12-16-2022

Writer Identity using Stylometry and Machine Learning

Abdulla Alshuweihi
ama6297@rit.edu

Sultan Alblooshi
sha4132@rit.edu

Writer Identity using Stylometry and Machine Learning

by

Abdulla Alshuweihi & Sultan alblooshi

A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies:

Data analytics

Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

December 16 , 2022

RIT

Master of Science in Professional Studies:

Data analytics

Graduate Capstone Approval

Student Name**:**

Abdulla Alshuweihi & Sultan alblooshi

Graduate Capstone Title**: Writer identity using stylometry & Machine Learning.**

Graduate Capstone Committee:

Name: Dr. Sanjay Modak **Date:**

Chair of committee

Name: Dr. Ehsan Warriach **Date:**

Member of committee

Acknowledgments

First and foremost, I would like to express my heartfelt appreciation to my mentor, Dr. Ehsan , for his unwavering support on my Masters capstone, as well as his enthusiasm, patience, and vast expertise. His advice was vital in the study and writing of this capstone.

My sincere gratitude also goes to Dr. Sanjay Modak, the committee chair, who’s office door was always open to provide assistance and guidance, which also played a huge role in making this capstone feasible.

Contents

Table of Figures

Abstract

This particular study will be based on linguistics and stylometric use to find or identify the legitimate authors of the text. For this purpose, the study is expected to use the machine learning approach or framework that consists of various features to sort out and find the style of writing belonging to the right author. The machine learning approach is accompanied by the support of SAS (Statistical Analysis System).

SAS covers the algorithms problems required for better accurate functioning of machine learning approach experiments. In this learning framework, the experiments use a large amount of data to sort and filter out the right person responsible for the content written. AI technologies and SAS are the two components that help the machine learning experiments work accurately and provide reliable results. Moreover, this study will be using the statistical methodology of CRISP-DM (Cross Industry Standard Process for Data Mining), which is suitable for such large data mining projects like this particular project of identifying authors from massive data of the document.

Moreover, the tweets dataset used for this research is already available on the Kaggle platform in the .csv format. The dataset contains textual data for five (05) authors and 10000 approximate tweets. Each author has 1900 to 2000 tweets related to his/her name. In the proposed model, the dataset is initially passed through certain preprocessing steps such as cleansing of data from null values and unnecessary details. Also, stop words, nouns, Adverbs, or other particular parts of speech are removed from the data.

After pre-processing, different SAS-based Machine learning models are applied to relate the specific text to the author. For this purpose, a specific CRISP-DM model is adopted and four (04) different machine learning algorithms are tested. For the training of each model, the train test split is set to be 80-20. Initially, the Bayesian Network is applied to the dataset followed by the classifier. It is observed from the results that the Decision Tree classifier outperforms Bayesian Network. Afterward, Gradient Boosting Trees and MBR are tested with the same data. The end results for each model are: MBR Model = 97.09, Gradient Boosting = 97.06, Decision Tree = 83.89, HPBNC = 81.36. The results are better from most of the state-of-the-art mechanisms with

the same dataset. Moreover, it is worth mentioning that MBR and Gradient Boosting have performed exceptionally well with the forensic texts.

This research may be utilized as a starting point for forensic examination of Twitter data to identify ownership and Stylometry style. The accuracy of the models is high, however, it might be improved in the future by utilizing different parameters and methodologies instead of current research. Lastly, this research will be extremely useful for any country's cybercrime unit in reducing bogus news, and postings, and determining which news truly belongs to them.

Keywords: Statistical Analysis System, Stylometry, Writer Identity, Twitter, , Machine Learning Algorithms, CRISP-DM

"Everyone is unique," we've been told for centuries. All individual possesses a distinct personality, identity, retina, and many other characteristics. These characteristics are critical in identifying persons for secure authentication. However, when it relates to the protection of a person's published literature or phrases, these basic distinct identities are useless. One cannot recognize an author from a published line of writing using retina or fingerprint scans, and occasionally indeed the signature could be falsified; in such cases, identifying the genuine author is critical for security and intellectual property rights (Khedkar et al. 2018). Stylometry is crucial in this regard. Every person has a distinct prose technique, and the measurement of such a style is known as Stylometry. Several instinctual patterns are included by the author when writing; these qualities have gone undetected until now, however, can serve as a significant influence in the reliable and swift recognition of content creators (Iyer and Rose 2019).

In this regard, different authors have proposed their work such as (Alonso-Fernandez et al. 2021; Anwar, Bajwa, and Ramzan 2019; D. Pavelec et al. 2009) have proposed their work in the field of stylometry. With invent of digitalization, the domain has seen a boost since in traditional writing authors can be identified through their handwriting and other procedures. However, in the digital world identifying an author is a difficult task. The stylometry can be helpful in many domains besides proprietary issues. For instance, (Alonso-Fernandez et al. 2021) work proposed the utilization of stylometry in forensics on the Twitter dataset. The author claimed that fake and false tweets have been a severe crime and come under cybersecurity so, identifying the author of the tweet is essential for cybersecurity crime agencies to identify the real culprit. For this purpose, the author proposed a novel framework for security agencies. Our work is quite similar to say study. However, the novelty of this research is firstly the targeted domain which is the copyright proprietary domain for the tweets, and secondly, the proposed methodology that has been based on the Statistical Analysis Systems (SAS). We compared four (04) machine learning algorithms for the relating authors to specific tweets on the given dataset. The set of algorithms has both regression and classification models. Our results showed that we have achieved reasonable improvements in terms of accuracy and our models can be utilized for future applications with reliable results.

Problem Statement

The writing style of a writer is supposed to remain the same for the author. So, stylometry is the study of different styles of writing with the aim of finding the authorship of the content or identifying the author (Anwar, Bajwa, and Ramzan 2019). As technology has progressed in the world of computer science and digitalization has immerged, some problems have also been witnessed. Such technological advancements have resulted in cybercrimes and cyber-attacks. Any digital document with a fake author can be involved in forgery and other digital crimes. There can also be a hacking problem or transfer of data. Digital documents with fake authors can influence many unethical issues in form of writing. So, the identification of the author is very important. Forensic control has to be done to identify the author with the support of artificial intelligence (AI). This situation has been a problem and should be addressed using stylometry and linguistic features to identify the authors with the use of a machine learning approach to avoid frauds, hoaxes and deception in writing style (Afroz, Brennan, and Greenstadt 2012).

Background of the Problem

Digitalization and advancement in technology have caused conflicts of authorship over the years. Some texts or contents require the authorship of the writer because of their importance. For example, some legal digital documents or lawsuits, or any business-related documents have to be owned by someone. There has been a history of writing harassing and threatening notes or letters (Daniel Pavelec, Justino, and Oliveira 2007). So, the identification of such writers should be known to avoid any criminal activity or legal activity. Hence, the investigation of such crimes requires the author's identification.

Business understanding of the Problem

In this first step of CRISP-DM, problem identification is made. In this case, the causes and need for writer identification with stylometry are done. Many digital documents require writer identification to reduce the impact of fraud and forgery. The stylometry or style of the writer has

to be detected. So, when the problem or goal is simplified then the machine learning approach can be used for data mining with this methodology to find out the authorization using stylometry. In other words, the objectives of the project are understood which is then transformed into a data mining problem definition, and finally, a project plan is designed.

Data understanding

In the data understanding process of CRISP-DM, data collection is first made. The data collection includes data verification, data description, and explanation of the data. Sub-sets are made to form the hypotheses for the hidden information (Wirth 2000). In this case, the writer's identification is hidden that is to be found using this method.

The biggest advantage of using this methodology is its low cost and especially its main objective is to evaluate large data mining projects. Moreover, CRISP-DM can be repeated, it is reliable, easy to use, and quick. Furthermore, this methodology works in six phases with a total of 24 tasks and outputs (Plotnikova, Dumas, and Milani 2019). Before data mining, business and data understanding must be known.

Project Definition and Goals

1. To use the stylometry in linguistics to identify the author of the content, which is the style adopted by the writer while writing (Daniel Pavelec, Justino, and Oliveira 2007).
2. To use a machine learning approach to find the legitimate author to avoid frauds and other cybercrimes, as authors with different and fake names do fraudulent activities on the internet using digital documents.
3. To use SAS-supported algorithms in using machine learning framework with help of CRISP-DM methodology. Moreover, this study will review the previous literature on writer identification using a machine learning approach, and find flaws or gaps left behind in the literature to fill these gaps in this approach for more effective writer identification.

Goals of this Project

1. To study the purpose of writer identification.

2. To understand stylometry and its use in in writer's identification.
3. To understand the role of machine learning in stylometry.
4. To implement and compare the best machine learning model on the given dataset of author's profiling

Research Questions

RQ 1. What is the purpose of writer identification?

RQ 2. How stylometry can be used to identify the writer?

RQ 3. What is the purpose of machine-based learning to find the writer's identity?

RQ 4. Which Machine learning model fits the best?

Limitations of the Project

1. The scope of this project is limited to the data acquired from the given dataset. Also, the experiments are conducted with five authors for the purpose of balanced data and reduced complexity.
2. The sentences data in the dataset related to every author has variant nature. Therefore, only textual data is selected that reduces the context of this project to 25 common sentences used by each author.
3. The evaluation is mainly measured with the accuracy scores for each model. There is no formal method given to calculate MAE or RMSE for textual data. However, other methods like ROUGE or ROUGE-2 usually applied for textual domain requires in depth domain information and understanding.

Literature Review

Linguistic identification and Stylometry

In the history of linguistics and writing, the identification of the author is very important. In this regard, linguistic and stylistic investigation for the identification of the author has been done since the nineteenth century. Previous literature has plenty of content available on a linguistic and stylistic investigation to identify the author, which has resulted in the formation of forensic linguistics that includes the analysis of the authorship for forensic purposes (D. Pavelec et al. 2009). The need for a writer's investigation for forensic purposes has gained so much attention due to its requirement in the criminal laws that contain harassment or any other threatening notes, and then it is needed in civil law that involves the copyright regulations and estate problems. Most importantly in today's world of innovation and technological advancements, computer security has become a necessity because of writer's identification in the mining of emails. This digital writing world has witnessed many crimes, which have to be identified (Daniel Pavelec, Justino, and Oliveira 2007).

The document source and writer's identification in digital documents must be extracted to stop such sort of digital crime. So, the legitimacy of the document written by the writer through the computer keyboard can be identified. Many ideas and ways have been developed in finding the legitimate author of the digital document such as the style in which the writer has produced the digital document. Style is an essential component of linguistics and is used in different ways by different authors. This study and identification of the style in writing are known as stylometry (Daniel Pavelec, Justino, and Oliveira 2007). Stylometry is the term basically used for the recovery of important features of the document using the style of writing in order to find the identity of the author. The identification of the author such as gender, native language, and if the writer has a problem with dementia could also do with the use of statistical techniques and analysis (June et al. 2020). Many social scientists, marketers, and analysts use statistical techniques to directly know the details of the author.

Lots of characteristics are there that distinguish the writing style or stylometry from others. It includes the use of vocabulary, grammatical errors, spelling mistakes, and repetition of words judged by the frequency of words used and length of sentences written (Daniel Pavelec, Justino, and Oliveira 2007). There has been a problem with the identification of the author. For this purpose, a writer-specific model or personal model was suggested to be used. It has two kinds of classes i.e. w1 and w2. W1 represents authorship while w2 shows the forgery. However, this model has some drawbacks, especially the issue of including new authors every time, and a large number of writing samples is also required, which decreases the reliability of this model. On another hand, with the introduction of the forensic document examination approach, the problem of writer identification has been eased down to a certain extent. There was also a model presented for author identification that contained the independent approach of the writer. The Portuguese language was used in this feature of writer identification. This approach counted the words used in conjunction with the fusion strategies with the use of Receiver Operating Characteristics (ROC). After that, a Support Vector Machine (SVM) that analyzed short articles databases has been tested (Daniel Pavelec, Justino, and Oliveira 2007).

Another strategy that has been introduced in recent years for the extraction of information about the author consists of a modern data compression algorithm. Such algorithm is called Prediction by Partial Matching (PPM) and has been widely accepted and used. PPM requires computer-based resources as it uses the latest technology for data storage (D. Pavelec et al. 2009). Such an algorithm has experimented with many tests that contained the workings or documents of the authors. SVM was used in this regard, which ultimately proved that PPM was a good substitute algorithm to detect and identify the author (D. Pavelec et al. 2009).

Forensic stylistics is formed from forensic linguistics in which the statistics are used for author identification. It is based on two basic assumptions to distinguish two authors from each other, these two authors from the same native language cannot write similar to each other. The second assumption is that one writer is not able to write in the same pattern every time he is asked to write. For this, SVM seemed to be the best choice to identify the author. The basic advantage of this machine is its ability to handle high-dimensional data. However, SVM is unable to work in the probabilistic model (Daniel Pavelec, Justino, and Oliveira 2007).

Machine Learning and SAS

Another study that advocated for the machine learning approach for author identification was conducted by (Pearl and Steyvers 2012). This approach is to reduce the number of cybercrimes from digital documentation or forgeries. Machine learning is used in internet search engines to filter out the emails like spam and junk, and sort out the data or emails (Mohammed, Khan, and Bashie 2016). The verification of the writer is important because a criminal-minded person can try to copy the writing style and pattern of other writers. So, authorship deception is identified using this machine learning approach. The said machine learning approach is a type of artificial intelligence (AI) that is used to increase the effectiveness of the software applications to find more accurate results with the help of machine learning algorithms. These algorithms can be programmed according to the mindset of humans. So, different algorithms are made for the performance of different tasks (Anwar, Bajwa, and Ramzan 2019). Machine learning using algorithms allows the working of machines in a better way, as they contain AI. Moreover, Statistical Analysis System (SAS) provides support to machine learning according to the programing set. It is mostly used in data mining and statistics (Mohammed, Khan, and Bashie 2016). So, it can be said that machine learning is the intersection of computer sciences and statistics. Thus, with the help of SAS algorithms are developed that are used for the mining or identification of the authors. The data miners that use SAS in the machine learning approach, provide support in running and using the statistical models like linear and logistic regression analysis (Parsad 2014). SAS is a comprehensive software and can handle multiple problems like complex statistical analysis data mining, data creation, sorting, graphics, etc. along with the support of its three components of a host, portable applications, and data (Parsad 2014).

A study by (Iyer and Rose 2019) also used a machine learning framework for authorship identification. The task was divided into single labeled multi-class text for the categorization and explanation of the features of stylometrics. As a sample, 50 different sample works of authors were collected to check and distinguish between these samples according to the features of stylometry. These features of stylometrics helped in providing more accuracy and reliability in the identification of the authors. Some of the features included the algorithm from LibLINEAR SVM (Iyer and Rose 2019).

Machine-based learning and its Features in stylometry

Another tool used along with stylometric statistical analysis is Writeprint characterization. It covers both stylometric and content features. Among these features, stylometrics has nine features and content has eighty-one features. So, these features are combined together in the machine learning approach (June et al. 2020). This machine has made designed to handle the problem of authorship in which a document is analyzed to find the correct author. To decide whether the document is written by the same author or not, the classifier is the feature of this machine that does the authorship detection of the document. On other hand, such classifiers cannot be made without the assistance of humans accurately from the large set of texts of different authors (Ramyaa, He, and Rasheed 2004).

However, the classifier is developed by making documents for the authors to be tested. The first document to make a classifier is denoted by A1, in which the single randomly selected document for the author is selected. Then it is A2 from where the remaining documents of the authors are collected. The third is X1 which analyze the single randomly chosen targeted document from the different author for the comparison (Pearl and Steyvers 2012). On gathering the said information or the dataset of content from authors, Sparse Multinomial Logistic Regression was used, which helped in identifying the authorship of two different case studies taken for this study.

The growing trend of digitalization has provided many ways for the authors to commit any criminal activity. The algorithms are going complicated along with time and technology which makes it difficult to find the author. So, stylometry can be introduced with AI technology. This AI technology could help to automatically read the document and detect the text and linguistics, and thus the author (Ramyaa, He, and Rasheed 2004). For this AI to work properly, some features of the author and text should be known. Again the style of two different authors writing would distinguish the authors from each other. Style may be more general rather specific. So, this study has used different forms of styles such as type-token ratio that highlight the vocabulary level of the author, in which high ratios would indicate variety in vocabulary use and also repetition of the words. Then it is mean word and sentence length, in which long sentences are written with some planning (Ramyaa, He, and Rasheed 2004). Standard deviation measurement of the sentence length will provide variation in the sentence length. Similarly, the writing of paragraphs and the

length of chapters provide meaningful information. Moreover, the amount of the use of commas, semicolons, quotations, exclamation marks per tokens, etc. provides the style of the writing by the author (Ramyaa, He, and Rasheed 2004).

(Anwar, Bajwa, and Ramzan 2019) also used a machine learning approach for the author identification. The first step for the experiment in this particular study was to gather the dataset in both English and Urdu languages. Dataset was taken from PAN12 and UrduCorpus and used author representation-specific documents. The documents from different authors were collected in one file i.e. one file for one author that contained different work samples. After that, document preprocessing was done for the review of the content from the authors. It was done to know about the style of using languages like grammar, spelling, sentences, phrases, abbreviations, sentence structuring, etc. Natural language Toolkit (NLTK) was applied as a tokenization process that changed the sentences into smaller words (Anwar, Bajwa, and Ramzan 2019). Similarly, N-gram generation was used that involved the grouping or compiling of words in n length. Moreover, the use of algorithms such as LDA was done in the experiment.

Furthermore, this study by (Ramyaa, He, and Rasheed 2004) also adopted decision trees as a tool for highlighting the stylometry of the writer. Under this, two decision trees are made that represent the style used in the texts. These are used for experimental purposes with the support of the basic ID3 algorithm of Quinlan. Moreover, neural networks are another powerful tool in machine learning techniques for stylometry. These consist of complex non-linear modeling equations and act as strong matching tools. Apart from the complex nature of these networks, they can be used as stylometric identifiers due to their nature of taking inputs simultaneously. For experiments, a statistical technique like Neuroshell made by Ward System Group Inc. was introduced (Ramyaa, He, and Rasheed 2004).

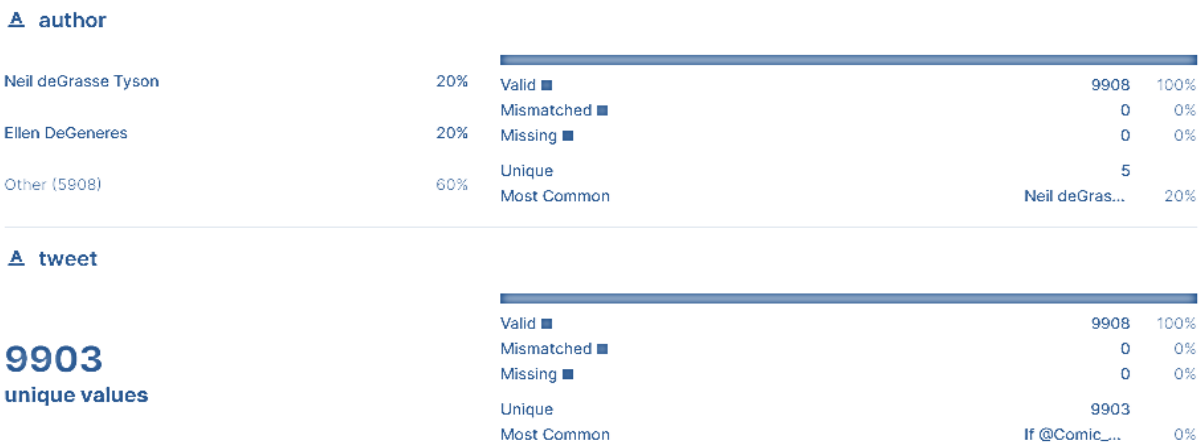
Methodology

Different methods especially the machine learning approach with the support of SAS are needed to extract the exact identity of the author. SAS provides support for the large data to be estimated using statistical models. The finding of authors using SAS programming and machine learning method is a process of data mining and statistics. Sometimes data mining and statistics may come in combination to generate a SAS-based model (Mohammed, Khan, and Bashie 2016). Although there is no specific data mining approach to handle the projects, however, for data mining of stylometry CRISP-DM (Cross Industry Standard Process for Data Mining) process seems a better methodology (Wirth 2000). The steps of the CRISP-DM are explained in the subsequent sections of this chapter.

Sources of Data

The dataset which is being used in this research is taken from the Kaggle which was already extracted and widely used for the research. The dataset is taken from the following link. https://www.kaggle.com/datasets/azimulh/tweets-data-for-authorship-attribution-modelling.

Dataset Overview



Analysis

A. Data preparation

Data preparation is used to make the final dataset for the analysis. The collected in its appropriate form is fed into the machine learning algorithm in the given sequence of steps:

i. Cleaning of the data, in which the null values are removed and outliers are found to be cleaned as well.
ii. Then the data is transformed where the numerical attributes are normalized (Wirth 2000).
iii. If there is more than one dataset, then all the data is integrated.

Modelling –

Different techniques of data modelling are selected and applied. Many techniques are present that are closely related to the data preparation. The models will be developed to find the authorization of writing using stylometry. This dataset will be used and added to the decision tree machine learning algorithm, which will then highlight the stylometry of different authors.

Evaluation -

The evaluation of data has to be done effectively before the calculations or final output. A good evaluation helps provide accurate results. Every step of the methodology should be followed for the right model setting to achieve the business objective (Wirth 2000).

Deployment -

This is the final step of the methodology or the final output in which the tested model is deployed as a part of an application. Usually, an application is built independently of the model keeping in mind the ease of use, performance, and security metrics.

B. Dataset Description

In this paper, the Twitter dataset has been used which was already generated and available on the Kaggle Datasource website. The dataset is available in .csv format. There are two variables within the dataset, author name, and tweets. Both dataset types are textual and stored as string data within

the tool. The primary statistics of the dataset is showing that the total num observations are more than 10000 and the total number of authors is five (5).

C. CRIPS Model

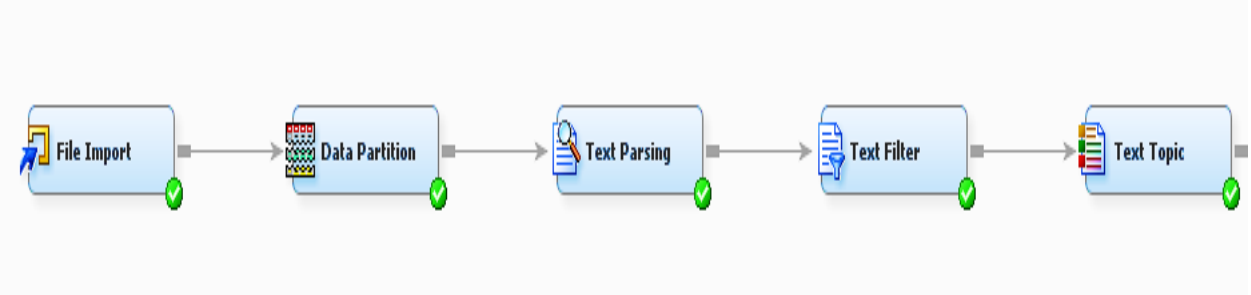
In this research paper, the CRIPS model is applied using SAS Enterprise Miner for Text Mining and Prediction of Author. SAS Enterprise Miner is supporting machine learning algorithms that are helpful for tokenization and analysis. The CRIPS-DM model consists of six different stages in which “Data Understanding”, “Data Preparation”, “Modeling”, “Evaluation”, “Deployment, and “Business Understanding”.

Hypotheses

A. Data Preparation

After importing the dataset within the SAS Enterprise Minner, the dataset is prepared for modeling. For the preparation of the dataset, the “Data Input” node helps to check the missing values within the dataset, and whether missing values are present in the dataset or not. In the next step, the feature extraction method is applied using “Text Parsing”, so that the dataset should be clear from all unnecessary words, vocabulary, and punctuation. Then features are extracted using the “Topic Mining” node. Topic mining again helps to purify the dataset. Then using categorize node, features are again extracted using matching words. At the end match of keywords is performed with the writer, and words are identified with the chosen topics to find the relevance score.

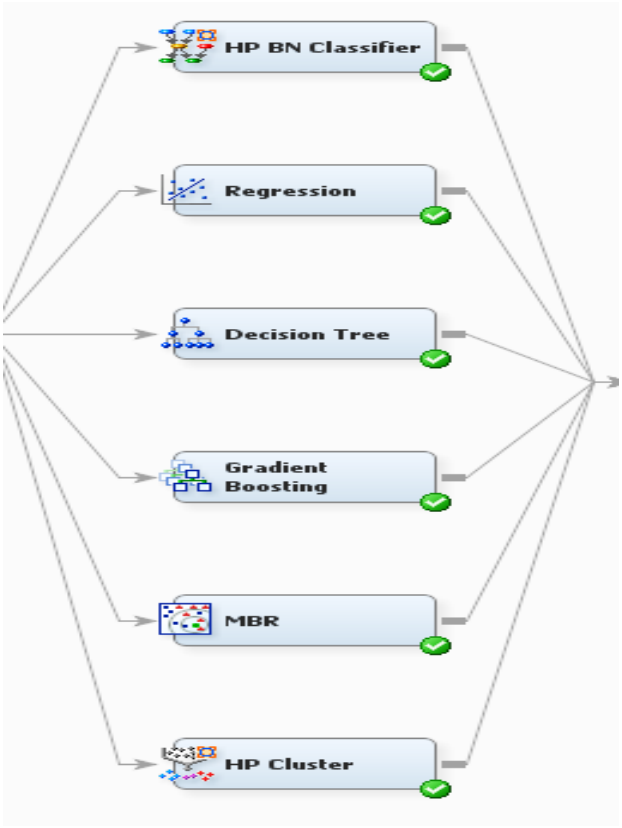
Process flow Nodes



B. Modeling

After the data preparation step, models are designed using the built-in features of the SAS Enterprise Minter. The following models are applied to test the dataset. The first BN classifier is used to classify the data according to the writer. After the Regression model is used. The decision tree helps to find the most correlated scores within the dataset. Gradian boost help to create a decision tree to find the correlation between the tweets and the writer. In the end, MBR and HP Clustering have been applied to test the data.

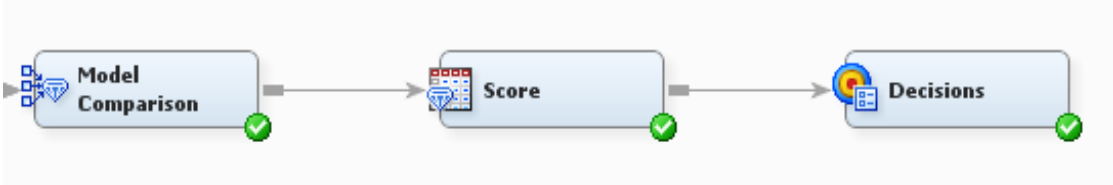
Models Implementation



C. Evaluation

By following the CRIPS-DM model, at this stage, all models' output is compared and is Figured out which one is best for implementation. The models are compared using Precision Rate, Mean Error Rate, Accuracy, Mean Square Error, and Valid data statistics. A comparison node is applied.

Results generation from Models

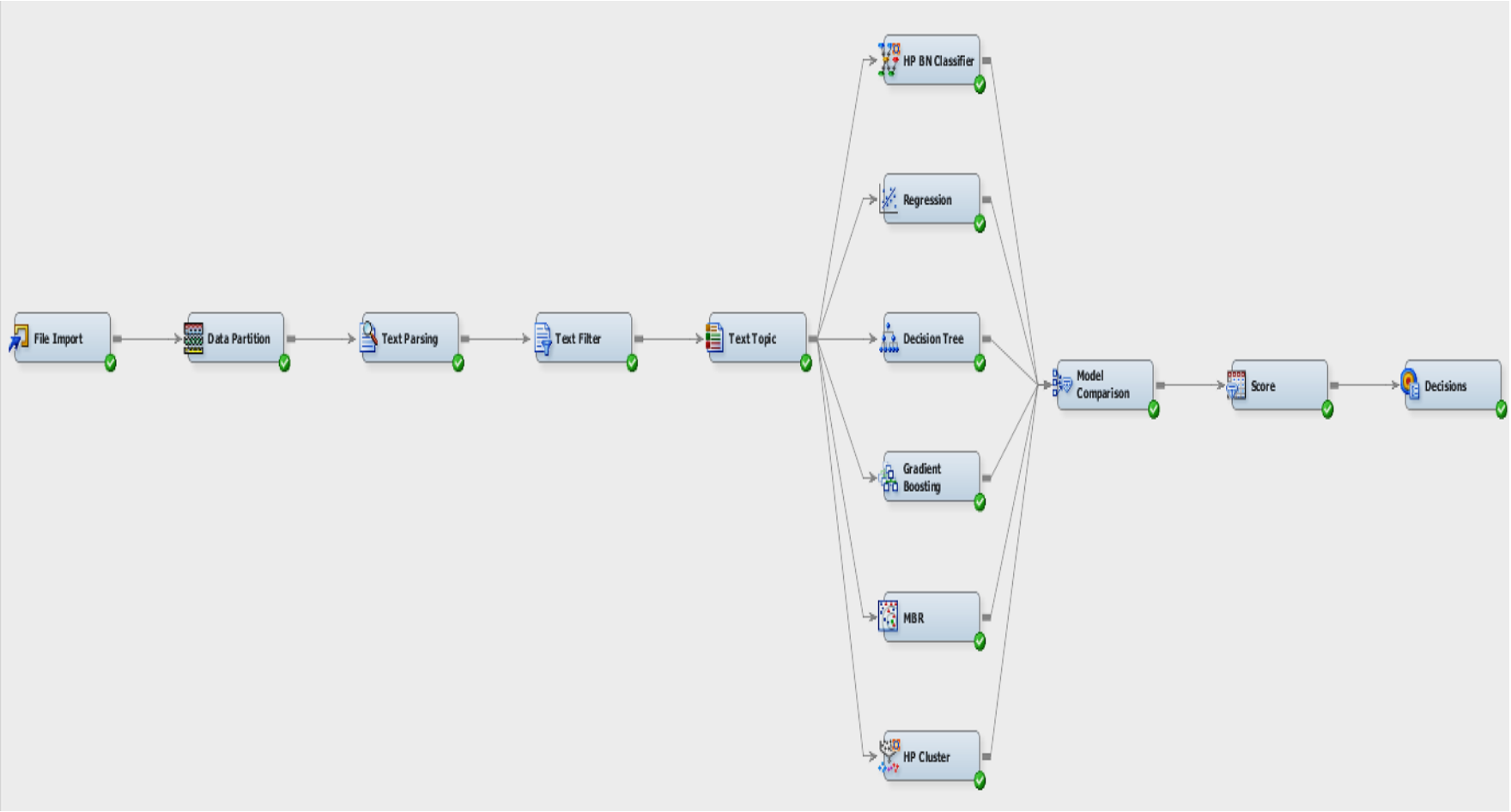


Once the models are compared their results are stored using the score node. These scores can be utilized for generating outcomes and discussing the reliability of the conducted research.

D. Complete Model

Below figure integrates all the three sub parts of the proposed methodology to depict the overall schema of the project. It represents all the phases of Data preparation, Modelling and Evaluation.

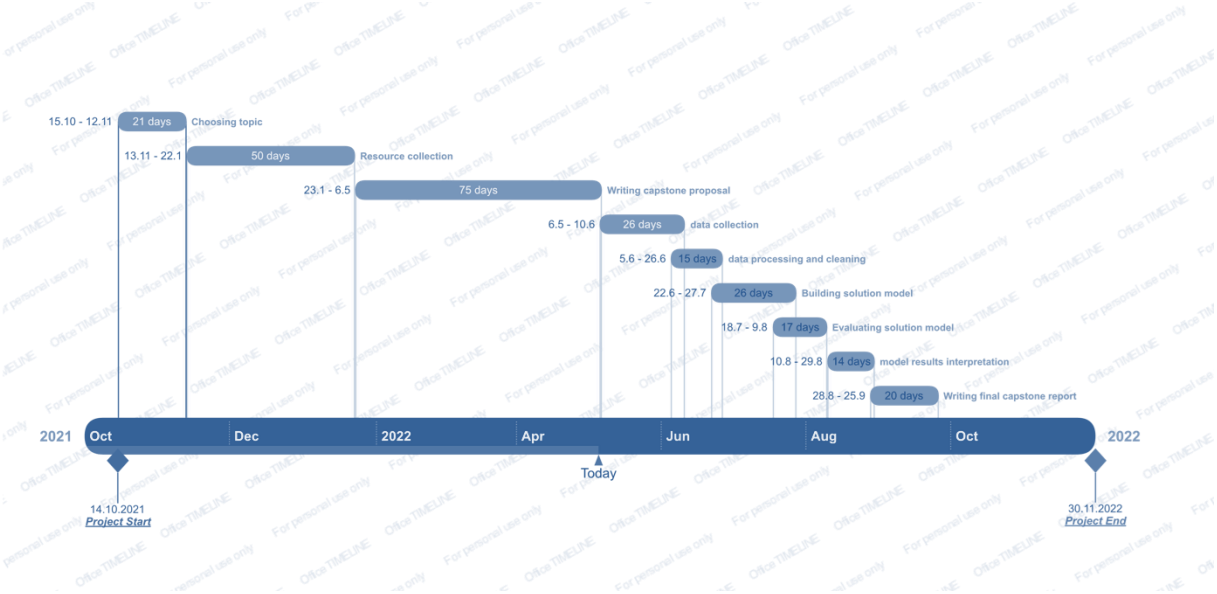
Complete Model



Project Deliverables

The main objective of this project is to find the writer's identity using stylometry with the support of a machine learning approach or algorithms and the most accurate machine learning model to predict the correct author. The first project deliverables included the research proposal. After that first draft of the project has been submitted upon the acceptance of the proposal. Similarly, the project has been completed in milestones to sort out any problem during the entire timeline of the final project. This document is the final thesis report submitted after the completion of the project.

Project Timeline



Project resources and budget estimate

There will be no cost since the tool is provided by my employer.

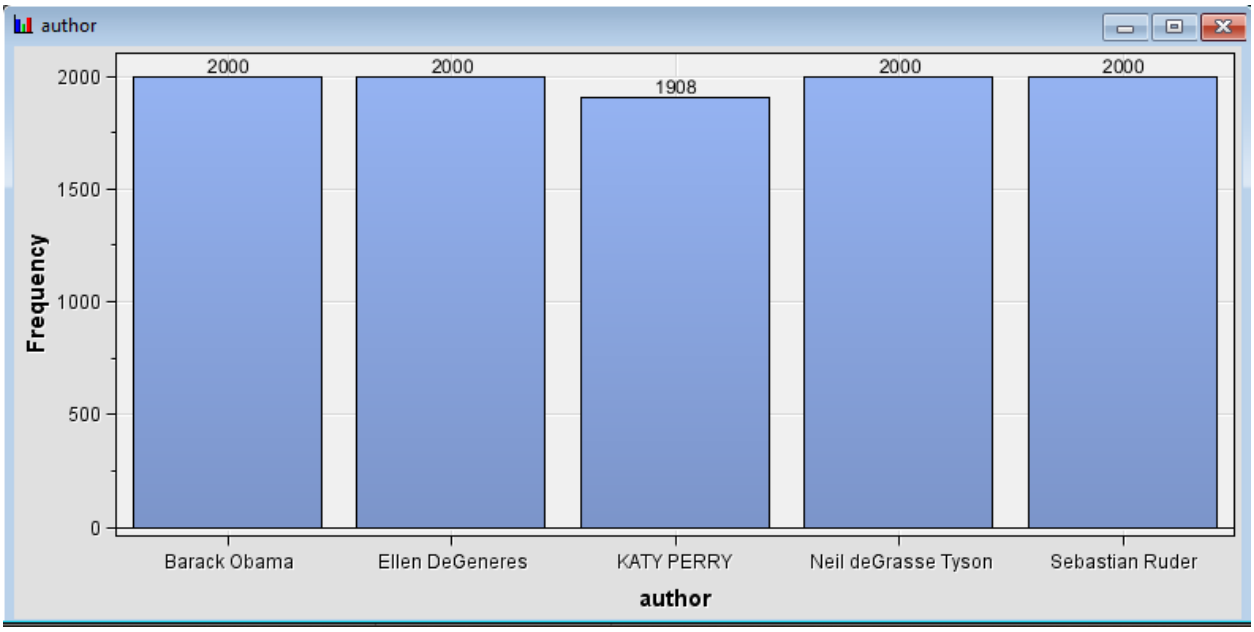
Result and Discussion

In this part, the results of the conducted research are discussed.

Descriptive Statistics

Descriptive statistics are used to understand the dataset and take an overview of it. It tells us the different basic statistics about the variable. In the first variable "Author" basic statistics show that there are (Count = 5) authors in the dataset. The total tweets tweet by the (Barack Obama = 2000, Ellen DeGeneres = 2000, Katy Perry = 1908, Neil DeGrasse Tyson = 2000, Sebastian Ruder = 2000). The total number of Tweets = 9908 where many tweets belong to Barak Obama.

Frequency Chart of Authors

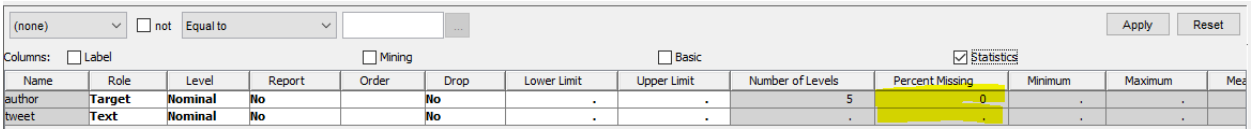


A. Checking Null Values

The dataset should be clean and in proper format before doing any kind of analysis. Missing values are much important in doing analysis. Dealing with missing values is important because, if missing

values exist within the dataset, the output is entirely different. While performing imputation with missing values. All missing values have been removed within the dataset using “Import Node”.

Null Values output



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit	Number of Levels	Percent Missing	Minimum	Maximum	Mea
author	Target	Nominal	No		No	.	.	5	0	.	.	
tweet	Text	Nominal	No		No	.	.	.	.	.	.	

B. Data Partition

To apply the model, it must be trained using a training dataset. For this purpose, a partition node is added to the workplace to divide the dataset into two parts. Training and Testing. The training dataset seeds = 80% of the whole data and the testing dataset consist of 20% seeds of the whole data. The output is given as under.

C. Summary Statistics for Class Targets

Table 1- Data=DATA

Numeric		Frequency		
Variable	**Value/ Formatted Value**	**Count**	**Percent**	**Label**
author	Barack Obama	2000	20.1857	author
author	Ellen DeGeneres	2000	20.1857	author
author	KATY PERRY	1908	19.2572	author
author	Neil deGrasse Tyson	2000	20.1857	author
author	Sebastian Ruder	2000	20.1857	author

Table 2- Data=TEST

Numeric		Frequency		
Variable	**Value/ Formatted Value**	**Count**	**Percent**	**Label**
author	Barack Obama	600	20.1545	author
author	Ellen DeGeneres	601	20.1881	author
author	KATY PERRY	574	19.2812	author
author	Neil deGrasse Tyson	601	20.1881	author
author	Sebastian Ruder	601	20.1881	author

Table 3- Data=TRAIN

Numeric		Frequency		
Variable	**Value/ Formatted Value**	**Count**	**Percent**	**Label**
author	Barack Obama	800	20.1969	author
author	Ellen DeGeneres	800	20.1969	author
author	KATY PERRY	762	19.2376	author
author	Neil deGrasse Tyson	799	20.1717	author
author	Sebastian Ruder	800	20.1969	author

Table 4- Data=VALIDATE

Numeric		Frequency		
Variable	**Value/ Formatted Value**	**Count**	**Percent**	**Label**
author	Barack Obama	600	20.2020	author
author	Ellen DeGeneres	599	20.1684	author
author	KATY PERRY	572	19.2593	author
author	Neil deGrasse Tyson	600	20.2020	author
author	Sebastian Ruder	599	20.1684	author

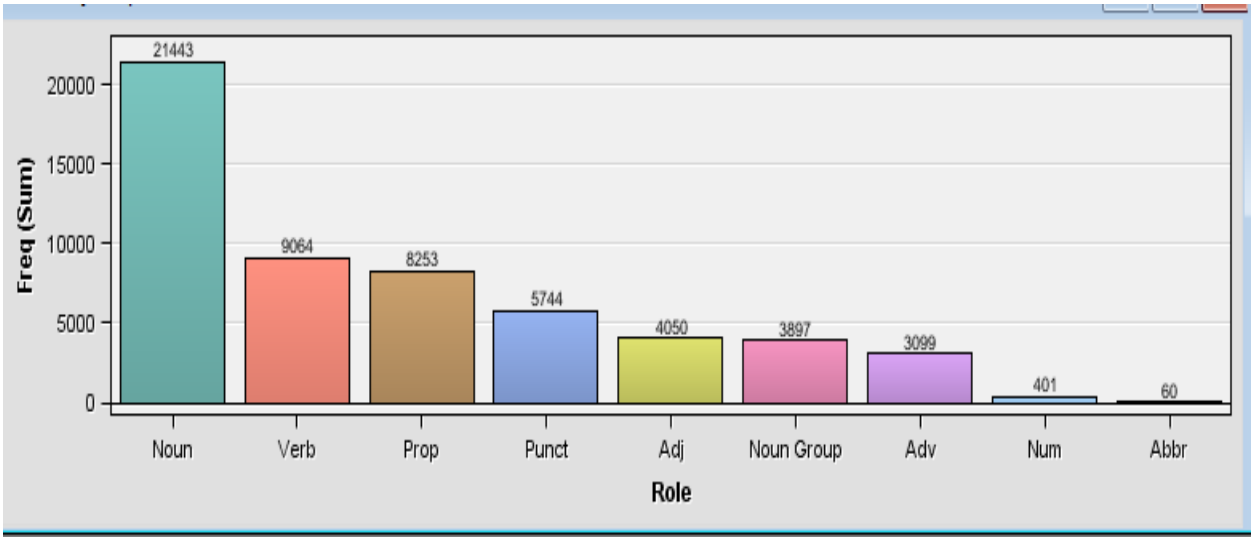
The output given above is showing the data partition. The dataset is divided into three subsets. Training (Total Observations = 3268), Testing (Total Observations = 3000) and validation (Total number of observations = 3000).

Data Transformation

A. Text Parsing

SAS Enterprise miner, assist in text mining using text parsing node. This node vectorized the text data. It also helps to quantify information about the text terms used in the dataset. It counts all the similar words and finds their frequency which helps in analysis. By analyzing the parsed data, we are able to select the specific term for analysis and also can remove all unnecessary terms from the data. The output of this node is as under. In the given figure, the output is showing that almost (Noun = 21443, Verbs = 9064, Prop = 8243, Punct = 5744, Adj = 4050, Non-Group = 3897, Adj = 3099, Number = 401, Abbreviation = 60). The statistics is showing that, our dataset is mostly consist with the nouns. By using this vactorizated text, text will be recognized

Text Parsing Frequency Chart



B. Text Filter

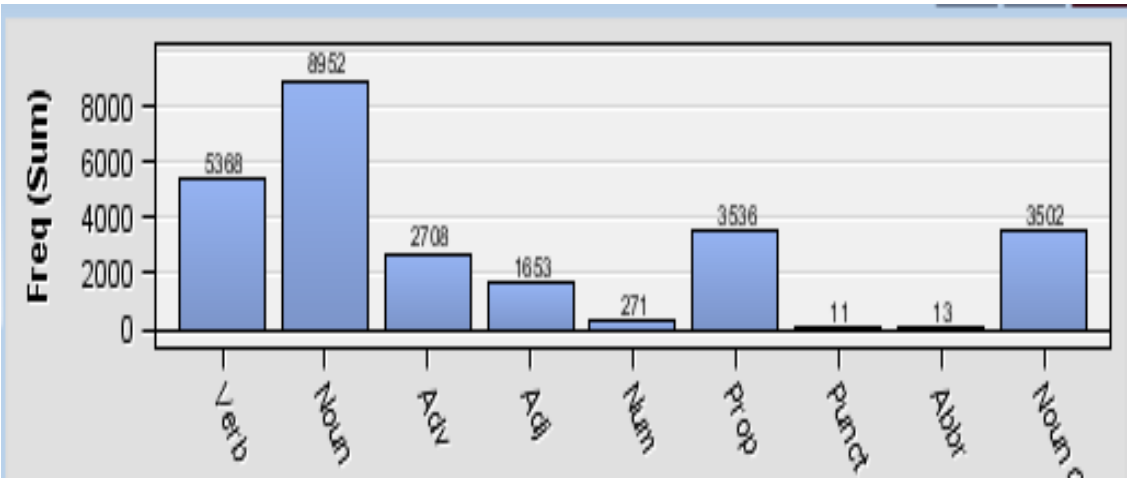
Once text parsing is applied in the next step, all unnecessary text should be removed from the data so that clean data is used for the model. In the below figure, the list of removed words is given. Almost (mean = 5938, Noun = 8952, Adv = 2708, Num = 271, Punct = 11, Abbreviation = 13 and None group words = 3502 has been removed from the data.

Text parsing and dropping all unnecessary words.

Terms

Term	Role	Attribute	Status	Weight	Imported Frequency	Freq	Number of Imported Documents	# Docs	Rank	Keep	Parent/Child Status	Parent ID	OLDROLE	OLDATTRIBUTE	Imported Parent/Child Status
+ be	Verb	Alpha	Drop	0.000	1416	1436	1195	1206	2N	+		20023	Verb	Alpha	+
+ have	Verb	Alpha	Drop	0.000	397	401	380	384	7N	+		19898	Verb	Alpha	+
+ s	Noun	Alpha	Drop	0.000	424	425	381	382	8N	+		19833	Noun	Alpha	
+ not	Adv	Alpha	Drop	0.000	373	374	347	348	10N	+		19916	Adv	Alpha	
+ do	Verb	Alpha	Drop	0.000	306	322	280	294	11N	+		20078	Verb	Alpha	+
+ get	Verb	Alpha	Drop	0.000	240	241	228	229	15N	+		19837	Verb	Alpha	+

Frequency Chart of words



C. Text Topics

After the data cleaning step, the next step is to select the topic which needs to find out the relationship between the author and their text.

List of Selected Topics

Topic ID	Topic	# Docs
1	ðÿ,¼,ƒ,¥,¦	367
2	president,obama,â,+live,america	265
3	™,itâ,youâ,€,gonna	618
4	rt,seb_ruder,¦,katyperry,nlpdublin	417
5	¤,ï,¸,â,ðÿ	190
6	amp,startalkradio,+post,â,itunes	247
7	+happy,+birthday,happy birthday,+hope,¦	134
8	¦,€,+model,data,jeremyphoward	631
9	actonclimate,climate,climate change,fight,+denier	96
10	americanidol,¼,+love,™,ðÿœ	80
11	americanidol,˜,™,¼,¯	80
12	earth,+sun,+moon,amp,+space	113
13	senate,doyourjob,+leader,hearing,garland	101
14	nlp,¦,learning,seb_ruder,news	147
15	˜,ðÿ,ðÿž,â,katyperry	211
16	iâ,™,€,¦,glad	423
17	+paper,¦,jeremyphoward,rt,deeplearning	88
18	gameofgames,™,knoworgo,¦,+app.	116
19	president,obama,ofa,+speak,+tune	346
20	+show,¦,ellentube,thankssponsor,+love	137
21	ðÿž,¯,âœ,¤,¦	127
22	â,whitehouse,potus,rt,€	375
23	health,+american,care,getcovered,obamacare	207
24	¸,ï,¯,âœ,ðÿ	166
25	+love,+post,+blog,learning,blog post	302

The figure given above is showing the list of 25 selected topics for analysis. These topis have no misclassification and will be used in the modeling of data.

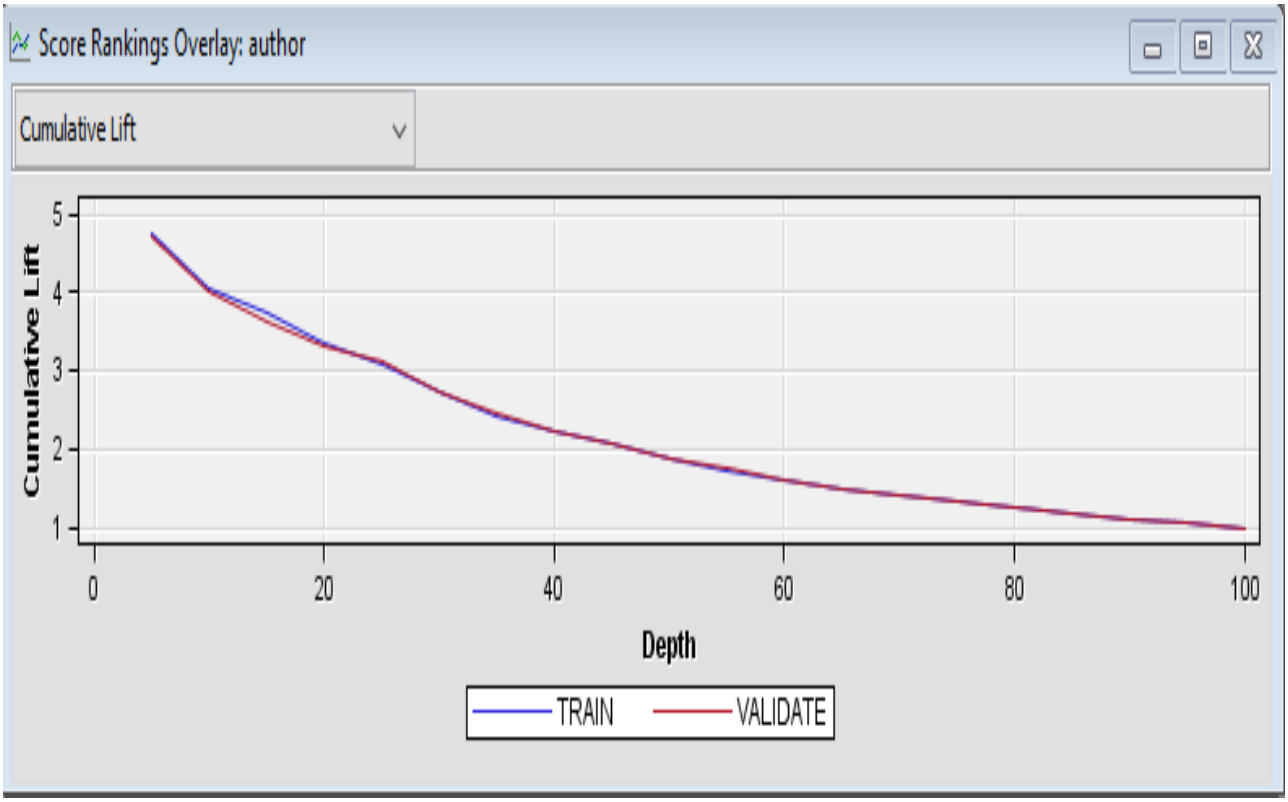
Models Implementation

In this step, all suitable models are applied to the data to identify the author. For this purpose, following models are used.

A. HP BN Classifier

HP BN Classifier helps to classify the keywords based on the categories. The categories in the given dataset are applied as “Author”. The below graph in figure shows that the model has no overfitting and the line is smoothly moving towards 100% for both train and validation accuracy. In simpler words, there is no difference or gap between both curves which means that model is fit and classified.

Training, Testing, Validation Model Curve



B. Decision Tree

A decision tree helps to classify the text based on the parameters and target variables. By applying this model following output is generated which is classifying the text with the author.

Decision Tree Output

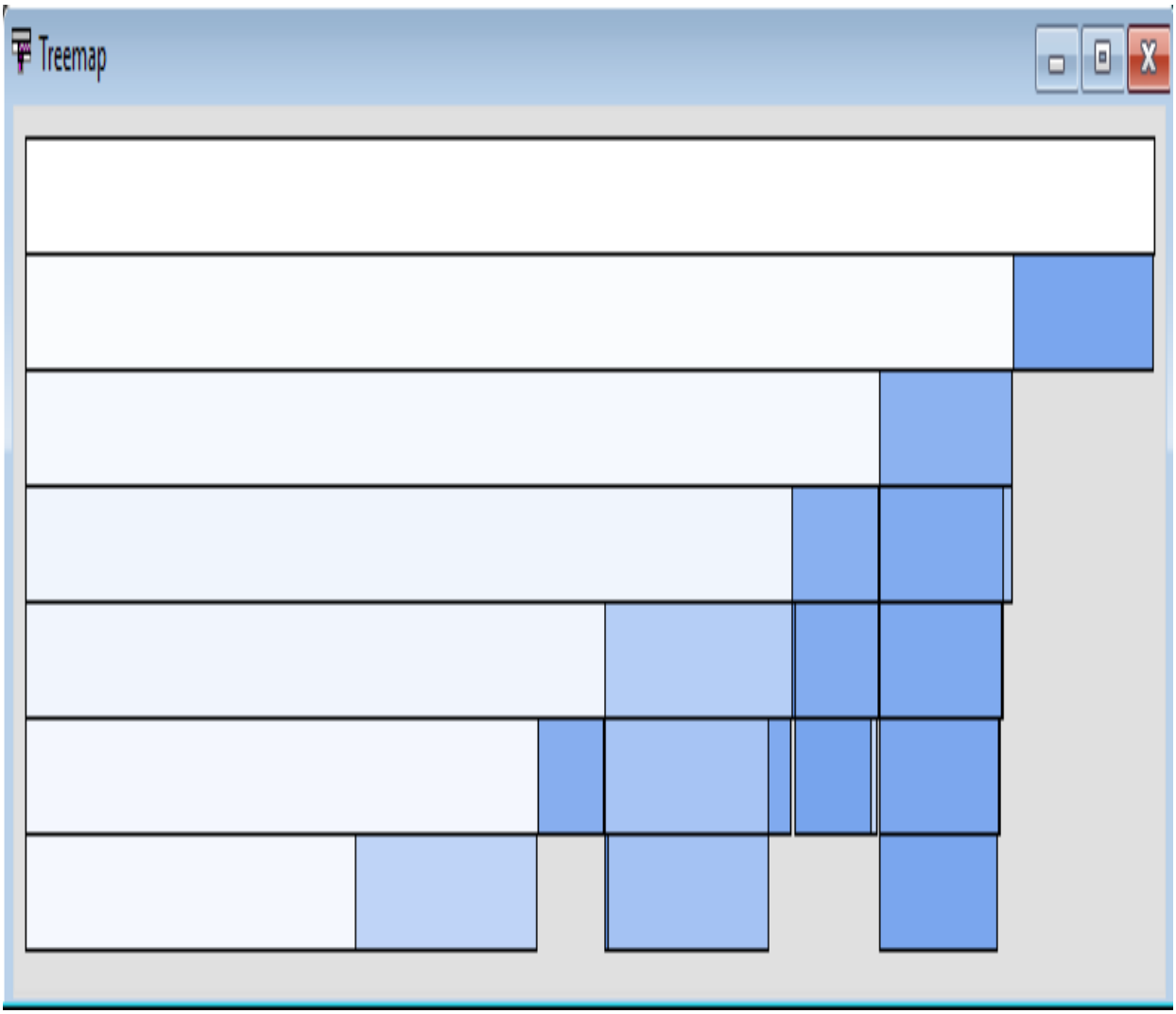
Classification Table

Data Role=TRAIN Target Variable=author Target Label=author

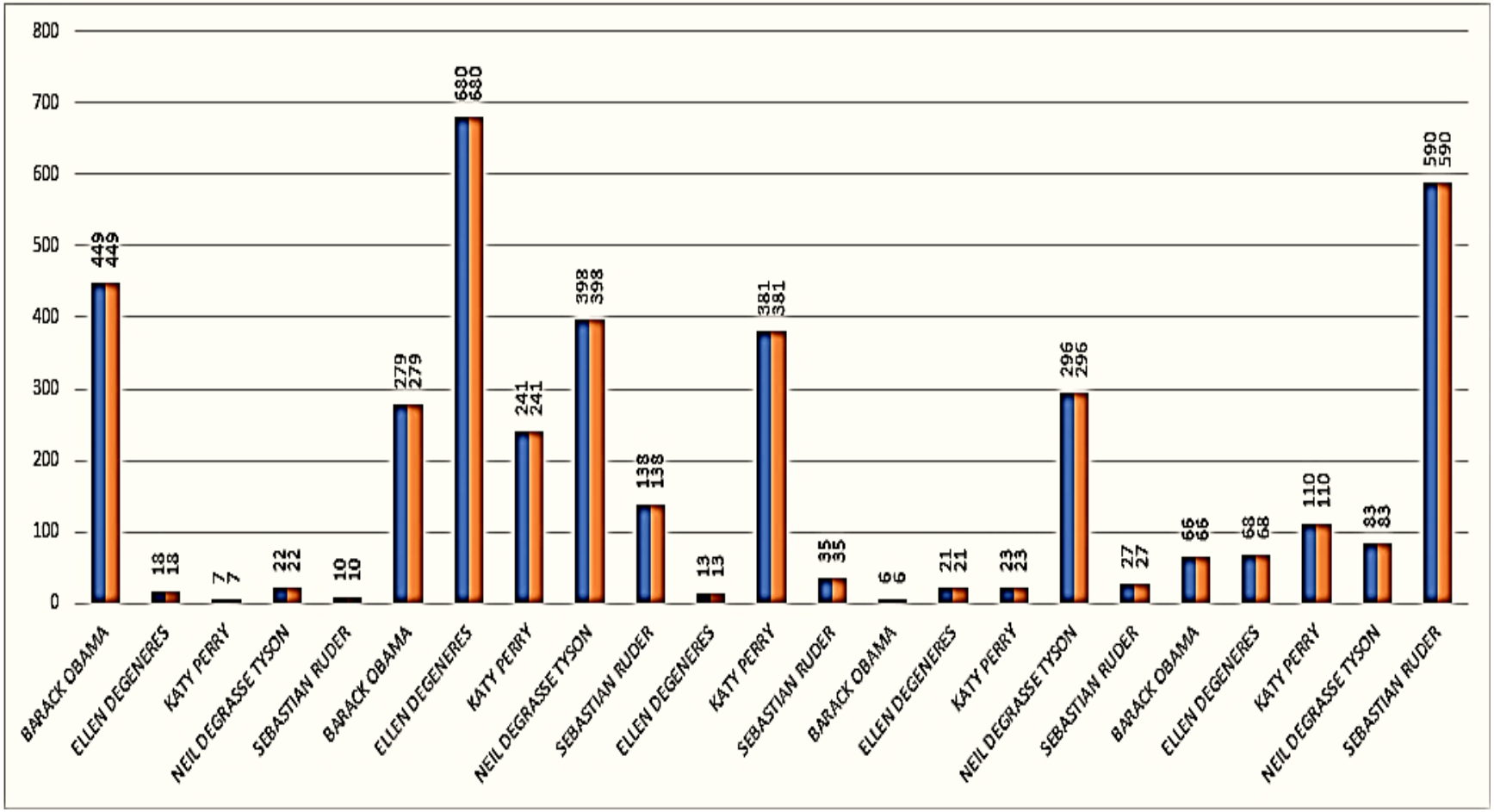
Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
BARACK OBAMA	BARACK OBAMA	88.7352	56.1250	449	11.3355
ELLEN DEGENERES	BARACK OBAMA	3.5573	2.2500	18	0.4544
KATY PERRY	BARACK OBAMA	1.3834	0.9186	7	0.1767
NEIL DEGRASSE TYSON	BARACK OBAMA	4.3478	2.7534	22	0.5554
SEBASTIAN RUDER	BARACK OBAMA	1.9763	1.2500	10	0.2525
BARACK OBAMA	ELLEN DEGENERES	16.0714	34.8750	279	7.0437
ELLEN DEGENERES	ELLEN DEGENERES	39.1705	85.0000	680	17.1674
KATY PERRY	ELLEN DEGENERES	13.8825	31.6273	241	6.0843
NEIL DEGRASSE TYSON	ELLEN DEGENERES	22.9263	49.8123	398	10.0480
SEBASTIAN RUDER	ELLEN DEGENERES	7.9493	17.2500	138	3.4840
ELLEN DEGENERES	KATY PERRY	3.0303	1.6250	13	0.3282
KATY PERRY	KATY PERRY	88.8112	50.0000	381	9.6188
SEBASTIAN RUDER	KATY PERRY	8.1585	4.3750	35	0.8836
BARACK OBAMA	NEIL DEGRASSE TYSON	1.6086	0.7500	6	0.1515
ELLEN DEGENERES	NEIL DEGRASSE TYSON	5.6300	2.6250	21	0.5302
KATY PERRY	NEIL DEGRASSE TYSON	6.1662	3.0184	23	0.5807
NEIL DEGRASSE TYSON	NEIL DEGRASSE TYSON	79.3566	37.0463	296	7.4729
SEBASTIAN RUDER	NEIL DEGRASSE TYSON	7.2386	3.3750	27	0.6816
BARACK OBAMA	SEBASTIAN RUDER	7.1974	8.2500	66	1.6662
ELLEN DEGENERES	SEBASTIAN RUDER	7.4155	8.5000	68	1.7167
KATY PERRY	SEBASTIAN RUDER	11.9956	14.4357	110	2.7771
NEIL DEGRASSE TYSON	SEBASTIAN RUDER	9.0513	10.3880	83	2.0954
SEBASTIAN RUDER	SEBASTIAN RUDER	64.3402	73.7500	590	14.8952

The model given in figure is showing the output. In the first column the name of the authors is used as the target variable while in the second column when the model is trained and tested, it understands the textual data, based on the scores, and classifies correctly. If we look deeply within the output (Barack Obama = Barak Obama, Ellen Degrasse = Barak Obama) it means that tweet was written by Ellen Degrasse using Barak Obama's writing style.

Tree Map of Decision Tree

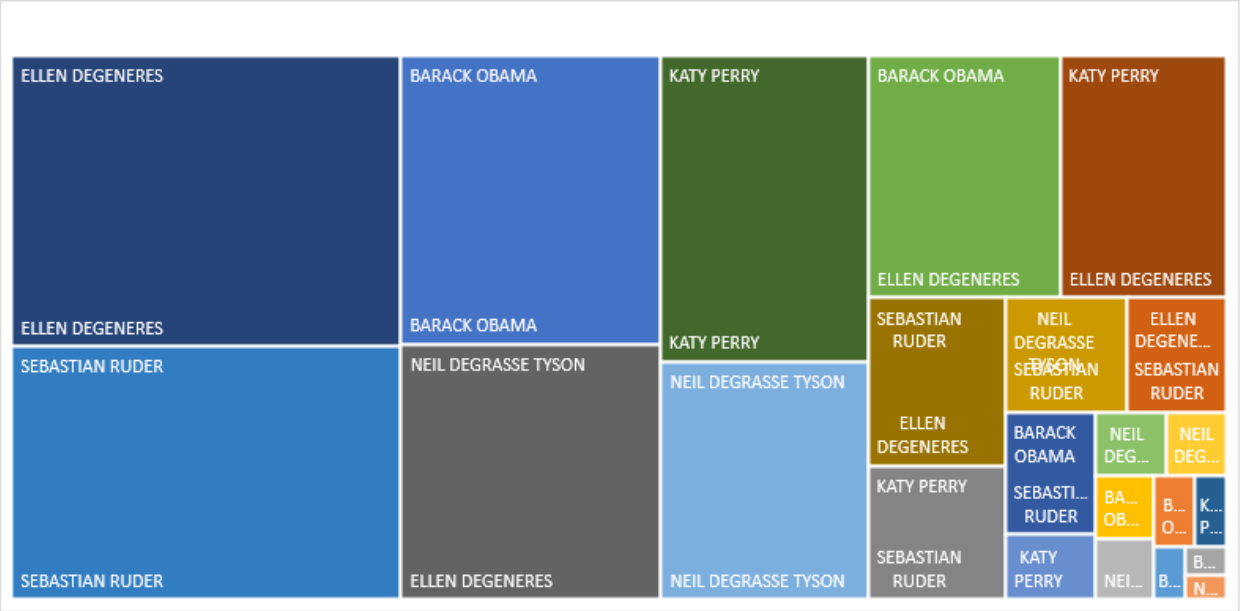


Comparative Author Chart



:

Heat Model



The models given in above figures are the visualization of the Decision Tree.

C. Gradient Boosting

The Gradient Boosting algorithm is another classification algorithm that classifies terms based on scores. This model uses the Greedy Function Approximation method for the classification of the data.

It searches the optimal partition of the data and then combined them to find out the best fit goodness rate. After evaluation, it generates a predictive model based on scores and resemblance. On applying the model, the following output is generated.

The model given in below figure is showing the output. In the first column the name of the authors is used as the target variable while in the second column when the model is trained and tested, it understands the textual data, based on the scores, and classifies correctly. If we look deeply within the output (Barack Obama = Barak Obama, Ellen Degrasse = Barak Obama) it means that tweet was written by Ellen Degrasse using Barak Obama's writing style

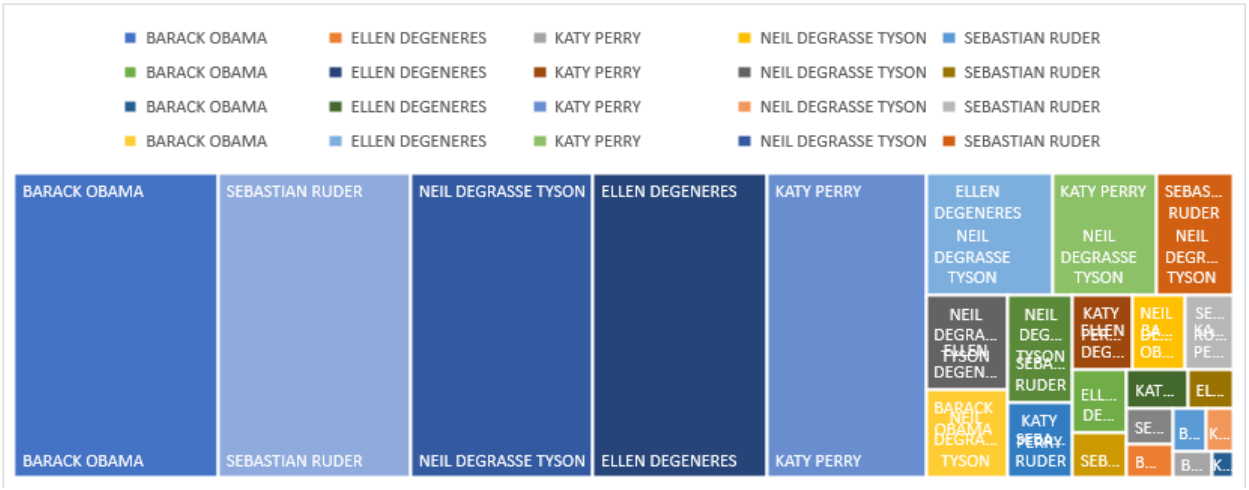
GB Model Output

Classification Table

Data Role=TRAIN Target Variable=author Target Label=author

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
BARACK OBAMA	BARACK OBAMA	88.7399	82.7500	662	16.7130
ELLEN DEGENERES	BARACK OBAMA	2.1448	2.0000	16	0.4039
KATY PERRY	BARACK OBAMA	1.4745	1.4436	11	0.2777
NEIL DEGRASSE TYSON	BARACK OBAMA	5.6300	5.2566	42	1.0603
SEBASTIAN RUDER	BARACK OBAMA	2.0107	1.8750	15	0.3787
BARACK OBAMA	ELLEN DEGENERES	4.9137	4.6250	37	0.9341
ELLEN DEGENERES	ELLEN DEGENERES	75.4316	71.0000	568	14.3398
KATY PERRY	ELLEN DEGENERES	6.2417	6.1680	47	1.1866
NEIL DEGRASSE TYSON	ELLEN DEGENERES	10.8898	10.2628	82	2.0702
SEBASTIAN RUDER	ELLEN DEGENERES	2.5232	2.3750	19	0.4797
BARACK OBAMA	KATY PERRY	1.0050	0.7500	6	0.1515
ELLEN DEGENERES	KATY PERRY	4.1876	3.1250	25	0.6312
KATY PERRY	KATY PERRY	86.2647	67.5853	515	13.0018
NEIL DEGRASSE TYSON	KATY PERRY	2.1776	1.6270	13	0.3282
SEBASTIAN RUDER	KATY PERRY	6.3652	4.7500	38	0.9594
BARACK OBAMA	NEIL DEGRASSE TYSON	7.2165	9.6250	77	1.9440
ELLEN DEGENERES	NEIL DEGRASSE TYSON	15.4639	20.6250	165	4.1656
KATY PERRY	NEIL DEGRASSE TYSON	12.8397	17.9790	137	3.4587
NEIL DEGRASSE TYSON	NEIL DEGRASSE TYSON	55.1078	73.5920	588	14.8447
SEBASTIAN RUDER	NEIL DEGRASSE TYSON	9.3721	12.5000	100	2.5246
BARACK OBAMA	SEBASTIAN RUDER	2.2556	2.2500	18	0.4544
ELLEN DEGENERES	SEBASTIAN RUDER	3.2581	3.2500	26	0.6564
KATY PERRY	SEBASTIAN RUDER	6.5163	6.8241	52	1.3128
NEIL DEGRASSE TYSON	SEBASTIAN RUDER	9.2732	9.2616	74	1.8682
SEBASTIAN RUDER	SEBASTIAN RUDER	78.6967	78.5000	628	15.8546

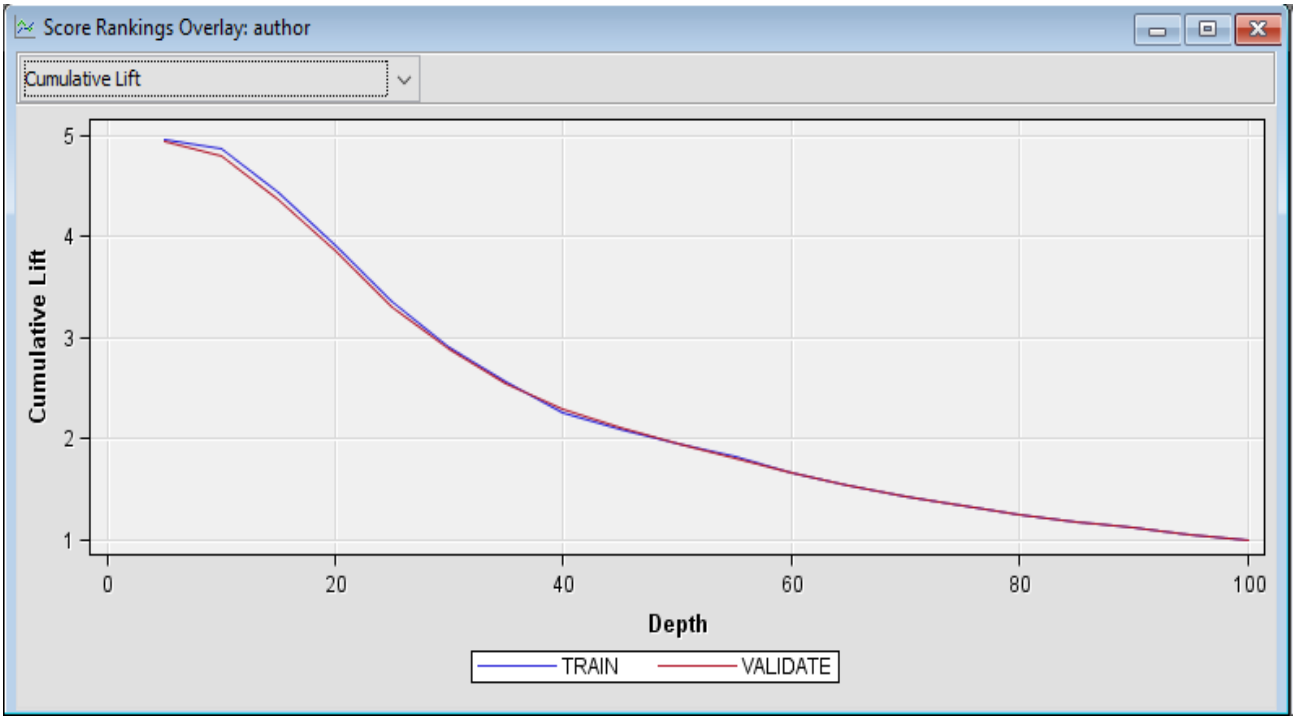
. Heat Graph of Model



D. MBR

MBR stands for memory base reasoning algorithm which is used to categorize the data according to given parameters. This modeling is working based on K-Mean clustering which categorized the variables according to scores. In the given dataset the MBR works by recognizing the pattern of the given topics from the data. After implementing this model the output is given as under.

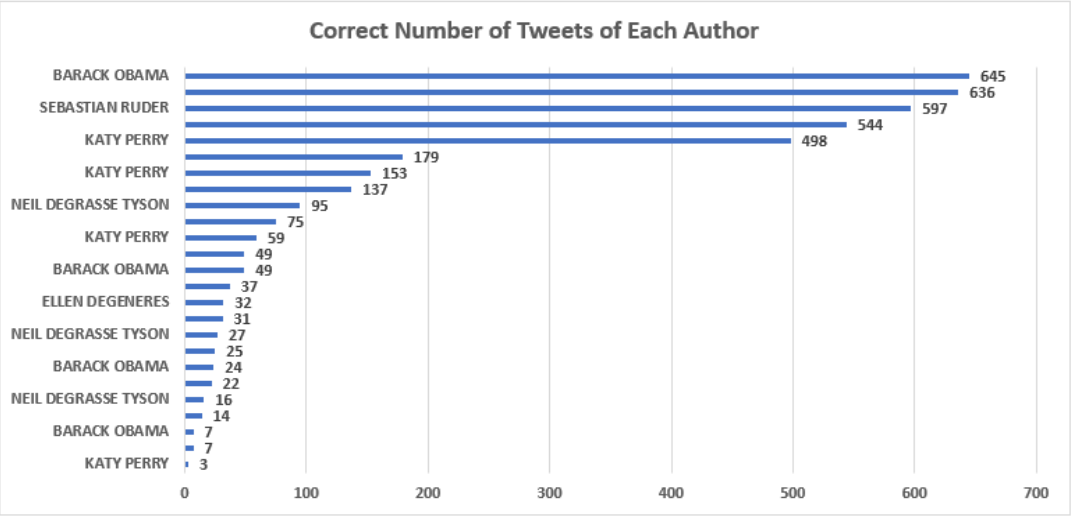
MBR Accuracy Curve



The graph given in figure is showing that the curve of the training model and testing model. The movement of the curve from top to bottom of both training and testing is equal and smooth. There is no gap between them which means the model which is trained is the best fit while validating it. The movement of the curve is showing the accuracy of both models is near 99.6%, which is considered the best fit. Upon further analysis, following

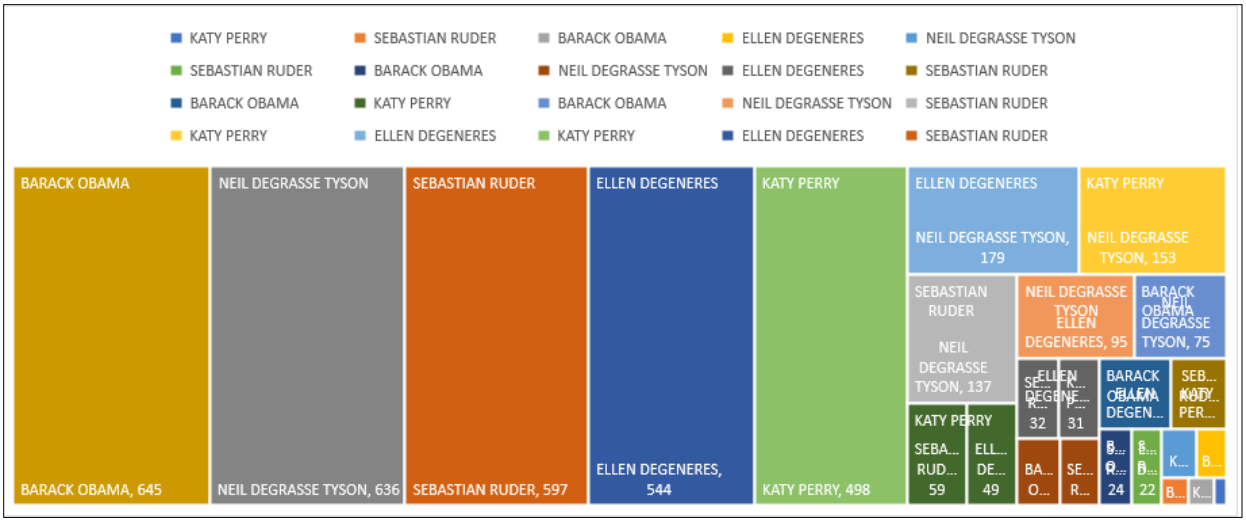
The graphs given below is showing the actual tweets made by the relevant personality, while at the beginning of this research the number of tweets was different made by each personality. This model is showing an exact number of tweets made by each person.

Tweets by Author

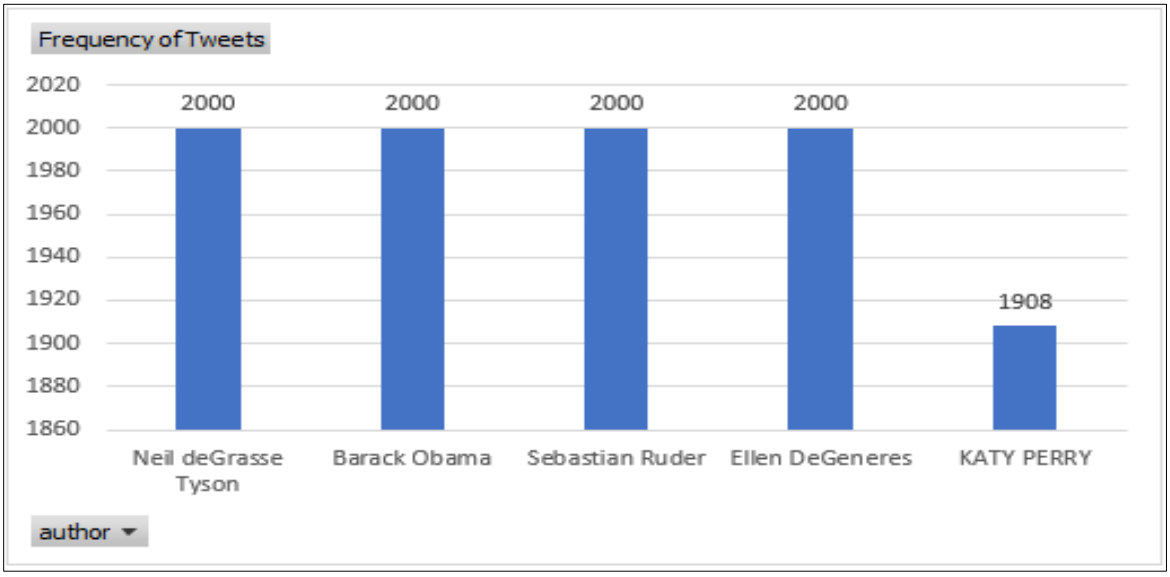


The graph in figure below is showing the actual distribution of tweets made by each person. While by checking very first tweets it is found that total (Neil deGrasse Tyson = 2000, Barack Obama = 2000, Sebastian Ruder = 2000, Ellen DeGeneres = 2000, Katy Perry = 1908). so these tweets were those which were not originally tweeted by them. By analyzing the above figure it is found that there was writing resemblance to tweets. These tweets were not originally belonged to the person by whom it was posted.

Pattern Chart of Tweets



Frequency of Tweets



E. HP Cluster

HP Cluster node is used for segmentation of the data based on the scores. These clusters contain similar types of data that belong to each entity within the dataset. By using disjoint cluster analysis on the basis of Euclidean distances. After performing the cluster the output is examined from the graphical output.

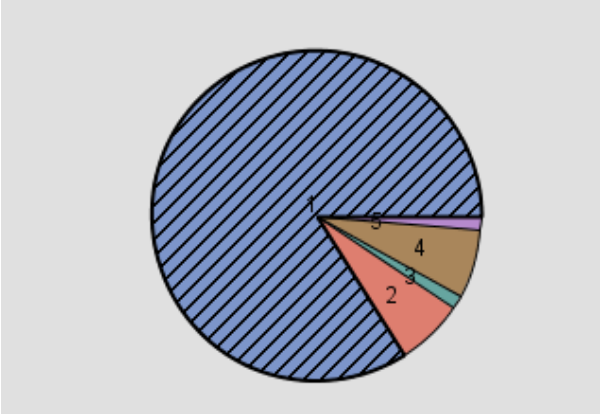
Model Information

Model Information

Parameter	Setting
Maximum Iteration	10
Stop Criterion	Cluster Change
Stop Criterion Value	20
Clusters	5
Seed Initialization	12345
Distance	Euclidean
Number of Cluster Estimation	ABC

figure is showing the setting of the model in which the total number of iteration is 10 which mean that the loop will run 10 times to cluster the data. And a total number of clusters is 5. The following output is generated from the cluster analysis.

Clusters Distribution



The above figure shows that the whole dataset is divided into 5 clusters in which most of the clusters belong to 1 whereas the least number of clusters belong to cluster 3.

Comparative Result of Models

After completing the analysis using different models, in this step, we are going to examine which model is the best fit for this type of analysis and dataset. For this purpose, we have used a comparison model node within SAS Enterprise Miner, which enables us to find the best model based on the accuracy score of all models. The output of the result is given in below figure, five different models are given with their accuracy scores. The first model is MBR with highest score of (MBR Model = 97.09, Gradient Boosting = 97.06, Decision Tree = 83.89, HPBNC = 81.36). So, it is concluded that MBR and Gradient Boosting are fit models for this analysis

Comparative Result of All Models

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Test: Cumulative Percent Response ▼
	MBR	MBR	MBR	author	author	97.09445
Y	Boost	Boost	Gradient Bo...	author	author	97.06376
	Tree	Tree	Decision Tr...	author	author	83.89262
	HPBNC	HPBNC	HP BN Cla...	author	author	81.36293

Conclusions/Future Work

The basic purpose of this research was forensic authorship identification using different machine learning models and techniques. This forensic analysis is done using Tweets data of relevant authors to know whether that tweet was published by its or someone else by using his writing style or method because with the spread of social media, it has become hard to find the original authority of any post without deep analysis. Many people try to use other writing styles, even sometimes using the same name for posting fake news, tweets, and other data. This analysis report helps in the identification of authorship which will reduce cybercrime and one can easily find the originality of the account in future applications.

The models are tested based on the accuracy score as well as the final output generated by the models. After performing different steps, to prepare data, selection of variables, and use accuracy scores, it is concluded that the "Gradient Boost and Memory Boost Regression" Models are the best fit for the forensic authorship identification process through the Stylometry technique. This study has been undertaken by following all research questions, and goals, and meeting all research objectives which are written in the research problem section. This research is also based on the previous knowledge published in different papers and by keeping in view those parameters that have already been published, I tried to make everything new by increasing the accuracy score.

This study can be used as starting point for the forensic analysis of Twitter data for the identification of ownership and Stylometry style. The accuracy of models is considerable but, in the future, might be increased by using other parameters and methods rather than this research. This research will be highly beneficial for the cybercrime unit of any country to reduce the fake news, and posts and trace which news it belongs to.

Bibliography

Afroz, Sadia, Michael Brennan, and Rachel Greenstadt. 2012. “Detecting Hoaxes, Frauds, and Deception in Writing Style Online.” *Proceedings - IEEE Symposium on Security and Privacy*, 461–75. https://doi.org/10.1109/SP.2012.34.

Alonso-Fernandez, Fernando, Nicole Mariah Sharon Belvisi, Kevin Hernandez-Diaz, Naveed Muhammad, and Josef Bigun. 2021. “Writer Identification Using Microblogging Texts for Social Media Forensics.” *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3 (3): 405–26. https://doi.org/10.1109/TBIOM.2021.3078073.

Anwar, Waheed, Imran Sarwar Bajwa, and Shabana Ramzan. 2019. “Design and Implementation of a Machine Learning-Based Authorship Identification Model.” *Scientific Programming* 2019: 12–14. https://doi.org/10.1155/2019/9431073.

Iyer, Rahul Radhakrishnan, and Carolyn Penstein Rose. 2019. “A Machine Learning Framework for Authorship Identification From Texts,” December. https://doi.org/10.48550/arxiv.1912.10204.

June, Conference Paper, David Yarowsky, Shane Bergsma, Matt Post, and David Yarowsky. 2020. “Stylometric Analysis of Scientific Articles Stylometric Analysis of Scientific Articles Department of Computer Science and Human Language Technology Center of Excellence Abstract We Present an Approach to Automatically Re- Cover Hidden Attributes of Scien,” no. June 2012.

Khedkar, Sujata, Shashank Agnihotri, Anshul Agarwal, Mahak Pancholi, and Pooja Hande. 2018. “Author Identification Using Stylometry.” *2018 International Conference on Smart City and Emerging Technology, ICSCET 2018*, November. https://doi.org/10.1109/ICSCET.2018.8537362.

Mohammed, Mohssen, Muhammad Badruddin Khan, and Eihab Bashier Mohammed Bashie. 2016. *Machine Learning: Algorithms and Applications*. *Machine Learning: Algorithms and Applications*. https://doi.org/10.1201/9781315371658.

Parsad, Rajender. 2014. “Sas for Statistical Procedures,” no. August 2010: 273–313.

Pavelec, D., L. S. Oliveira, E. Justino, F. D.Nobre Neto, and L. V. Batista. 2009. “Compression and Stylometry for Author Identification.” *Proceedings of the International Joint Conference on Neural Networks,* no. June: 2445–50. https://doi.org/10.1109/IJCNN.2009.5178675.

Pavelec, Daniel, Edson Justino, and Luiz S. Oliveira. 2007. “Author Identification Using Stylometric Features.” *Inteligencia Artificial* 11 (36): 59–65. https://doi.org/10.4114/ia.v11i36.892.

Pearl, Lisa, and Mark Steyvers. 2012. “Detecting Authorship Deception: A Supervised Machine Learning Approach Using Author Writeprints.” *Literary and Linguistic Computing* 27 (2): 183–96. https://doi.org/10.1093/llc/fqs003.

Plotnikova, Veronika, Marlon Dumas, and Fredrik P. Milani. 2019. “Data Mining Methodologies in the Banking Domain: A Systematic Literature Review.” *Lecture Notes in Business Information Processing* 365: 104–18. https://doi.org/10.1007/978-3-030-31143-8_8.

Ramyaa, Congzhou He, and Khaled Rasheed. 2004. “Using Machine Learning Techniques for Stylometry.” *Proceedings of the International Conference on Artificial Intelligence, IC-AI’04* 2: 897–903.

Wirth, Rüdiger. 2000. “CRISP-DM : Towards a Standard Process Model for Data Mining.” *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining,* no. 24959: 29–39.