Daniel Kiely
Professor Leberknight
Authorship Analysis Graduate Research Project
05/11 /2022

Abstract

The purpose of this paper is to conduct an experiment in which stylometric features can be extracted and further analyzed to accurately predict the author of a document when comparing them to a set of known authors within a test data set. This discourse first provides an introduction to the importance of cybersecurity and cybercrime investigation. The introduction supplies a foundation for Authorship Analysis by describing what it is historically and further comparing the disciple to its modern utility. The research is then followed by a literature review of existing work on Authorship Analysis. The project is further broken down into two sections which were found to be most pivotal for Authorship Analysis: Feature Selection and the Techniques used for Authorship Analysis. It was found that both of these categories lead to varying results depending on the features selected and methods chosen. This project was influenced by the work of R. Ramezani, and was adapted by using differing features. The primary purpose of this project is to use the TF_IDF algorithm to accurately identify an unknown text document when compared to the set of known authors.

Keywords

CyberSecurity, Digital Forensic, Authorship Analysis, Stylometric Features, Feature Selection, Stylometry, TF_IDF, CIA triad, Integrity, NLP, Natural Language Processing

Introduction

As the boom in technology and use of computers is increasing exponentially, the escalation and prevalence of cybercrime are also on the rise [4]. The trend of cybercrime increasing is not stopping anytime soon according to Verizon 2021 Data Breach Investigations Report as 5,258 confirmed data breaches have occurred spanning 16 industries across four world regions. To further solidify the increasing cybercrime trend, in 2020 ransom payment peaked at $233,817 and phishing scams had peaked at 510% [11]. As the field of cyber investigations is evolving there are evidently going to be difficulties that arise while investigations take place. As new technologies come to fruition cybercriminals will seek to exploit these for their personal gain. These difficulties in the cyber realm can be equated similarly to the field of Forensic and the way that traditional scientific inquiry influences the investigation process. There are inherent difficulties that are present in all investigations with the rapid advancements in information communication technologies (ICTs) which makes these difficulties highlighted for investigators, prosecutors, and other responding individuals to cybercrime. This is why it is pivotal to explore all areas of cybercrime investigations and the hurdles that must be overcome in order to make the investigation process much more simplistic in nature. An area of research that has been pivotal throughout history, not only in cybercrime investigations, is authorship analysis.

This area of the field has a clear focus on anonymity that is still present in the modern-day. On October 22, 2021, an article was published that discussed the European Union's (E.U) potential ban on using an anonymous domain. The reason behind this potential ban that E.U is considering is because it is understood premise that an anonymous domain is typically used for conducting illegal activities and cybercrime including authorized distribution of copyright-protected works [8]. This consideration to ban anonymity could potentially prevent cybercrimes if enforced properly. The primary purpose of this potential ban is to make cybercrimes more difficult and more expensive to commit, which in turn would deter the use of

anonymous activity, which could potentially pave the way for more legislation to be drafted and enforced. This is important in the realm of authorship analysis because the widespread number of copyrighted documents would drop on the global scale because it would be much more difficult and expensive, as previously stated, to provide.

The unethical spread of Copyright information and documents is not easily detectable to the average person, as it is difficult to pinpoint what is the exact source of the information. More specifically, it is certainly no shock of the imagination to say that each individual author and person alike have their own unique handwriting and writing style. This concept is evident through each individual's everyday actions. With everything individuals do, ranging from writing research documents in academia to drafting casual emails with friends and family, each individual subconsciously uses their own writing techniques. The study of this is the field of Stylometry and more specifically authorship analysis. To pinpoint exactly what Authorship Analysis is, previous research has defined it as examining the characteristics of a piece of work in order to draw a conclusion based on the author of this body of work. The implications of this area of research are pivotal in investigations. As Cameron S. D. Brown has proposed in his work, for an act to be characterized as a cybercrime an act must be motivated by the *intent* to commit harm against a *person* or *organization* and is facilitated by information and communications technology [5], amongst other criteria. Authorship Analysis is specifically related to the above criteria because, in order to prove intent, it can be helpful to analyze documentation that is considered to be circumstantial evidence, which can include information that is stating the intent of the crime.

Authorship analysis can narrow down the number of suspects that were on the network, or as Chaski states, who was at the keyboard when incriminating documents were produced [6]. It is evident that authorship analysis can lead to breakthroughs and give accurate identification processes during investigations, however some methods of determining the author of text documents are far more accurate than other methods. As Rudman argues, the results of most nontraditional authorship attribution studies are not universally accepted as definitive. One major indication that there are problems in any field is when there is no consensus on results, no consensus as to accepted or correct methodology, and no consensus as to accepted or correct techniques [3]. It will prove evident that there are countless methods, results, and techniques used to attempt to identify the author of a text. To list only but a few examples of approaches that have been used in authorship analysis are biometric, qualitative, and stylometric approaches that are common in the field of authorship analysis. The degree to which each method is used is dependent upon the research topic of the author(s).

With the dominance of cybercrime in today's world, it is of the utmost importance to find ways to stop and prosecute those individuals who commit these acts. It is particularly difficult to pinpoint exactly who the offender is based on modern techniques used to avoid this from occurring. Cybercriminals will routinely forge the sender's address which is routed through an anonymous server and use multiple usernames in hopes to distribute online messages via different anonymous channels [7].

Research Problem

The purpose of this body of work is to implement the TF_IDF algorithm which

will analyze a set of text documents, referred to as our 'train' data set, in order to identify an unknown author of an introduced document. With regards to the overarching scope of this research, in terms of the Confidentiality, Integrity, and Availability (CIA) triad within the Cyber Security discipline, this research is primarily focused on the integrity aspect of data transmission security. The CIA triad is a fundamental principle that forms the basis for CyberSecurity practices and when all of the segments have been met, the system is considered secure. Ensuring the integrity of data is primarily focused on ensuring the transmission of data was not tampered with. The integrity of data must remain as was originally created by its true curator and if it is not, the data was tampered with and cannot be considered reliable. This is seen very evidently with Military Decision making processes with memos being sent about crucial, time sensitive data. If this data has been tampered with, it cannot be trusted and could be catastrophic. Within an academic setting, integrity of information could be the difference between a student graduating or not. This is seen countless times with students plagiarizing others' work and using it as their own. This approach allows us to propose a resolution to this. The train data set includes letters, chapters of novels, articles, and various other forms of text documentation making it cross-compatible with nearly any discipline involving text documentation. It will prove clear that this model can be utilized in the numerous applications that were previously discussed. This work has been inspired by the ongoing issue of authorship identification analysis in the realm of Cyber Crime identification.

Literature Review

The primary focus of this body of research is to analyze methods of authorship identification utilizing text documents and further analyze the results while looking through a variety of applications. The authorship analysis problem has been broken down into three main components which are used to study the area: Authorship Identification (otherwise known as Authorship Attribution), Authorship Characterization, and Similarity Detection [7]. The three categories act as general steps/guidelines which are used to identify an author of any text and not simply in the cyber realm. Authorship analysis has been researched historically and can be traced back as far as the works of William Shakespeare and the Federalist Papers [various sources]. However, the more recent work has been focused predominantly with respect to two key aspects of this broad field: feature selection and similarly the over-arching techniques that have been used to analyze the authorship of a piece of work [1]. When discussing these two main features of authorship analysis, Feature Selection and Techniques used for Authorship Analysis, previous research conducted by Holmes and Forsyth [2] has shown that word based methods of Feature selection for an author analysis is extremely extensive and dependent on new research, which makes it extremely difficult to apply it to an extended application. However, in order to focus the research in a way that would be more reliable than using style markers [1], Joseph Rudman [3] had proposed the idea of style markers and, more specifically, the concept of a finite amount of style markers should be focused on to analyze and 'solve' attribution problems. However, when discussing this concept of a finite number of style markers to analyze, 1000 as [1] claims, there can be an endless amount of studies done and each can claim that the chosen combination of style markers is the most effective combination of markers chosen, which is not always accurate to

reality. Another area of focus within the realm of authorship analysis investigation is studying the techniques that are used for the analysis of text documents. It has been understood that the majority of older studies in this area are focused primarily on statistical tests of word usage for a given author. This has since been proven unreliable and inaccurate in terms of identifying authors to their text with high levels of accuracy across multiple texts of the same author [1]. Although Rudman does not conduct an experiment based upon authorship analysis, his work provides some useful examples of the issues with feature selection and methods used for authorship analysis.

A. Feature Selection

Feature selection is the process of focusing on stylistic patterns in a chosen author's writing [9]. The goal of feature selection is to reduce the number of unnecessary features in classification rules [15]. Choosing features to focus on is entirely dependent upon the researcher's preference. The features that are chosen can depend upon language, sex, time period, and countless other factors. Pavelec, Justino, and Oliveira focused their research on a feature set predominantly based upon conjunctions of the Portuguese language [12]. They used Linguistic Features of the Portuguese language to determine the feature set and applied the Support Vector Machine method in order to determine the author of their text data sets. It has been seen that these features can be found in various forms according to Houvardas and Stamatatos, as they claim that these stylistic features can take the form of lexical, syntactical, and structural features [9]. Using the structural features as a starting point for feature selection, as this feature is seemingly self-explanatory in its area of focus, Chaski further explains the structural features that were used throughout her research. These structural features include punctuation, sentential complexity, grammatical and spelling errors. However, there were shortcomings within this research in that Chaski did not focus on a gender-neutral pool of authors as she only chose female authors [10].

A more modern approach to feature selection is outlined throughout the work of [17]. These authors had an interesting approach to classifying text documents based on whether or not these documents are considered hate speech or not hate speech. These features were based upon function word usage and emotion-based features to determine which category the text documents fall into, hate speech or not hate speech. The authors used social media data sets for their analysis. More importantly, the researchers had to narrow their features down. They did this by understanding the principles of *n*-grams, which provide good results for their specific model of detecting abusive language. However, the authors were unaware of the impact of the subsets of the chosen features which lead them to use broader POS n-gram features. These features include Part of Speech (POS), Stylometric, and Emotion-based features, with the authors putting more emphasis on the Emotion-based features and function word usage [17]. The remaining portion of the work analyzes the results of their statistical analysis. The results of this study added much useful information to the state-of-the-art hate speech detection methods that have already been in place by various projects. The addition of stylometric features coupled with the emotion-based features added a comprehensive and reliable result when analyzing documents for hate speech.

Another approach that has been widely studied in the literature of the past has been the *Rough-Sets* approach to feature selection as discussed by Zhang and Yao

[15]. Rough-Sets, simply put, is a branch of sets-theory. This technique was adapted to the formation of PASH, Parameterized Average Support Heuristic. The benefits to using this method of Feature Selection, according to [15] include consideration of all average support rules for every decision class, as well as consideration of predictive instances that have been excluded by the previous research. The reason this PASH technique is useful to look into is because of the concept that it takes predictive instances into consideration. This allows for a much more narrow feature set to be determined by having a lower approximation. Although this method is different from other feature selection forms, it is not without flaws. As with most concepts in statistics, it is impossible to have 100% accuracy in these atmospheres, meaning that there would need to be countless years of testing to prove this 100% accurate.

B. Techniques for Authorship Analysis

Many techniques have been developed and employed to analyze who the author of a text is most likely to be. Techniques such as statistical analysis utilizing the widely known Chi-Square test have been used by numerous authors including Chaski [10]. The Chi-Square test allows researchers and the audience to quantify linguistic features, in terms of authorship analysis, and transform them into something statistically tangible. However, this is not the only method that can be used to determine authorship analysis, although statistics does play a pivotal role in nearly all techniques. An area that is also heavily researched and utilized in the area of authorship identification is machine learning.

Pavelec, Justino, and Oliveira focused their research by using stylistic features but the method in which the authors chose to analyze these features was using Support Vector Machines (SVMs), as mentioned in the previous section on feature selection [12]. SVM's were chosen because they are designed to deal with two classes coupled with the idea that they can deal well with outliers in a set of data. More simply put, SVMs take a set of data input and predict which of the possible classes the input more narrowly falls into. The shortcomings of using SVMs are noted as being they do not deal with Max, Min, Average, and Median of the statistical data produced, which would be useful in authorship analysis considering all data must be converted into a statistical approach that allows the audience to quantify the data and understand it on a more simplistic level. It is clear that this method has its shortcomings but it is also abundantly clear that this is not the only method used for the Techniques used in Authorship Analysis.

Another more simplistic method is the method chosen by the author R. Ramezani, which is by creating an *n*-gram model of an author's writing using techniques from statistical language modeling [16]. In more common terms these authors chose to use a language modeling approach for capturing regularities within the natural language which can be further used to make predictions about authorship analysis. Within this work, the author describes the importance of *n*-Gram Language modeling as being widely used in various applications such as spelling correction, character recognition, machine translation, along with numerous other applications. However, he further discusses the formulas used to predict, with high probability, the natural word sequence (word sequences that actually occur within a text document). The formulas/models chosen for this body of research were the empirical perplexity and entropy scores. The authors used the Bayesian decision theory to perform their experiment while analyzing

the text/characters between authors of English, Greek, and Chinese nationalities. It is important to note that throughout this body of research, the authors are assuming each piece of text is written by a single author as opposed to multiple individual authors. [16] compared their results with previous research conducted upon the same data set and the method seen within [16] was more simplistic, but showed improved results within the same Greek data sample. However, one conclusion drawn from this method is that this language modeling approach is more effective at capturing author-specific idiosyncrasies in a homogeneous collection of text, meaning that the text is of the same style (academic research for the Greek Data set described above).

Within the English and Chinese data sets, the authors focused on the characters in the language, which allowed them to have 98% accuracy using the 6-gram model using the absolute soothing method for the English data set. The Chinese dataset analysis was successfully conducted using the 3-Gram model using Witten-Bell smoothing and was able to attain 94% accuracy in correctly identifying the author of the text. The drawback to the English method was that [16] used a data set and specifically chose the most well-known/prolific authors. Meaning that each of these authors has an extremely distinct writing style. It would be beneficial for the authors to choose a more diverse, and challenging data set for both the English and Chinese language data sets due to the results that were obtained from the authors and text documents chosen. This would give the author's method more credibility and limit the number of potential biases within the discussed method.

However, when discussing this notion of a finite number of style markers, 1000 [1], there can be thousands of studies done and claim that the chosen combination of style markers is the most effective combination of markers chosen. Another area of focus in the realm of authorship analysis investigation is studying the techniques used for the analysis of text documents. It has been proven that the majority of older studies in this area are focused primarily on statistical tests of word usage for a given author. This has since been proven unreliable and inaccurate in terms of identifying authors to their text with high levels of accuracy across multiple texts of the same author [1].

As was seen throughout the research in the field of authorship analysis, each body of work has its shortcomings. Many of these shortcomings can be attributed to the scope, or lack thereof, within the model that was implemented throughout each experiment. Many of the models were not language independent and would work more effectively while analyzing the text/document which is written in a specific language. The focus of this body of research is to adopt a language-independent model and accurately predict the author of a text document. The results will then be further analyzed against the original experiment. This original experiment in question is the work of Fuchen Peng, Dale Schuurmans, Vlado Keselj, and Shoajun Wang [17]. However, one of the areas of critique that was evident throughout this work was the length of the documents being analyzed. The research briefly mentions the length of the document as a stylometric feature, however does not go into much more detail or place importance on the length of the document. The basis of this notion is the idea that the model presented throughout [17] may react differently and yield results that differ based upon the length of the document. For instance, if the documents analyzed through [17] were each over 500 characters long, this body of work would seek to investigate documents, such as tweets, with a much

smaller character count and see if this model holds the same results. The focus of the duration of the research will primarily seek to identify the author of small sample text documents such as tweets, text messages, or emails while using the same model to further test the methods used throughout [17].

In order to recreate this experiment all of the components should be further investigated. The data sets used were found online via public information such as Reuter C50 Corpus as well as fifty short stories in Persian. To analyze these text documents the author then used multiple algorithms including SVM (Support Vector Machine), NN (Neural Network), RF (Random Forest), KNN (K-Nearest Neighbors), NB (Naive Bayes), Cos (Cosine Similarity Measure), and Hel (Hellinger Distance Measure). After utilizing the algorithms listed, the authors then narrowed their feature selection to a total of 13 features listed, however, some of the 13 have different cases and aspects which leads to 30 stylometric features chosen and evaluated separately. They then used N-grams 1,2, and 3 to further break down the results of their experiment for the English language. The selected methodology for the Persian data set was as follows. The authors used what is known as Term Frequency Inverse Document Frequency Scheme (TF_IDF). This statistical method highlights the importance of a term that is used to calculate the similarity between an anonymous document and known documents. More simply put, this TF_IDF method is used to determine the importance of a specific term within a document with respect to a corpus that is being studied. With this TF_IDF method, there is no NLP tool needed which makes it language-independent, hence why it was used to study the Persian Data set.

The shortcomings of this research were seen in the vast amount of features used that all did not prove to be perfect in identifying an author. Numerous features simply had average or below average results when compared to other features selected. Had Ramezani simply focused his research on finding the most productive features for authorship analysis, it would have made the research much stronger and could have yielded much more research to follow from it. However, simply because he researched thirty stylometric features, it did not leave much room for further research in that aspect.

Prototype

When constructing and implementing this model, multiple functions were used to successfully carry out the operation. What makes the model being constructed throughout this experiment different from that of the work of R. Ramezani is that only the English data set was in focus as well as the TF_IDF algorithm and not multiple algorithms and a Persian dataset. The reason that the TF_IDF algorithm was used is because of the lack of natural language processing (NLP) tools, as previously mentioned. This makes the algorithm and model seemingly language independent. TF_IDF which stands for Term Frequency–Inverse Document Frequency has been frequently used for information retrieval for organizations and researchers alike. It has been a technique used to quantify words that are found within a given set of documents. The TF_IDF algorithm has been repeatedly used to weigh and assign importance to keywords that can aid in the mining of a document and can further be used to more accurately predict the author of a given anonymous text document. The principle behind this algorithm is rooted in the number of times a phrase occurs in a given document. TF_IDF looks at the frequency of the term and compares it to a corpus, which is the

collective set of all texts being analyzed. This algorithm works by correlatively increasing the number of times that a given word or set of words are seen within a document. At the same time this frequency is then offset by the total number of documents that contain the given word/ set of words. In more common language; words and phrases that are common in all documents such as “the”, “these”, “what”, all will provide low ranking because they do not have much impact on the paper's substance despite the fact that they appear consistently throughout the document(s). This also works in an inverse fashion as well, meaning that the lower the TF_IDF score, the more common the term is in that given text document. Once this score was computed, we were able to construct a web application that is constructed on a local machine hosted at Montclair State University. This web applications interface can be seen in Figure 1 listed below. The Drop-Down-List, DDL, allows users to navigate to a specific known user they would like to compare.

In order to get the implementation/back end program to display in a web browser, Djangoproject was used. *Djangoproject* (*Django* for short) is a web framework that allows the development and integration of *Python* into web designs. The primary rationale behind using Django was based on the principle that the programming language *Python* is not fully supported by web browsers and does not integrate properly if used on its own when programming a web application. Simplified, *Django* is a software that allows *Python* programs to be run for web applications development. The r

A. Dataset

The dataset used for this experiment was a set of American authors from Reuter C50 Corpus. This corpus contains American authors, some more well known than others, and have varying lengths for each text document provided. Each text was written in the American English language and is not topic/genre specific meaning that each text document being analyzed could be covering a wide array of topics depending upon the author. Within the data set the authors were divided into train data and test data. The training data allowed the program to learn and recognize patterns within the data. The test data set is then introduced and further used to test the program's accuracy. Both the training data set and the test data set were functions of the program implemented. The other function used to execute this experiment was the use of multinomial classifiers.

B. Multinomial Classifier

This model used multinomial classifiers to test if term length would alter the accuracy of the model. The term multinomial in this instance is used to describe the amount of terms used to analyze in a phrase. For example, a classifier can be *bi-gram*, two word phrases, and range to any number of words used to analyze called *n-gram*. The use of n-grams and n-gram probability is seen frequently in auto-completion of sentences that is consistently implemented in Gmail, automated spell-checking, and can even be attributed to grammar checking on specific platforms. In order to get the implementation/back end program to display in a web browser, Djangoproject was used. *Djangoproject* (*Django* for short) is a web framework that allows the development and integration of *Python* into web designs. The primary rationale behind using Django was based on the principle that the programming language *Python* is not fully supported by web browsers and does not integrate properly if used on its own when programming a web application. Simplified,

Django is a software that allows *Python* programs to be run for web applications development.

Results

The results of this experiment were underwhelming in the scope of attempting to replicate previous work. We were not able to get the exact precision rate as the authors mentioned above when looking at the exact features, however we were able to get close to the same precision. We were able to produce a precision of 84%, a recall of 80.4%, and an F value of 82.3%. Recall answers the question: what proportion of actual positives was identified correctly? Comparatively, Precision answers the question of: what proportion of positive identifications was actually correct? When attempting to finalize the overall positive percentage rate of this experiment, we were able look at our F1 value. This value is the mean of the overall performance for the model based on precision and recall. Image 2 shows a confusion matrix that displays our results from the experiment conducted.

These results could be further expanded by looking at a new test variable, such as different features extracted from the text documents. Some potential new features to be extracted would be sentence length, verb usage, verb tense, etc. This would further the findings of our model and allow the research to be broadened and more accurate.

Conclusion

The purpose of this research was to construct a web-based software in which stylometric features can be extracted and further analyzed to accurately predict the author of an unknown document when comparing them to a set of known authors within a test data set. We were able to construct this model by utilizing Django and Python to display our results and create an interface. It was seen that n-grams were the most useful feature to extract when attempting to predict the author as it yielded the best results. It has proven evident that authorship analysis has multiple applications including academic, military decision making, and even automatic sentence prediction. The widespread use of authorship analysis tools such as the one constructed throughout this discourse could be a useful tool for all aspects of modern life if applied in the proper context. Further research can be conducted by exploring other features such as length of text, punctuation usage, sentence structure, etc. This will improve the constructed model by narrowing down the scope of features that are useful or broadening them. However, other applications for authorship analysis and other NLP tools.

Appendix

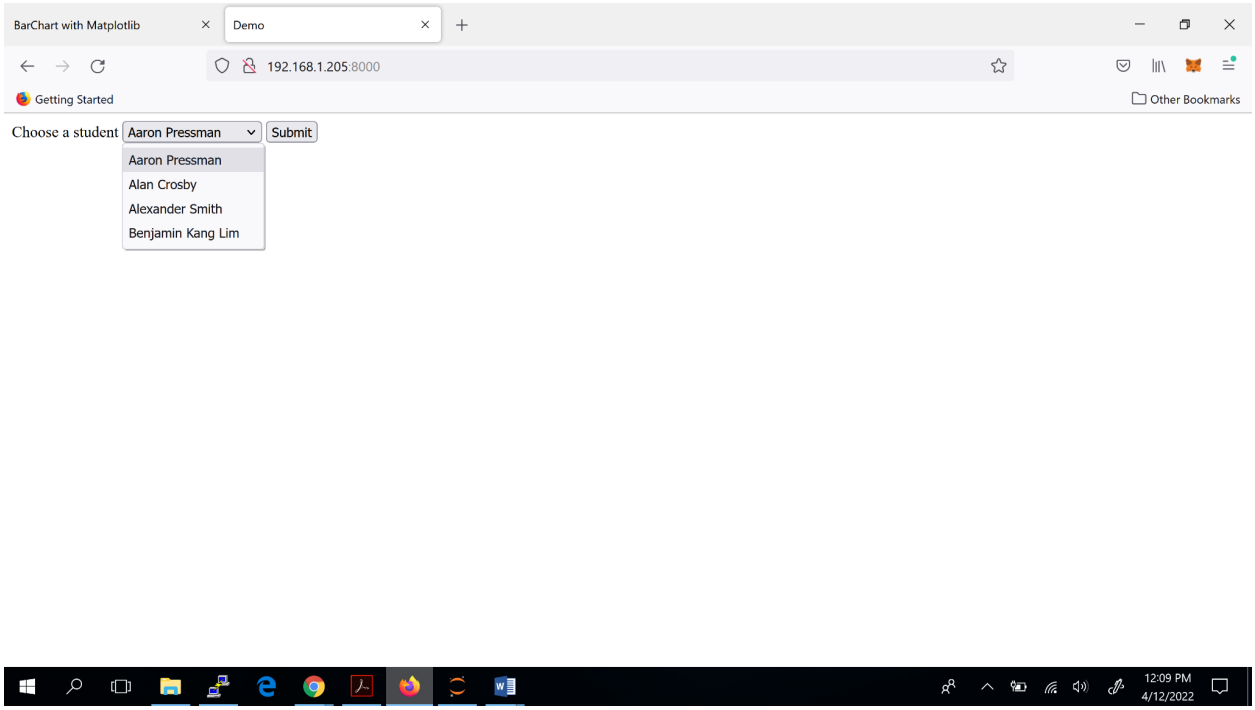

Figure 1. Prototype Web-Application interface

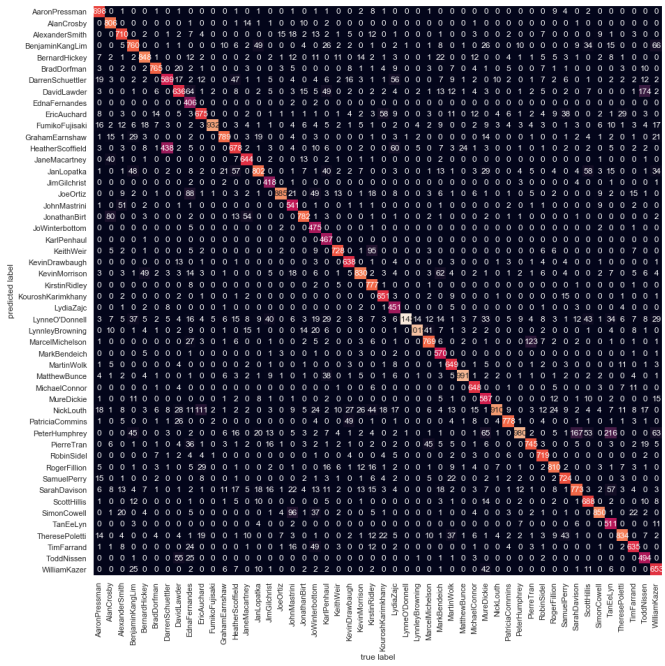

Figure 2. Confusion Matrix from experiment

References

1. Zeheng, R., Quin,Y., Huang, Z., & Chen, H. (2003, June). Authorship analysis in cybercrime investigation. *In Proceedings of the 1st NFS/NIJ Conference on Intelligence and security informatics* (pp. 59-73).
2. Holmes, D., I., & Forsyth R., S., (1995) The federalist Revisited: New Directions in Authorship Attribution
3. J Rudman, The State of Authorship Attribution Studies: Some Problems and Solution. Computers and the Humanities, 31, 351-365. (1998). http://staff.um.edu.mt/albert.gatt/teaching/dl/rudman97_status-of-authorship-attribution.pdf
4. Broadhurst, R. Grabosky, P., Alazab, M., & Chon, S. (2014). Organizations and CyberCrime: An Analysis of the Nature of Croups engaged in Cyber Crime. *International Journal of Cyber Criminology*, *8*(1), 1-20. Retrieved on 22nd February 2015 from http://www.cybercrimejournal.com/broadhurstetalijcc2014vol8issue1.pdf
5. Brown, C.S. (2015). Investigating and Prosecuting cybercrime: Forensic Dependencies and barriers to justice. International Journal of Cyber Criminology, 9(1), 55.
6. Chaski, C. E. (2005). Who's At The Keyboard? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence*, *4,* 1-13. Retrieved on 27th October 2014 from https://www.utica.edu/academic/institutes/ecii/publications/articles/B49F9C4A-0362-765C-6A235CB8ABDFACFF.pdf
7. https://cryptome.org/2014/02/author-cyber-forensics.pdf
8. https://portswigger.net/daily-swig/eu-ban-on-anonymous-domain-registration-welcomed-by-threat-intel-firm
9. https://link.springer.com/chapter/10.1007/11861461_10
10. https://journals.equinoxpub.com/IJSLL/article/view/1690/1151
11. https://www.csoonline.com/article/3634869/top-cybersecurity-statistics-trends-and-facts.html
12. D. Pavelec, E. Justino, L.s. Oliveira (2007) Author Identification using Stylometric Features. *Inteligencia Artificial Vol. 11, No. 36, 2007.*
13. *http://nyc.lti.cs.cmu.edu/yiming/Publications/yang-icml97.pdf -- Comparative study on feature selection in Text categorization*

14. *https://www.researchgate.net/publication/336375764_Authorship_Attribution_using_Rough_Sets_based_Feature_Selection_Techniques -- Authorship Attribution using rough sets based Feature Selection Techniques*
15. *http://www2.cs.uregina.ca/~jtyao/Papers/306_Nafips04.pdf -- Rough Sets based approach to feature selection*
16. Ramezani, R. (2021). *A language-independent authorship attribution approach for author identification of text documents. Expert Systems with Applications, 180, 115139.*
17. Markov, I., Ljubešić, N., Fišer, D., & Daelemans, W. (2021, April). *Exploring Stylometric and Emotion-Based Features for Multilingual Cross-Domain Hate Speech Detection. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 149-159).