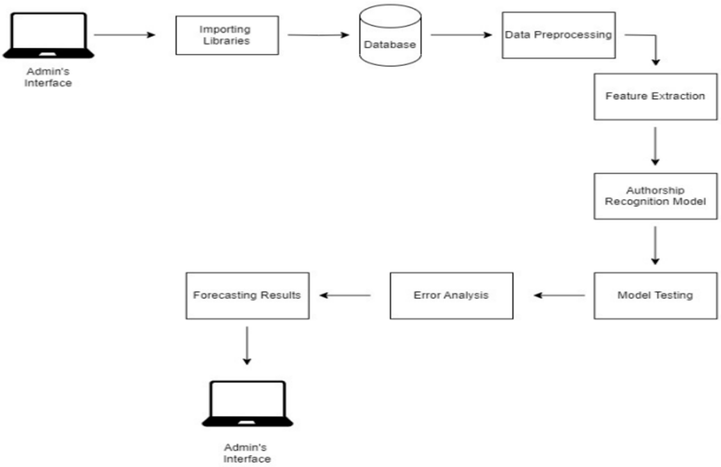
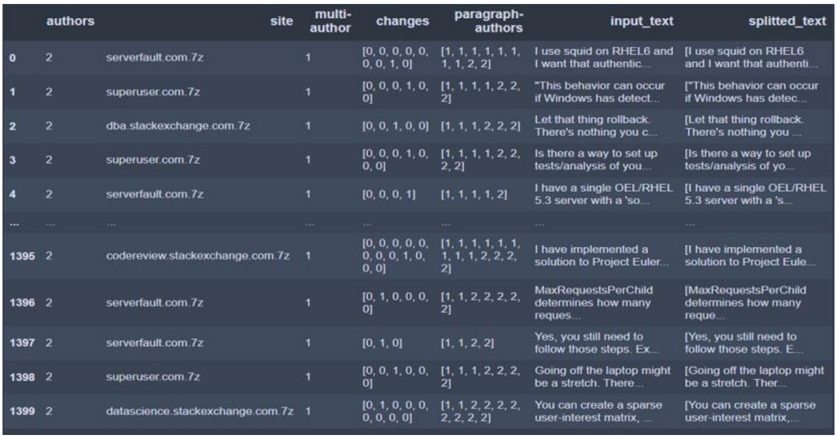


Fig. 1. System Architecture

4 Methodology

4.1 Dataset

Dataset is collected from the zenodo platform. The dataset contains training dataset, validation dataset, test dataset. Training and validation dataset contains two types of files in it .txt and .json files .txt files contain the text collected from a website written by multiple authors and .json files contain the truth values of the .txt files. This dataset is then loaded into our code for further processing of it.



	authors	site	multi-author	changes	paragraph-authors	input_text	splitted_text
0	2	serverfault.com.7z	1	[0, 0, 0, 0, 0, 0, 0, 1, 0]	[1, 1, 1, 1, 1, 1, 1, 1, 2, 2]	I use squid on RHEL6 and I want that authentic...	[I use squid on RHEL6 and I want that authenti...
1	2	superuser.com.7z	1	[0, 0, 0, 1, 0, 0]	[1, 1, 1, 1, 2, 2, 2]	"This behavior can occur if Windows has detect...	["This behavior can occur if Windows has detec...
2	2	dba.stackexchange.com.7z	1	[0, 0, 1, 0, 0]	[1, 1, 1, 2, 2, 2]	Let that thing rollback. There's nothing you c...	[Let that thing rollback. There's nothing you ...
3	2	superuser.com.7z	1	[0, 0, 0, 1, 0, 0, 0]	[1, 1, 1, 1, 2, 2, 2, 2]	Is there a way to set up tests/analysis of you...	[Is there a way to set up tests/analysis of yo...
4	2	serverfault.com.7z	1	[0, 0, 0, 1]	[1, 1, 1, 1, 2]	I have a single OEL/RHEL 5.3 server with a 'so...	[I have a single OEL/RHEL 5.3 server with a 's...
...	...	...	...	...	...	...	...
1395	2	codereview.stackexchange.com.7z	1	[0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0]	[1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2]	I have implemented a solution to Project Euler...	[I have implemented a solution to Project Eule...
1396	2	serverfault.com.7z	1	[0, 1, 0, 0, 0, 0]	[1, 1, 2, 2, 2, 2, 2]	MaxRequestsPerChild determines how many reques...	[MaxRequestsPerChild determines how many reque...
1397	2	serverfault.com.7z	1	[0, 1, 0]	[1, 1, 2, 2]	Yes, you still need to follow those steps. Ex...	[Yes, you still need to follow those steps. E...
1398	2	superuser.com.7z	1	[0, 0, 1, 0, 0, 0]	[1, 1, 1, 2, 2, 2, 2]	Going off the laptop might be a stretch. There...	[Going off the laptop might be a stretch. Ther...
1399	2	datascience.stackexchange.com.7z	1	[0, 1, 0, 0, 0, 0, 0, 0, 0]	[1, 1, 2, 2, 2, 2, 2, 2, 2, 2]	You can create a sparse user-interest matrix, ...	[You can create a sparse user-interest matrix,...

Fig. 2. Dataframe of provided dataset

4.2 Preprocessing of Data:

We will read all the files in the dataset both .txt and .json files. We then do pre-process of our data with data augmentation to our training dataset and do preprocessing of our data without data augmentation to our validation data.

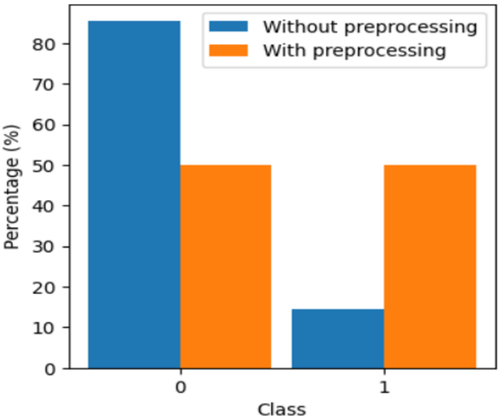


Fig. 3. Graph on how the dataset is pre-processed

4.3 Data Feature Extraction:

The pre-processed text can be used to extract important linguistic characteristics such as sentence length, word frequency, vocabulary richness, paragraph structure, and punctuation usage. Create a numerical representation of the text data for analysis.

4.4 BERT:

The pre-trained deep learning language model BERT, also known as Bidirectional Encoder Representations from Transformers, was created by Google. By taking into account the words in both left and right contexts, it aims to comprehend the context and meaning of words in a phrase. The Transformer design, on which BERT is based, enables it to effectively capture long-range dependencies in text. Numerous Natural Language Processing (NLP) activities, such as text categorization, named entity recognition, sentiment analysis, and question answering, have made extensive use of it. The state-of-the-art in many NLP applications has considerably improved because to BERT's capacity to handle contextual information.

4.5 Workflow and Algorithm:

1. The dataset for our project is collected from Zenodo platform. Our dataset contains 3,700 files.
2. Our dataset consists of two types of files .txt and .json files. All .txt files contain text data in them and all .json files contains the ground truth, i.e., the correct solution in JSON format.
3. The text in the files is then tokenized with the help of AutoTokenizer.
4. Now these sentences are pre-processed and token_id, attention_mask, token_type_id are generated for the tokenized text.
5. The pre-processing involves:
   - Removing special characters
   - Padding
   - Lower cases of all letters in the sentence
6. Now these sentences are stored in the dataframe using pandas, Dataframe consists:
   - Number of Authors in the text file
   - From which site was the text taken
   - Author change which is denoted by 0,1
   - Paragraph author
   - Input text
   - Splitted text
7. We are using BERT architecture to train our model, we are training our model based on 3 different types of optimizers they are:
   - ADAM
   - SGD
   - ADAMAX
8. The above used model is trained by using the fit() function and this model is evaluated on the test dataset.

4.6 Figures and Tables

Table 1. ADAM classification report

	Classification Report			
	Precision	**Recall**	**f1-score**	**support**
0.0	0.96	0.82	0.88	1841
1.0	0.42	0.81	0.55	300
accuracy			0.81	2141
macro avg	0.69	0.81	0.72	2141
weighted avg	0.89	0.81	0.84	2141

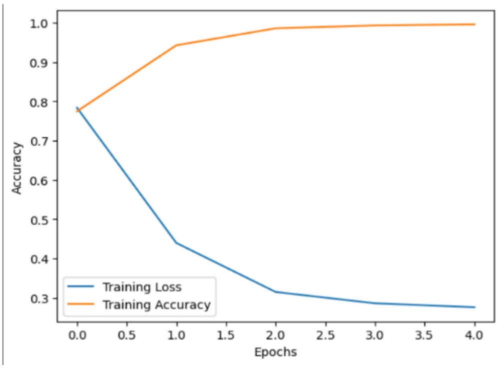


Fig. 4. ADAM Accuracy vs Loss

Table 2. SGD classification report

	Classification Report			
	Precision	**Recall**	**f1-score**	**support**
0.0	0.88	0.63	0.73	1841
1.0	0.17	0.48	0.26	300
accuracy			0.61	2141
macro avg	0.53	0.56	0.50	2141
weighted avg	0.78	0.61	0.67	2141

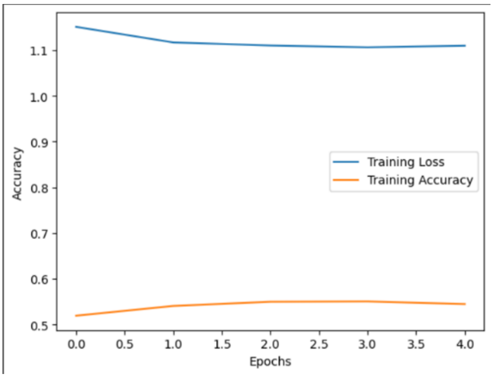


Fig. 5. SGD Accuracy vs Loss

Table 3. ADAMAX classification report

	Classification Report			
	Precision	**Recall**	**f1-score**	**support**
0.0	0.97	0.81	0.88	1841
1.0	0.42	0.85	0.56	300
accuracy			0.81	2141
macro avg	0.70	0.83	0.72	2141
weighted avg	0.89	0.81	0.84	2141

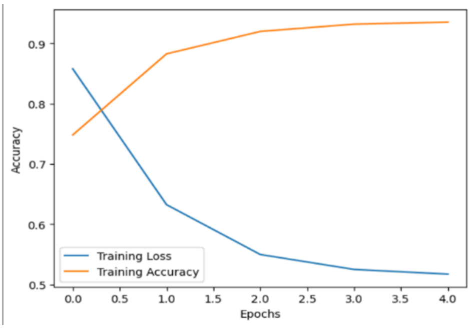


Fig. 6. ADAMAX Accuracy vs Loss

5 Results Analysis

Analysing results entails evaluating how well the model accomplishes the project's goals. Based on the particular NLP task, pertinent evaluation measures are selected for result analysis, such as accuracy, precision, recall, F1-score, or perplexity. These metrics give precise evaluations of the model's performance. To ensure an objective assessment of the model's generalisation to real-world circumstances, a separate test dataset that was not used during training is used. To identify the model's relative strengths and improvements, performance comparisons with baseline models or advanced methods are useful. Graphs and plots are visualisation tools that help present the results clearly. The real-world application and impact of the NLP project are also taken into account in the result analysis.

6 Conclusion

In this study, our goal was to analyze the writing styles of several writers using the BERT model. Three distinct optimizers were used to achieve this goal: Adam, SGD, and Adamax. These optimizers produced accuracy rates of 79%, 60%, and 81%, respectively. Adamax outperformed the other two optimizers examined, achieving the maximum accuracy of 81% in the analysis of the writing styles of many authors. It demonstrated great memory (81%) and precision (97%) demonstrating its capacity to successfully identify positive cases and reduce false positives. The F1 score of 88% showed that precision and recall were harmoniously balanced. The macro F1 score, which measures performance across all classes, was 76%, which is considered satisfactory. SGD optimizer, on the other hand, produced the lowest accuracy (60%) and displayed inferior precision, recall, and F1 score. Adam optimizer earned a 79% accuracy, a 96% precision, a 79% recall, and an 87% F1 score. In conclusion, Adamax showed to be the best accurate and efficient optimizer for the multi-author writing style analysis.

7 Future Enhancements

The multi-author writing style analysis project can be improved in the future in a lot of ways. Among them are improving the model architecture by investigating deeper or transformer-based models, fine-tuning hyperparameters to achieve optimal performance, utilizing data augmentation techniques to increase dataset diversity, using ensemble methods to capitalize on the strengths of multiple models, incorporating pre-trained language models for transfer learning, incorporating additional features to capture more in-depth writing style patterns, and addressing dataset imbalance through The multi-author writing style analysis system's capabilities and efficacy may be further improved by these changes. The model should be expanded to check the text for writers beyond the existing limit of two authors as part of the multi-author writing style analysis project. This would entail updating the architecture to account for the increasing complexity of detecting and differentiating between various writing styles, training the model on a larger dataset with samples from several writers, and so on.

References

1. Dimitrina Zlatkova, Daniel Kopev, Kristiyan Mitov, and Atanas Atana. 2018. An ensemblerich multi-aspect approach for robust style change detection. PAN at CLEF (2018)
2. Janith Weerasinghe and Rachel Greenstadt. Feature Vector Difference based Neural 50 Network and Logistic Regression Models for Authorship Verification. Vol-2696, (2020)
3. Robert Deibel , Denise Löfflad. Style Change Detection on Real-World Data using an LSTM-powered Attribution Algorithm. PAN AT CLEF, Vol-2936, (2021)
4. Eivind Strøm. Multi-label Style Change Detection by Solving a Binary Classification Problem, Vol-2936, (2021)
5. Shams Alshamasi , Mohamed Menai.Ensemble-Based Clustering for Writing Style
6. Change Detection in Multi-Authored Textual Documents, Vol-3180, (2022)

7. Qidi Lao, Li Ma, Wenyin Yang, Zexian Yang, Dong Yuan, Zhenlin Tan and Langzhang
8. Liang. Style Change Detection Based on Bert And Conv1d, Vol-3180, (2022)
9. Raju, NV Ganapathi, V. Vijay Kumar, and O. Srinivasa Rao. "Author based rank vector coordinates (ARVC) Model for Authorship Attribution." International Journal of Image, Graphics and Signal Processing 8.5 (2016): 68.
10. Raju, NV Ganapathi, V. Vijay Kumar, and O. Srinivasa Rao. "AUTHORSHIP ATTRIBUTION OF TELUGU TEXTS BASED ON SYNTACTIC FEATURES AND MACHINE LEARNING TECHNIQUES." Journal of Theoretical & Applied Information Technology 85.1 (2016).
11. Raju, NV Ganapathi, and Someswara Rao Chinta. "Region based instance document (rid) approach using compression features for authorship attribution." Annals of Data Science 5.3 (2018): 437-451.
12. Kumar, V. Vijay, NV Ganapathi Raju, and O. Srinivasa Rao. "Histograms of Term Weight Feature (HTWF) model for Authorship Attribution‖." International Journal of applied Engineering Research 10.16 (2015): 36622-36628.
13. H kanaka Durga Bella, DR.S. Vasundra ,CSE, JNTUA,"A study of security threats and attacks in Cloud Computing " IEEE-4th International conference on smart systems and inventive technology (ICSSIT-2022), DOI: 10.1109/ICSSIT53264.2022.
14. Ledalla, Sukanya & Mahalakshmi, Tummala. (2018). Multilingual Sentiment Analysis of Hinglish Tweets. Indian Journal of Public Health Research & Development. 9. 1627. 10.5958/0976-5506.2018.02092.2.
15. Prasanna Lakshmi, K., Reddy, C.R.K. A survey on different trends in Data Streams (2010) ICNIT 2010 - 2010 International Conference on Networking and Information Technology, art. no. 5508473, pp. 451-455.
16. Jeevan Nagendra Kumar, Y., Spandana, V., Vaishnavi, V.S., Neha, K., Devi, V.G.R.R. Supervised machine learning Approach for crop yield prediction in agriculture sector (2020) Proceedings of the 5th International Conference on Communication and Electronics Systems, ICCES 2020, art. no. 09137868, pp. 736-741.
17. Sankara Babu, B., Suneetha, A., Charles Babu, G., Jeevan Nagendra Kumar, Y., Karuna, G. Medical disease prediction using grey wolf optimization and auto encoder based recurrent neural network (2018) Periodicals of Engineering and Natural Sciences, 6 (1), pp. 229-240.
18. Nagaraja, A., Boregowda, U., Khatatneh, K., Vangipuram, R., Nuvvusetty, R., Sravan Kiran, V. Similarity Based Feature Transformation for Network Anomaly Detection (2020) IEEE Access, 8, art. no. 9006824, pp. 39184-39196.
19. Sri Lalitha Y., Prashanthi G., Sravani Puranam, Sheethal Reddy Vemula, Preethi Doulathbaji and Anusha Bellamkonda, "Natural Language to SQL: Automated Query Formation Using NLP Techniques", E3S Web of Conferences Volume 391, 2023.
20. Y. Sri Lalitha, G. V. Reddy, K. Swapnika, R. Akunuri and H. K. Jahagirdar, "Analysis of Customer Reviews using Deep Neural Network," 2022 International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR), Hyderabad, India, 2022, pp. 1-5, doi: 10.1109/ICAITPR51569.2022.9844183. EISBN : 978-1-6654-2521-6