The North African Journal of Scientific Publishing (NAJSP)
مجلة شمال إفريقيا للنشر العلمي (NAJSP)
EISSN: 2959-4820
Volume 1, Issue 4, October-December 2023, Page No: 100-108
Website: https://najsp.com/index.php/home/index
SJIFactor 2023: 3.733 0.63 :2023 (AIF) معامل التأثير العربي ISI 2023: 0.383

Analyzing Dataset (Email) by Using Classification Approach to Authorship Identification

Khaled Ahmed Adbeib *
Department of Computer Science, Faculty of Education, Bani Waleed University, Libya

*Corresponding author: khaled.adbeib@gmail.com

Received: September 30, 2023	Accepted: November 23, 2023	Published: November 30, 2023

Abstract:
With the widespread adoption of internet technologies and applications, the misuse of online emails for illicit purposes has become a significant concern. Authorship identification, a crucial text analysis task, involves determining the likely author of a given document. In the context of emails, this methodology proves valuable in attributing a particular email to its originator based on factors such as writing style, word choice, and other linguistic features. Classification analysis emerges as a prevalent approach for authorship identification, employing machine learning models trained on a dataset of known authors to predict the authorship of unknown texts.

The anonymous nature of online emails presents challenges in tracing identities, escalating the gravity of the issue. The internet has unfortunately become a playground for cybercriminals engaging in activities ranging from simple spamming to sophisticated phishing attacks. Authorship analysis stands out as a pivotal measure to counter such illicit cyber activities. This study delves into authorship identification, focusing on a dataset of emails to ascertain whether an anonymous email is created by a suspect [1].

The primary objective of this project is to discern the authorship of anonymous emails by leveraging stylometric features. These features encompass vocabulary richness, sentence length, and writing style. Through an examination of a dataset comprising known emails, the study aims to distinguish and confirm the authorship of anonymous emails. Authorship analysis has demonstrated effectiveness not only in countering illegal cyber activities but also in revealing the true authorship of anonymous emails. This research contributes to the ongoing efforts to bolster cybersecurity measures and address the challenges posed by the misuse of online communication

Keywords: RapidMiner, K-Nearest Neighbor (K-NN), Authorship Analysis, Stylometric Features, Classification Analysis, Authorship Categorization.

Cite this article as: K. A. Adbeib, “Authorship Identification technique used to determine the structure and categories within a given dataset (E-mails) by using Classification Analysis,” *The North African Journal of Scientific Publishing (NAJSP),* vol. 1, no. 4, pp. 100–108, October-December 2023

تحليل مجموعة البيانات (البريد الإلكتروني) باستخدام نهج التصنيف لتحديد الكتّاب

خالد احمد الدبيب *
قسم علوم الحاسوب، كلية التربية، جامعة بني وليد، ليبيا

الملخص

مع انتشار تقنيات الإنترنت والتطبيقات، ظهور سوء استخدام البريد الإلكتروني عبر الإنترنت لأغراض غير قانونية أصبح مصدر قلق كبير. تحديد الكاتب هو مهمة تحليل نصي حاسمة تنطوي على تحديد الكاتب المحتمل لوثيقة معينة. في سياق الرسائل الإلكترونية، يمكن استخدام هذه الطريقة لتحديد الشخص الذي يقف وراء رسالة بريد إلكتروني معينة استنادًا إلى أسلوب الكتابة، واختيار الكلمات، وميزات اللغة الأخرى. يعتبر تحليل التصنيف نهجًا شائعًا لتحديد الكتاب، حيث يتم تدريب نماذج التعلم الآلي على مجموعة بيانات تحتوي على أسماء كتّاب معروفين لتوقع الكتابة لنصوص غير معروفة.

طابع الهوية المجهولة للبريد الإلكتروني عبر الإنترنت يشكل تحديات في تتبع الهويات، مما يجعلها مشكلة رئيسية. الإنترنت يشكل ملعبًا للمجرمين الإلكترونيين الذين يشاركون في أنشطة تتراوح بين الرسائل المزعجة البسيطة إلى هجمات الاحتيال المعقدة. يبرز تحليل الكتابة كإجراء رئيسي لمواجهة مثل هذه الأنشطة الإلكترونية غير القانونية. يستكشف هذا البحث تحديد هوية الكتاب، مركزًا على مجموعة بيانات من رسائل البريد الإلكتروني للتحقق مما إذا كان البريد الإلكتروني المجهول قد تم إنشاؤه بواسطة مشتبه به.

الهدف الرئيسي من هذا المشروع هو تحديد الكتابة للرسائل الإلكترونية المجهولة باستخدام ميزات الأسلوب. تتضمن هذه الميزات غنى المفردات، وطول الجمل، وأسلوب الكتابة. من خلال فحص مجموعة بيانات تضم رسائل بريد إلكتروني معروفة، يهدف البحث إلى التمييز والتحقق من الكتابة للرسائل الإلكترونية المجهولة. أثبت تحليل الكتابة فعاليته ليس فقط في مواجهة الأنشطة الإلكترونية غير القانونية ولكن أيضًا في تحديد الكاتب الحقيقي للبريد الإلكتروني المجهول. يساهم هذا البحث في الجهود المستمرة لتعزيز التدابير الامنية عبر الإنترنت ومواجهة التحديات الناجمة عن سوء استخدام الاتصالات عبر الإنترنت.

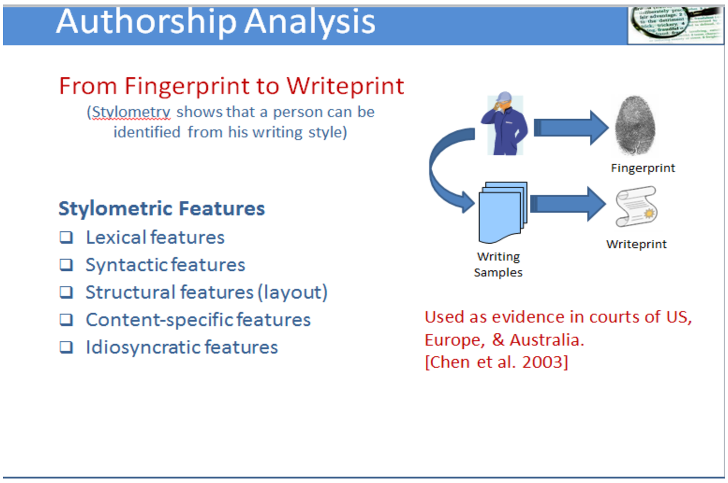
الكلمات المفتاحية: RapidMiner، K-Nearest Neighbor (K-NN)، تحليل الكتابة، ميزات الأسلوب، تحليل التصنيف، تصنيف الكتابة.

1. Introduction

Electronic communication has become one of the most prominent communications in the world and the anonymous nature of online emails is one of these communications, unfortunately, emails may contain viruses, malicious programs, or spyware. Therefore, this issue makes the experts try to find the best solution to prevent or detect the suspect authors. [5].

People have different ways of writing; each one has his/her own writing style. In recent years, the linguistic and stylistic investigation into authorship identification has a long history of determining the authorship of specific work [4]. The authorship analysis of computer-mediated communication (CMC) has attracted much attention from other disciplines due to its application in the prosecution of criminals in the court of law [1].

Briefly, what composes authorship analysis are “authorship attribution, characterization/profiling and verification or similarity detection” [1]. As mentioned, what this project concentrates on is the problem of authorship verification in correctly detecting the true author of a suspect email, an exemplary case would be the detection and investigation of a true perpetrator of illegitimate cyber activities, particularly, the creation of textual spam or phishing e-mails [1]. To do so, a precise analysis supported by plausible evidence is necessary to prove whether a suspect is the true author or not by using stylometric features which will lead us to the true suspect. Authorship identification refers to the relationship between the anonymous email and the true author of that anonymous email by using some features as evidence of the writing style of the suspect [2]. furthermore, by taking advantage of the fact that most emails, unlike books and traditional methods of publication, are short/medium-length, contain more grammatical and structural errors, and include non-documental factors, such as timestamp, subject, and file attachment [1]. In addition, it might include the signature of the email, so by this fact, we can easily detect the authorship of an anonymous email from his/her writing style or fingerprint. The goal of this project is to identify the authorship of a specific email in a large collection of emails based on stylometric features. Figure 1 shows that the person can be identified by his/her writing style based on the stylometric features.



Authorship Analysis

From Fingerprint to Writeprint

(Stylometry shows that a person can be identified from his writing style)

Stylometric Features

- ❑ Lexical features
- ❑ Syntactic features
- ❑ Structural features (layout)
- ❑ Content-specific features
- ❑ Idiosyncratic features

Used as evidence in courts of US, Europe, & Australia.
[Chen et al. 2003]

Figure 1 Authorship analysis [6].

Initially, the creation of the dataset will employ the vp.net language, adhering to the previously defined features. This process aims to extract information about the author of the email. Subsequently, the classification function will be implemented through Rapid Miner to identify the suspect email within the dataset. Graphs will be incorporated to enhance clarity and facilitate understanding of the explanation. Another facet addressed in this project involves the utilization of stylometric features to demonstrate their predictive capability in identifying the suspect email.

2. Authorship Categorization

Although e-mail authorship analysis has produced only a handful of studies, there are interesting studies, such as that of de Vel Anderson et al, and Thomson et al. De Vel uses a simple metric and subset of structural and stylometric features to investigate a set of authors while disregarding personal profiles, such as the age and gender, of the authors. His usage of stylometric features is considered rather narrow and limited. In other studies, such as that of Anderson et al, there has been a more comprehensive use of stylometric features in determining several parameters that range from text size to the number of documents per author that may affect the author categorization performance of e-mails and text documents. Using these features has yielded a satisfactory result of categorization for different paragraph and text chunk sizes. Nonetheless, it is believed that what has caused such results is the built-in bias of N-graphs, which, using N =2, represent adjacent sequences of characters and punctuations, toward content and not just style. The study of Anderson et al renders two important observations: that for text chunks that exceed 100 words, text chunk size does not have a considerable effect on the categorization performance; and that there is no need to have many documents since enough documents for a proper categorization can be as few as 20 [3]. What these observations imply for the author's categorization of e-mail is that small text size and small number of e-mails do not necessarily prevent a categorization from extracting a good result. Furthermore, to compensate for the fact that the study that has been mentioned so far does not consider, Anderson et al provide a study on multi-topic categorizations. In detail, their results dealing with a set of newsgroup e-mails with different topic categories indicate that while categorizations for some authors are successful, others are not. Similarly, studies from Thomson et al manifest that determining and predicting the gender of e-mail authors are possible through systemic examination of styles and vocabularies that are gender-specific or gender-preferential. For example, female characteristics include more frequent usage of adjectives or adverbs and more references to emotions than male counterparts that, on the other hand, may include jargon, insults, slang, and other vocabularies that are male-inherent. The study of Thomson et al suggests that extracting such hypothesized features, and performing discriminant analysis may determine the gender of e-mail authors correctly [3].

3. Data collection

Gather a dataset of emails with labelled authors. Each email should be labelled with the corresponding author's identity.

Ensure that the dataset is diverse, and representative of the author's needs to identify.
As previously mentioned, the primary objective is to correlate unknown emails with the correct suspected author. Therefore, we must gather several emails from each suspect, saving them into a text file. Additionally, it is essential to adhere to a specific format, as illustrated in the following example.
From: Abdullah AL-Barakati <mr.albarakati@gmail.com>
Subject: The presentation time and meeting time
Date: 29 March, 2012 6:11:24 PM EDT
To: Abdullah AL-Barakati
Cc:
Bcc:

Hi guys,

We have meeting today.

Thank you
Abdullah Albarakati

$signiture:
Abdullah AlBarakati
Halifax, NS, Canada
Cell phone :001-(902) 999-4430

From: Abdullah AL-Barakati <mr.albarakati@gmail.com>
Subject: Project News
Date: 28 March, 2012 2:52:14 AM EDT
To: Abdullah AL-Barakati

Hi guys,
The meeting is canceled.

$signiture:
Abdullah AlBarakati
Halifax, NS, Canada
Cell phone :001-(902) 999-4430

Therefore, it is necessary to prepend the "$" symbol before the signature section " $signature:" to ensure it is correctly identified as a signature by our software (refer to Figure 1).

4. Creating Dataset

Creating the dataset is based on the stylometric features (see 1.1.1). Extrication of these features can be done by using our software (see Figure 2). This software replaces all the results in the CSV file which will contain about 662 attributes.

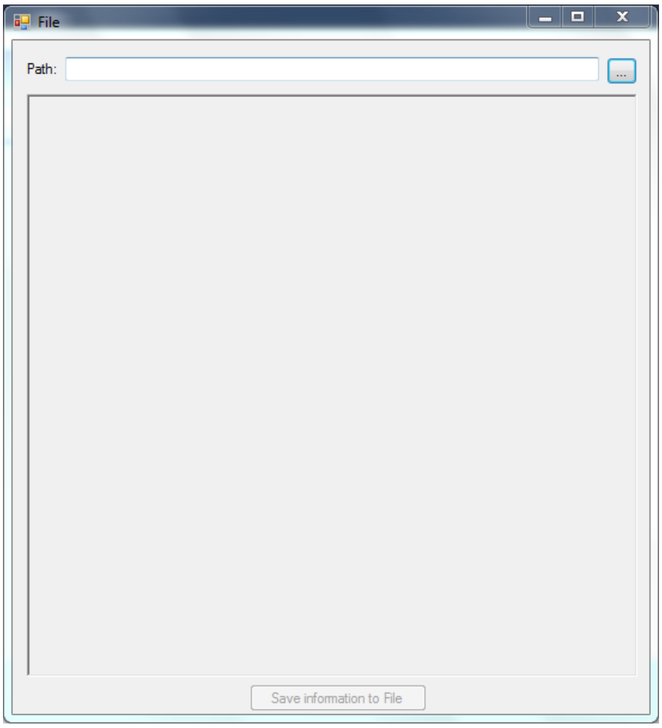


Figure 2 User face of the software.

The dataset underwent division into the training set and the testing set. The training set encompasses the known emails for all suspects, while the testing set incorporates the unknown emails. In the testing set, we substituted the email entries with "?" to be interpreted as missing values by "Rapid Miner." Consequently, the dataset is now prepared to serve as input for "Rapid Miner."

4.1. Stylometric features

One of the characteristics of online documents that CMC can utilize is stylometric features. There are hundreds of stylometric features, each defining distinctive characteristics of writing style. To be specific, they are divided into 5 main types: lexical or token-based, syntactic, structural, content-specific, and idiosyncratic features. The following paragraphs provide brief explanations for each kind [1].

First, lexical or token-based features interpret the linguistic preference of an author, such as the preferred choice of using certain words or special characters and symbols. In short, therefore, token-based features utilize characters, words, and other characteristics or metrics that are related to them, such as the frequency of capital letters or the number of words per sentence. These features can render useful information on diction, pattern, and style of writing to identify the true author [1].

Second, syntactic features are extracted from punctuation and function words to differentiate authors and find their preferences or styles. It is noted by Baayen et al. that such elements of writing are independent of the context of documents and treating them as an individual value allows differentiation of authors [1].

Third, the general appearance, format, organization and layout of the documents yield structural features. While token-based features are micro-analysis of each paragraph, looking into each character and word, structural features are macro-analysis of the whole document, looking into each paragraph and the general organization and outline: for example, number of the paragraphs per document, the average length of each paragraph, etc. are metric of structural features. The location of the header, greeting, or signature in e-mail is also considered [1].

Fourth, content-specific features take the advantage of an abundance of usage of certain keywords for a specific domain.

Fifth, as the name suggests, idiosyncratic features collect information on grammatical and spelling errors. However, since mistakes are not always idiosyncratic or habitual, it is difficult to characterize or differentiate authors by their mistakes: for example, typing mistakes are an independent variable that may have no relation to the idiosyncrasy or habit of a writer, though repeated grammatical mistake may still be able to characterize a certain author [1].

Table 1 contains a list of the Stylometric features which will be extracted in order from emails. Some of these features will be divided by the total number of words which are called “M”. The last 635 attributes are the function words, or it called sometimes stop words.

Table 1: Stylometric features

	Attribute No.	Stylometric features
Lexical Features	1	Total number of words (M)
	2	Total number of short words (less than four characters)/M e.g., and, or
	3	Average word length
	4	Average number of characters per sentence
	5	Average number of words per sentence
	6	Total different words/M
Structural Features	7	Total number of lines
	8	Total number of sentences
	9	Total number of paragraphs
	10	Average of number of sentences per paragraph
	11	Average of number of characters per paragraph
	12	Average of number of words per paragraph
	13	Has a greeting
	14	Has separators between paragraphs
	15	Indentation of paragraph Has indentation before paragraphs
	16	Use e-mail as a signature
	17	Use the telephone as a signature
	18	Use URL as signature

Syntactic Features	19-26	Frequency of punctuations (8 features) “,”, “.”, “?”, “!”, “:”, “;”, “ ’ ” ,“ ” ”
	27-662	Frequency of function words (635 features) who, while, above, what, below, for, between, both, but, by, can etc.

5. Classification

To perform classification on the dataset (for training/testing), we employed "Rapid Miner" to predict the missing values. The process involves two inputs: the training set and the testing set. Figure 3 illustrates the classification procedure, outlining the steps involved in the classification process.

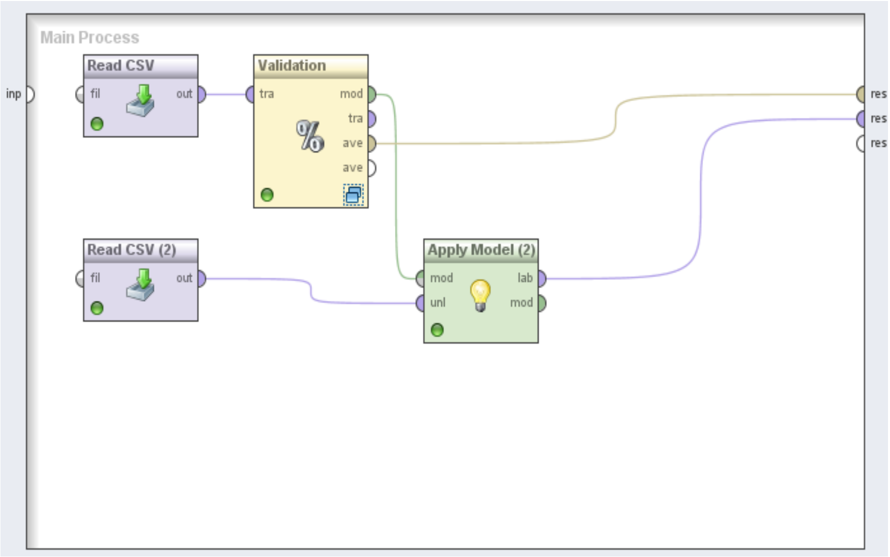


Figure 3 the main process.

Thus, we have two read CSV. The first one reads the training set and then sends it to the model inside the X-Validation. The second one reads the testing set and sends it to the applied model. Here the comparison happens to predict the missing value. Figure 4 shows inside the X-Validation. The used model here is “K-Nearest Neighbor (K-NN)”

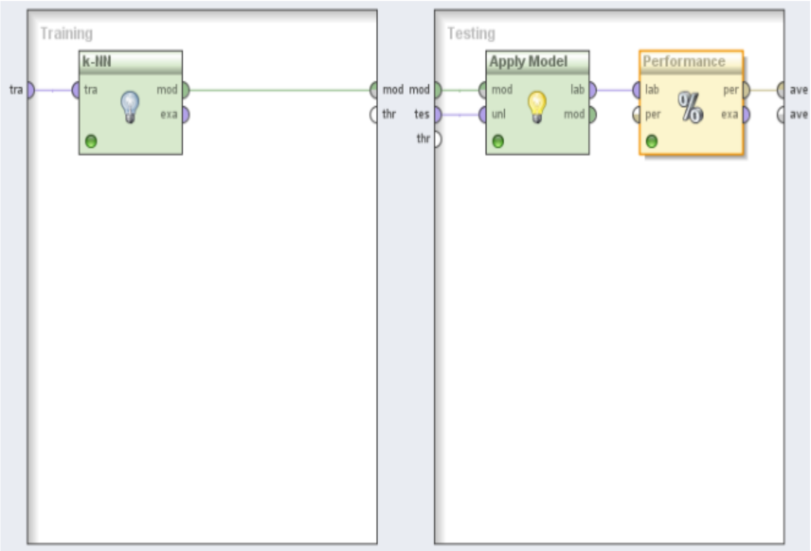


Figure 4 the used model.

5.1. K-Nearest Neighbor (K-NN)

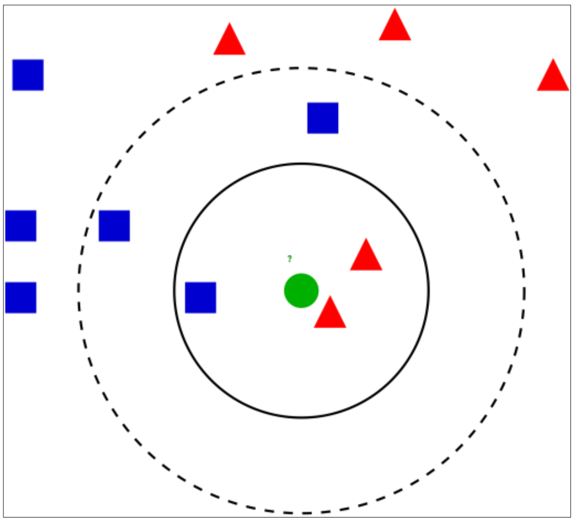


Figure 5 Example of k-NN classification. [1]

Figure 5 shows an example of k-NN. We have two classes the first class is red, and the second class is blue. The green one is unclassified. K-NN can find the nearest class to add the green one to which depends on the K value. If K=1 the green one will be with the red class, and if K=3 the green one will be with the red class because 2 reds vs. 1 blue, so the nearest class is red, but if K=5 the nearest class is blue because 3 blues vs. 2 reds. The same thing will happen to predict the author.

5.2. Prediction

Here is an explanation of how the prediction can be done:

Here is a part of the training set with only two attributes for two suspects and two known emails for each suspect (see Table 2).

Table 2: training set

le_total_words	st_lines	Author name
128	35	A
35	8	B
114	32	A
29	9	B

And here is the testing set which contains just one unknown email for one of the suspects (A or B). We know who the author (B) is, but let's assume that we do not know him (see Table 3).

Table 3: testing set

le_total_words	st_lines	author_name
47	15	?

To find or predict the author we need to follow the following formula which is based on Euclidian distance:

$$d_x^2 = \sqrt{(x_1 - x_1^d)^2 + (x_2 - x_2^d)^2}$$

x_1 = First attribute from training set which is "le_total_words"
x_2 = Second attribute from training set which is "st_lines"
x_1^d = First attribute from testing set which is "le_total_words"
x_2^d = Second attribute from testing set which is "st_lines"

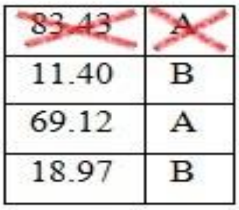
Now let's calculate each email in the training set with the testing set (see Table 4).

Table 4: calculation.

le_total_words	st_lines	$(x_1 - x_1^d)^2 + (x_2 - x_2^d)^2$	$\sqrt{}$	
128	35	$(128 - 47)^2 + (35 - 15)^2$=6961	83.43	A
35	8	$(38 - 47)^2 + (8 - 15)^2$=130	11.40	B
114	32	$(114 - 47)^2 + (32 - 15)^2$=4778	69.12	A
29	9	$(29 - 47)^2 + (9 - 15)^2$=360	18.97	B

The prediction depends on the K value, so let's set K to 3. From table 4 we need the last two columns to classifier or predict the unknown email. The nearest 3 emails (row) are required. (see table 5).

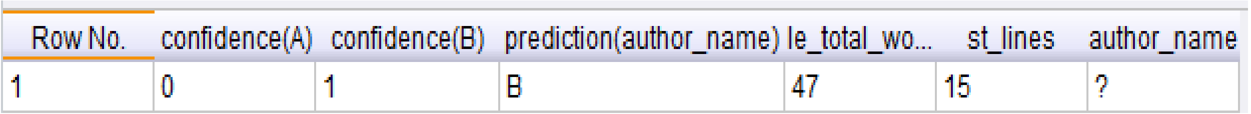
Table 5: the final step to predict the unknown email.



~~83.43~~	~~A~~
11.40	B
69.12	A
18.97	B

Now we have two Bs vs. one A which means the unknown email classifier is B.

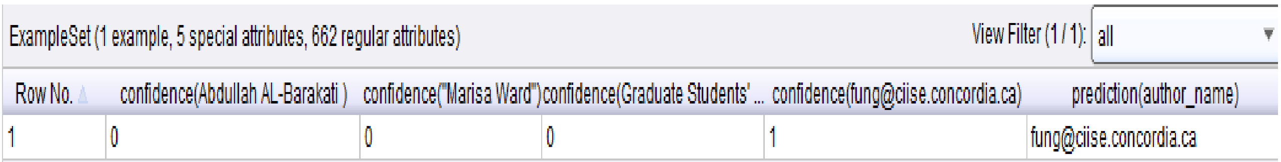
Figure 6 shows the result of the prediction using "Rapid Miner" which gave the same result as of the manual way that we did.



Row No.	confidence(A)	confidence(B)	prediction(author_name)	le_total_wo...	st_lines	author_name
1	0	1	B	47	15	?

Figure 6 the result of the prediction.

Figure 7 shows the result of prediction using "Rapid Miner" but this time with a larger data set with 100 known emails for the 4 suspects and one unknown email.



ExampleSet (1 example, 5 special attributes, 662 regular attributes)

View Filter (1 / 1): all

Row No.	confidence(Abdullah AL-Barakati)	confidence("Marisa Ward")	confidence(Graduate Students' ...	confidence(fung@ciise.concordia.ca)	prediction(author_name)
1	0	0	0	1	fung@ciise.concordia.ca

Figure 7 the result of the prediction for larger data.

6. RapidMiner

The RapidMiner tool stands out as one of the preeminent software applications in the realm of data mining, renowned for its widespread utilization. Notably, it operates as an open-source platform, offering a plethora of techniques that empower users in the domains of data mining, predictive analytics, and business analytics [7].

RapidMiner serves as a comprehensive solution, presenting an accessible methodology for extracting valuable insights from datasets. This is achieved through the application of diverse functions, including but not limited to classification, clustering, and association. By leveraging these functions and algorithms, users can effectively discern specific information amidst vast datasets, facilitating informed decision-making.

Developed in the Java programming language, RapidMiner not only excels in functionality but also in providing graphical representations such as decision trees. This visual approach enhances clarity and comprehension, making it user-friendly. Moreover, RapidMiner extends its utility to the analysis of data generated by "high-throughput instruments used in processes such as genotyping, proteomics, and mass spectrometry" [8].

7. Related work

According to Benjamin C. M. Fung, Mourad Debbabi, Liaquat A. Khan, and Farkhund Iqbal in their paper "E-mail Authorship Verification for Forensic Investigation", they discussed the problem of e-mail authorship verification and suggested a solution that handles the problem by making two models: the first model is constructed from e-mails of the potential author and the second or "universal background model" from a large dataset of e-mails from other individuals. As for the experiment on a real-life dataset, it yields an equal error rate of 17% "by employing support vector machines with RBF kernel, a regression function" [1]. Although the results are very satisfactory, the experiment has revealed some limits too. In the absence of a sample e-mail required to build a completely universal background model, the study is yet limited to an arbitrary universal model, like the one used in this study. Also, considering not only possible changes in an author's writing styles and contexts but also random and textual variations, it is only obvious that the framework, such as the SRE verification of NIST used in this study, that deals with non-textual data must be adjusted to be able to verify online documents and textual works more accurately. However, as such satisfactory results imply, there is huge potential and ground for improvement [1].

In addition, according to O. de Vel, A. Anderson, M. Corney, and G. Mohay in their paper" Mining Email Content for Author Identification Forensics ", they discussed the learning of authorship categories for aggregated and multi-topic e-mail documents, and they used some features such as structural characteristics and linguistic patterns. The classifier that they used was the Support Vector Machine learning algorithm. They applied this algorithm to a set of different emails, which belong to the authors of these emails, and they got a good result. On the other hand, they found just one author produced a bad result in performance and categorization, and they mentioned these results due to the small number of emails belonging to that author. Also, they noticed there was no improvement in the classification performance when the word collection was included, and the classification performance was decreased when they increased the function word [3].

8. Conclusion

Within the scope of this research, we elucidated the process of authorship identification and detailed our ability to anticipate anonymous emails within a cluster of communications through the application of stylometric features. This methodology guides us toward uncovering the authentic author behind these emails. Furthermore, a comprehensive analysis of all the features was conducted, systematically delving into each element to elucidate how these attributes play a role in delineating an individual's writing style.

Moreover, the vp.net language was employed to both construct and extract data from the dataset, sourced from text emails utilizing the features. These features, in turn, provide valuable insights into the identity of the email's author. Ultimately, the Rapid Miner program was utilized to forecast anonymous emails by implementing a classification function on the dataset. The outcome of this predictive analysis successfully unveiled the identity of the anonymous email correspondences.

References

[1] B. C. M. Fung, M. Debbabi, L. A. Khan and F. Iqbal, "E-mail Authorship Verification for Forensic Investigation," in Symposium on Applied Computing, New York, NY, USA, 2010.

[2] G. K. Mikros and K. Perifanos, "Authorship identification in large email collections," in CLEF, 2011.

[3] O. d. Vel, M. C. A. Anderson and G. Mohay, "Mining E-mail Content for Author Identfication Forensics," in ACM SIGMOD Record, New York, 2001.

[4] D. Madigan, A. Genkin, D. D. Lewis, S. Argamon, D. Fradkin and L. Ye, "Author Identi¯cation on the Large Scale," in the Meeting of the Classification Society of North America , North America, 2005 .

[5] P. K. Pateriya, S. Mishra and S. G. Samaddar, "Author Identification of Email Forensic in Service," India, 2010.

[6] B. Fung, "INSE 6150: E-mail Forensics," 2012.

[7] Rapid-I, "About Rapid-I," [Online]. Available: http://rapid-i.com/content/view/114/134/lang,en/. [Accessed 5 6 2012].

[8] Wikipedia, "RapidMiner," [Online]. Available: http://en.wikipedia.org/wiki/RapidMiner. [Accessed 5 6 2012].