Eurasian J. Sci. Eng., 2023, 9(2), 178-194
https://doi.org/10.23918/eajse.v9i2p14

Published by Tishk International University
Available at: https://eajse.tiu.edu.iq/

RESEARCH ARTICLE Open Access

INVESTIGATING THE EFFICIENCY OF MACHINE LEARNING METHODS IN AUTHORSHIP DETECTION FOR LOW-RESOURCED LANGUAGES: THE CASE OF KURDISH AUTHORS

Saia Hasan[1*] and Hossein Hassani[1]

[1*] Computer Science and Engineering, University of Kurdistan Hewlêr, Kurdistan Region, Iraq

Article History
Received: 13.04.2023
Revised: 06.06.2023
Accepted: 07.06.2023
Communicated by: Dr. Orhan Tug
**Email address:*
saia.hasan@ukh.edu.krd
**Corresponding Author*

Abstract:

Textual data continues to multiply with time, Alongside the exponential growth of textual information, an increase in anonymous material has also been seen. Authorship detection has significant potential for usage in numerous applications of authorship analysis, such as history and literary science, Forensic examination, or Plagiarism detection. We manually collected 2798 documents by 150 authors for this study in order to investigate how effectively existing machine learning algorithms can differentiate Kurdish authors from unidentified writings. The approach that has been developed uses a TF-IDF technique to calculate the weight of each token and extracts the token frequency of each token, ranging from 1 to 5 grams, as a feature to find a pattern in each author's text. We train SVM, CNB, MNB, and K-NN classifiers with a collection of available documents because an unknown document's essential tokens are similar to a known document's crucial tokens. Then we give it a mysterious document so it may assess how closely it resembles the known document. We achieved an accuracy of 80% by SVM with both O-V-O and O-V-R approaches for the token 1-gram, also a promising results in precision, recall, and F1-score measures. Furthermore, to our knowledge, this is the first study to investigate authorship detection for the Kurdish language.

Keywords: *Authorship Detection; NLP; Authorship Analysis; KLPT; ML; TF-IDF*

1. Introduction

Author detection is an application of Natural Language Processing (NLP). NLP is the study and application of how computers can understand and change text or speech in natural language to perform valuable tasks [1]. Authorship detection is a technique to detect the writer of the anonymous text based on training data. Kurdish is an official language of Iraq and the Kurdistan region of Iraq; besides, the Kurdish language is in use by other Kurdish people in Iran, Turkey, Syria, and other Kurdish communities around the world. However, it is considered a low-resourced language because it has few resources for natural language processing cases. This study investigates the effectiveness of existing machine learning techniques for detecting the author of the anonymous Kurdish Language Sorani dialect documents. The present chapter delivers an introduction to authorship detection by first discussing the background of the topic and the background of the Kurdish language.

1.2 Background

Authorship analysis is the art and science of distinguishing different writers' writing styles by determining the characteristics of their personas and studying articles written by them [2]. Moreover, it seeks to determine biographic information of the author, such as age, gender, and cognitive psychology. Author attribution, author verification, and author profiling are the three main tasks in authorship analysis [3; 4]. Artificial Intelligence (AI), especially Machine Learning (ML), Psychological, and Linguistics, are the three mains area of study combined in authorship analysis. ML enables the machine to learn from its previous experiment to predict the author for the current experiment. For example, word choice and sentence structure encode the author's psychological states [5]. Moreover, Linguistics, in which the combination of features is used to determine the author's writing style, such as vocabulary richness or any other properties related to linguistics [6].

Authorship detection, in its broadest terms, is both an old and a novel problem in information retrieval [7]. Since a large volume of text is now available in digital format, it is vital to identify any text to its owner to evaluate its validity [8]. The importance of the problem of authorship analysis stems from its usage in forensic investigation, humanities scholarship, and electronic commerce [8]. In addition to that, authorship detection was taken into consideration as a single-label multi-class classification. Authorship detection is applied to a document with both statistical and computational methods. The statistical approach has a long history in the literature on natural language processing. The process was first practiced in the plays of Shakespeare in the 19th century [9]. Author detection is now commonly performed using statistical and machine learning methods [10]. Author detection tries to find a pattern in each document that can help predict the document's writers' known as Stylometry, because content alone is insufficient to decide the ownership of available work [11]. This method analyzes an individual's writing style, which can be assessed to identify patterns within a text. The pattern can be the sentence's length, the words' size, or anything else that can refer to the author or act as a person's fingerprint [10]. Many scientists believe meticulous investigators can detect the author of anonymous text even if the author hides their information, including the name and affiliation [12] Authorship detection relies on a predefined set of trained data from each author to learn how to detect the patterns. Also this task can be done by statically and neural network model.

1.2 Kurdish Language

Kurdish is a language of more than 45 million of the population [13] who live in the area named Kurdistan, which is located in Eastern Asia, in the Middle East. Kurdistan borders Iraq, Iran, Syria, Turkey, and Armenia [14]. Moreover, Kurdish is a member of the Indo-European language family, namely the Irano-Aryan group [15]. Information about Kurds and the Kurdish language was written in other languages like Persian, Arabic, and Turkish such as "History of the Kurdish nation" written by Sharaf Khan (responsible for the principality of Bitlis) in the 16th century. Also, Ell Herirl, the very first well-known Kurdish poet, was birthed in 1425 in the Hakkari district and died in 1495; his favorite topics were those that his fellow citizens will discussed the most: love for the homeland, its natural beauty, and the allure of its girls [16]. The Kurdish area was the land of many battles, the Kurds have been involved in several fights over the centuries, and their geopolitical condition has always been a source of concern for international policymakers [17]. This situation affects Kurds in many areas, including economic and language growth.

Having an independent region and a safe area that was improved a few years ago allowed the Kurds to take their steps for more investigation and establishment. However, their many limitations still make the Kurdish language hard to investigate, such as the explanation given by Hassani and Medjedovic [17].

1.3 Kurdish Dialects

The Kurdish language is considered a multi-dialect language, and Sociolinguistics defines dialect as a variety of methods by which speakers of specific languages communicate [18]. Kurdish people express themselves in diverse dialects, such as Northern Kurdish (Kurmanji); speakers of this dialect are mainly in Turkey, Syria, Iraq, Iran, and Armenia. Also, Central Kurdish (Sorani) is used in Iran and Iraq [14; 19]. Gorani (also known as "Hawrami" or "Hawramani") is mainly spoken in Iran and Iraq. At the same time, Zaza (also known as "Zazaki") is used in Turkey. Most Western philologists believe that Gorani and Zazaki are linguistically different from other Kurdish dialects, but most people who speak these dialects identify themselves as Kurdish speakers [20]. The border of those dialects can be investigated and understood by a dialect continuum which is well- explained by [18]. Kurdish dialects are not different in vocabulary only but also in grammatical structure, such as the case of using gender differentiation. For instance, the Kurmanji dialect uses gender differentiation while Sorani does not. Alternatively, sometimes, root dialects do not get the benefit of gender differentiation. However, their sub-dialect does, such as the speaker of Jafar-Aabadi that uses (Laki), a sub-dialect of southern Kurdish. More grammatical issues were explained by Hassani and Medjedovic [17]; and Sheykh Esmaili [21].

Kurdish Scripts from Armenian through Cyrillic to Latin, Kurmanji has been made up of various alphabets. Today, Turkish and Syrian Kurds use a modified Turkish alphabet developed by Bedir Khan in the 1930s and 1940s, while former Soviet Union Kurds use a modified Cyrillic script. The Latin alphabet is shown in figure 3. The Arabic-based (or modified Persian-Arabic [17]) writing system for Sorani was founded in the 1920s and has undergone significant alterations on several occasions since then. As indicated by Ahmadi [22], Arabic-based has thirty-four letters. Another Script is Yekgirtû (unified) and Cyrillic, mainly used in the Kurdish area in Armenia and former Soviet countries. Like Sorani, the Gorani dialect employs modified Persian-Arabic script [17].

The paper structured as follow: the second section represents the literature review in the context of authorship detection as a branch of authorship analysis. The third section clarifies methodology to solve that current problem. The fourth section represents the result of the experiments and analysis of the data used to gain that result. Also, the findings of the research will be discussed. The fifth section concludes the current research and reviews all steps taken to solve our problem.

2. Literature Review

2.1 Rich Languages

A novel computer method for identifying the most probable author of text documents done by Ramezani [23]. He utilized the Term Frequency- Inverse Document Frequency (TF-IDF) scheme emphasizes lazy profile-based classification and adds a novel metric for detecting significant terms in texts. Consequently, the similarity between an anonymous paper and known publications is computed based on the significance of the phrases. Furthermore, the proposed method is language-independent because it operates on unprocessed textual data and does not require NLP tools for preparation. The usefulness of the Ramezani's [23] proposed solution was evaluated utilizing two English and Persian datasets, each containing six corpora with varying authors, to test the method's efficacy. Results reveal that the proposed

Technique achieves higher accuracy rates (0.902 for the English and 0.931 for the Persian datasets) compared to seven popular classifiers employing State-of-the-art stylometric features. Furthermore, additional experiments were conducted to evaluate document length's effects on the suggested solution's

Efficiency, the computation time of the proposed solution and competing classifiers, and the most efficient stylometric characteristics and classifiers. The proposed solution functions for a profile-based approach where all of an author's training documents are merged into a single document but not for a content-based approach where each document is considered a separate instance.

Luyckx and Daelemans [24] demonstrated the impact of many authors and inadequate data in authorship attribution and verification. They observed a considerable reduction in performance when they systematically expanded the number of authors. So, similar qualities function well for varying numbers of authors in their corpus; however, generalizations concerning certain characteristics are unhelpful. When dealing with restricted data, such as in forensics, memory-based learning proves to be reliable. Luyckx and Daelemans [24] used the Personae corpus, a collection of Dutch essays, as well as the classification algorithms such as K-NN and SVM. Furthermore, the results of authorship attribution trials on 145 authors show that a text from one of the 145 authors is properly classified in nearly half of the cases. However, Significant improvements are achieved by combining excellent working lexical and syntactic features. Therefore, studies claiming more than 95% accuracy on a two-author dataset are grossly inflating their efficiency and the relevance of the selected features.

They used a 145-author corpus, a more systematic investigation of the impact of different learning methods (including feature selection and other optimal solutions) on this problem. In addition, they included an examination of theprocessing of unbalanced data and experiments with alternative machine learning techniques for authorship attribution and verification.

2.2 Arabic, Persian, and Urdu

Nazir et al. [25] studied authorship attribution for Urdu, a low-resource language. However, this research also created a massive Urdu News Authorship Attribution Corpus (UNAAC-20). Because the corpus has 26,118,475 tokens derived from 21,938 news items published by 94 different authors, it can provide a more accurate portrayal of situations that may occur in the real world. They also used a snowballing strategy to find an extensive range of stylometry features, starting with a well-known study.

They then synthesized these characteristics to provide a shortlist of 194 features applicable to the Urdu language. These features are then separated into five groups: character, word, phrase, paragraph, and document-level features. They hoped many features from various groups would be beneficial as a reference for various text classification use cases. Finally, they ran 66 tests to compare the performance of classical and deep learning techniques.

A baseline features set, the five types of features, and all 194 features are employed in typical supervised learning techniques. In contrast, they employed the newly released three kinds of Urdu word embeddings for deep learning techniques. These word embeddings were developed using a vast corpus, including more than 28 million Urdu news tokens and three well-known methods: GloVe, Word2vec, and fastText. The following significant observations emerged from the examination of the experimental results:

- Using word-based features is the most successful method for authorship attribution in Urdu.

- In most situations, deep learning approaches perform far better than traditional supervised learning techniques.
- A convolutional neural network (CNN) is the most successful approach

With an F1 score of 0.99 for both datasets and an accuracy of over 99 percent, making it almost flawless. Having feature selection strategies can aid in identifying the most appropriate features.

Ramezani et al. [26] offered the findings of analyzing the effect of various textual components on Authorship attribution (AA) accuracy. This research was conducted on a massive corpus of some Persian literature that was carefully produced. For the first time in the history of quantitative AA, Iranian inspectors now have access to reliable data on which textual elements are the most beneficial for author identification in Persian. Experiments were carried out in a total of thirty distinct scenarios to collect reliable data regarding the impact that various textual elements have on the accuracy of AA. In these thirty examples, three different classification algorithms (SVM, K-NN, and C5) were utilized in two distinct learning procedures (Integrated and Disjointed), and they were applied to five different corpora with various author counts (2, 5, 10, 20, and 40). In each case, analysis was performed on twenty-nine different textual features. The most reliable criteria for AA tasks include information about the words and verbs used. The results showed that NLP-based features (Syntactic and Semantic) are more reliable than BOW-based features (Lexical and Character) when applied in real-world applications. In addition to the primary influence that textual features have on the accuracy of AA, other parameters like the number of authors, the number of documents per author, the method used to extract and weight features, and the frequency with which textual features are employed all

Have a role. In conclusion, the classification method applied to the text is one of the factors that determines how accurate the findings were. The expansion of this study can be accomplished in the future by using a mixture of the recommended features (those with a higher efficacy) to acquire more accuracy for the AA task while considering the scaling difficulty associated with the features.

Abbasi and Chen [27] applied authorship detection strategies to Arabic web- forum postings in this particular research project. They used lexical, syntactic, structural, and content-specific writing style traits to determine who the author was. To develop a model of the Arabic language that provides an appropriate amount of classification accuracy for authorship identification, they addressed some problematic features of Arabic. They also performed tests on their dataset to determine the effectiveness of various feature types and classification algorithms. These experiments are based on the data. In each test, a random selection of five writers was made, and the twenty messages were each put through a 30-fold cross-validation test using C4.5 and SVM. In the context of a multilingual online environment, emphasizing the linguistic differences between Arabic and English may give further insight into potential ways to boost the

Efficacy of authorship identification methodologies. The inclusion of other authors and data may also affect the degree of precision achieved.

3. Methodology

3.1 Research Methodology

The training dataset of supervised learning contains input and correct outputs, allowing the model to learn and improve over time. The classification algorithm uses the model to identify the category of incoming data into one of several classes or groupings, such as spam or not spam. Targets/ labels or categories are all terms that can be used to describe classes. Classification procedures with more than two class are called multi-class classifications. This study employs supervised machine learning and multi class classification to solve the challenge of authorship detection. We build a word n-gram language model from each author's labeled documents and use it to train the classification algorithms before giving the test dataset to the algorithms to predict the most likely author. Figure 1 shows the diagram for all process of Authorship detection.

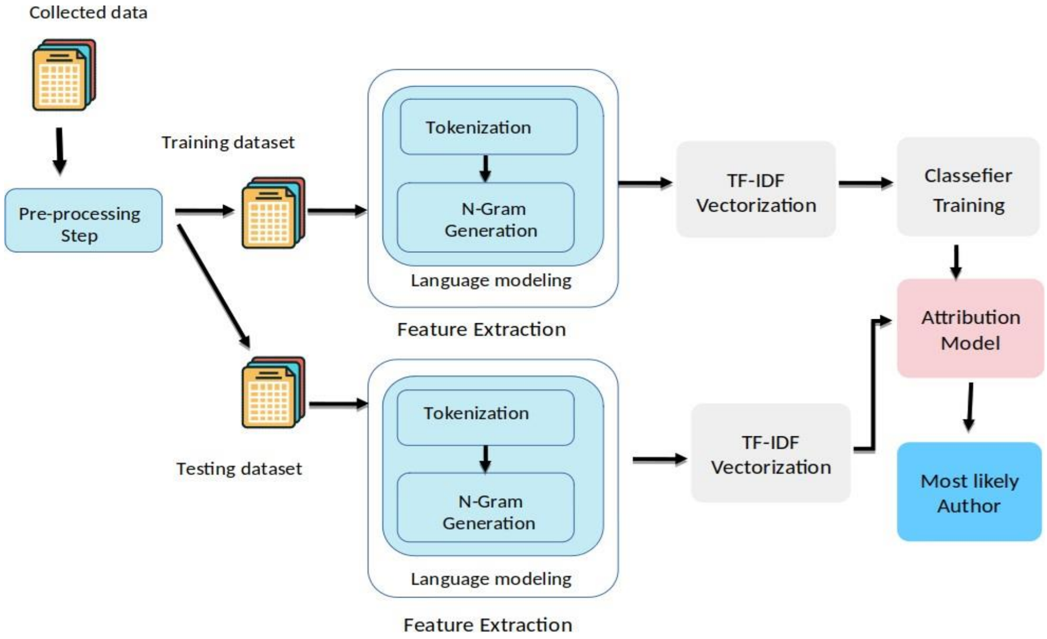

Figure 1: An overview of the authorship detection process

3.2 Data Collection

We create a corpus for the Kurdish language using news articles. The text documents used for this research may be obtained on the websites of Kurdish news agencies and magazines. we collect data from 150 authors. Also, to avoid a shift in style, we gather data for each author over three years.

Additionally, we collect data manually because there is not an API for the Kurdish language that can carry out this function. However, dataset we created will be made public after publishing this research. The dataset is unbalanced since not all authors produce the same number of documents. However, the number of words contained within the documents and the number of words for all documents for each author is also out of balance. Table 1 provides a detailed description of the dataset.

Thirteen categories are present in the dataset which are (Politic, Health and Medicine, Social live, Economy,

Religion, Philosophy, Law, Forensic Medicine, Culture, History, Geography, Critics, Environment protection, and Media & publication). Some authors provide a variety of content, like writing in the field of economy and politics, and some are experts in one fields. The authors come from various geographical areas which could affect their writing.

Table 1: Specification of the Kurdish news article dataset

Total number of articles	2798
Total number of authors	150
Total number of words	2055469
Total number of tokens	2291136
Lowest number of articles	10
Highest number of articles	34
Lowest total number of words after cleaning	3010
Highest total number of words after cleaning	78581
Lowest total number of tokens after cleaning	3447
Highest total number of tokens after cleaning	86502
Average number of tokens after cleaning	15291
Average number of articles	19
Average number of words after cleaning	13703
Authors with >=15 articles	58
Authors with < 15 articles	92
Authors with <10,000 words	112
Authors with > 10,000 words	38
Authors with <10,000 tokens	102
Authors with > 10,000 tokens	48

3.3 Pre-processing Step

However, the documents need some cleaning to match the current research requirements. The first step in the cleaning phase is to remove any additional metadata that was included during the data collection phase, such as the author's name and the document's publication date. Then, since reference lists are regarded as background noise in this research and poetry falls beyond the scope of this research, we remove reference lists and poems from any documents that contain them. After such procedures, the documents are ready for the following stages. We divided the dataset into two parts; one for training the classifiers and the other for testing the classifiers. The training consisted of a random selection of 80 percent of each author's documents, while the testing consisted of a random sample of 20 percent of each author's documents.

It is important to note that our study's current approach does not use NLP pre- processing tools such as stemming, Lemmatisation, and POS tagging.

3.4 Language Model

Using various statistical and probabilistic methodologies, language modeling determines the probability that a particular sequence of words may appear in a sentence. We construct a language model based on the probabilities of n-gram words and their sequences. Several procedures, detailed below, are carried out to get the model ready.

3.4.1 Tokenization

In Tokenization, sentences are parsed out into their component words and letters. We use the Kurdish Language Processing Toolkit (KLPT) by Ahmadi [28] to tokenize all documents.

3.4.2 N-gram Generation

N-grams, a type of statistical language modeling, are groups of contiguous words or letters that are n lengths long; n may be any number. We can determine a writer's linguistic structure, such as which character or word should come after the present one. We chose n for the n-gram as n ∈ {1, . . ., 5} because this range is used in most studies. Using the number n is dependent on the application [29].

3.5 Attribution Algorithm

This section provides the details required to determine how the new document's genuine author can be determined. We applied the following steps to resolve the attribution issue that arose from this research.

3.5.1 TF_IDF

"Term Frequency-Inverse Document Frequency" is abbreviated as TF-IDF; this is a metric used in statistics to determine how relevant a word is to a particular document among a group of documents.

This research aims to find key terms that serve as indicators for detecting a document's author. As a first assumption, we might believe that words with a high frequency in the documents are more significant and can serve as indicators. Nevertheless, this is untrue because words like (بە لە، بۆ، کە،....) that frequently occur in one document also frequently occur in other documents and cannot be the proper term for our purpose. Therefore, finding terms that frequently appear in a specific author's documents but are uncommon in those of other authors is the correct approach. The TF-IDF approach can be used to do this task.

Finding the frequency of each term in a single document is done using the TF part. The IDF part is to find out how uncommon a term is over the whole corpus. If a term (هەنار) is used frequently in the documents of one author but only sometimes in the other documents in the entire corpus, then it is crucial to recognize that author. The formula for IF-IDF is as follows:

$$TF\_IDFij = \frac{count\ of\ (i)\ inside\ (j)}{count\ of\ words\ \ inside\ (j)} \times \frac{1+M}{1+DFi} + 1 \tag{1}$$

- represents the token, and (j) represents the document.
- (M) is the total number of documents.
- DFi: number of documents that have a token (I) contained within them
- Zero divisions are avoided by adding the constant "1" to the IDF's numerator and denominator as though an additional document were viewed that included each word in the collection exactly once.

If the term (I) has a high (TF) and a low (IDF_{ij}), it is significant since it serves as an indicator for the document's owner.

3.5.2 Attribution

Using the TF-IDF technique, we provide a text document similarity measure to develop a predictor for the authorship detection problem in this work. The proposed method is predicated on the idea that a given author uses particular collocations of terms such as nouns and verbs in their work, which are referred to as tokens in this study. Consequently, the probable author of an unidentified text document can be identified by detecting the genuine author's distinctive writing patterns. In this sense, necessary tokens in an untitled document should also be required in other papers written by the unknown document's actual author.

Based on this approach, a strategy is proposed that evaluates and compares the significant tokens of the unknown document and the known documents to identify the candidate author of an anonymous text document. In the current problem, tokens prevalent in one text but rare in another effectively differentiate writing patterns. The anonymous document and the candidate document are compared for resemblance until every known document is selected as the candidate document for the first time. Then, the author of the known document with the highest similarity value is identified as the most likely author of the anonymous material. We present the solution's pseudo-code in algorithm1:

Algorithm 1: Attribution algorithm

```
Start
    Input N-gram as token in Unknown-document
    Input N-gram as token in Known-documents []
    Candidate-document = choose a document from Known-documents []
    maximum_similarity = 0
    true_author = ""
    while Candidate-document available do
        similarity = calculate similarity between Candidate-document and Unknown-document
        if similarity > maximum_similarity then
            maximum_similarity = similarity
            true_author = author of Candidate-document
        end if
        Candidate-document = choose another document from Known-documents []
    end while
    Return true_author
Stop
```

An indication of an unknown document is Unknown-document. Known-documents [] refers to the grouping of available documents. However, a Candidate-document is a candidate document whose similarity to an untitled document is evaluated.

To further distinguish between writing styles, the tokens in this study are made up of a collocation of n phrases known as N-grams rather than single terms. In other words, when identifying relevant tokens, the frequency of collocation of terms is employed.

3.6 Evaluation and Testing

We utilized mainly non-linear classifiers in our study because our problem is non-linear. In non-linear issues, data points of different classes can blend and cannot be separated linearly. In contrast, data points of different classes can be separated in linear issues by a straight line.

Three well-known non-linear classifiers, kernel SVM [30], K-NN, and NB [31], were utilized. The reason for choosing these classifiers is that they are used more frequently in previous authorship attribution studies, and their results are superior to those of other non-leaner classifiers.

3.6.1 Support Vector Machine

Support Vector Machine, or SVM, is a popular supervised learning technique used to address problems of classification and regression. However, it is most commonly used for classification problems in machine learning.

Multi-class classification is not naturally supported by SVM. It supports only binary classification by categorizing data elements into two classes. The same method is used for multi-class classification. After dividing the multi-class classification problem into several binary classification problems, we have two methods to do this task either we can do by One-to-One (O-v-O) approach or One-to-Rest (O-v-R) approach. In the ongoing study that we are conducting, we apply SVM with kernel for the O-v-R and O-v-O techniques.

3.6.2 K-Nearest Neighbours

The K-Nearest Neighbour algorithm is one of the most straightforward examples of machine learning algorithms. It is founded on the supervised learning methodology. The K-NN algorithm assumes that the newly collected and previously collected data are similar. Then, it places the newly collected data in the category most similar to the previously collected classes [30]. We employed K- NN with euclidean distance, while other distances, such as Manhattan distance and Minkowski distance, are available.

3.6.3 Naive Bayes

The Naive Bayesian Classifier uses the Bayes theorem for classification purposes, assuming that features are unrelated. Due to its simple structure, the Naive Bayesian classifier did better than some of the more complicated classification methods discussed in many research articles [32]. There are three different kinds of naive Bayes models, and they are as follows: Gaussian, Multinomial, and Bernoulli. We use Multinomial and Complement Naive Bayes (CNB) for the current research. CNB is a modification of the MNB technique that works best with unbalanced data sets [33].

3.6.4 Evaluation Measures

To determine whether or not the model is valid, we employ the confusion matrix, accuracy, precision, recall, and F1 Score metrics. Nevertheless, we offer a classification report that displays the outcomes for each class using the measurements mentioned earlier.

Accuracy, the percentage of correctly predicted observations relative to the total number of observations and is the most intuitive and elementary indicator of performance.

(2) $$Accuracy = \frac{number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

Precision is the fraction of correctly predicted positive observations as a percentage of all correctly predicted positive observations.

(3) $$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

The proportion of correctly predicted positive observations relative to the total number of observations in the actual class is known as the recall.

(4) $$\text{Recall} = \frac{\mathit{True\ Positive}}{\mathit{True\ Positive} + \mathit{False\ Negative}}$$

The weighted average of Precision and Recall is the F1 Score.

(5) $$\text{F1} - \text{Score} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right)$$

4. Result and Discussion

4.1 Result

We used TF-IDF to vectorize each documents in both tearing and testing portions. Then we trained the classifiers using the TF-IDF results of the training part. The output was (224488) features for the 1-gram, (1045933) features for the 2-gram, (1622045) features for the 3-gram, (1758361) features for the 4-gram, and the 5-gram (1779042) features. The SVM classifiers with both O-v-O and O-v-R technique produced the best accuracy, macro and weighted precision, macro and weighted recall, and weighted F1 score for the 1- gram feature. The accuracy rate was 80%. 83% for macro precision and 84% for weighted precision, 76% for macro and 80% for weighted recall, and 77% for macro and 79% for weighted F1-score as shown in figure 3. On the other hand, the MNB classifiers produced the lowest accuracy, macro, and weighted precision, macro and weighted recall, and macro and weighted F1 score for the 5-gram features. The accuracy was %05, %09, and %13 for macro precision and weighted precision, respectively, %03, and %05 for macro recall and weighted recall, respectively, as well as %04 and %06 for macro and weighted F1-score, respectively. Table 2 lists the 1-gram to 5-gram accuracy, precision, and recall for each of the four classifiers (CNB, MNB, SVC (O-v-O), SVC (O-v-R), and K- NN). In addition, we select K = 4 as the number of neighbours for the K-NN classifier. For SVC, we utilize a linear kernel with C set to 2 (both O-v-O and O- v-R technique).

With a substantial difference in all grams, the CNB outperforms the MNB in all measurements. SVM with the O-v-R approach gave the same result as the O-v- O technique. Differences in gram-level accuracy between classifiers are depicted in Figure 2.

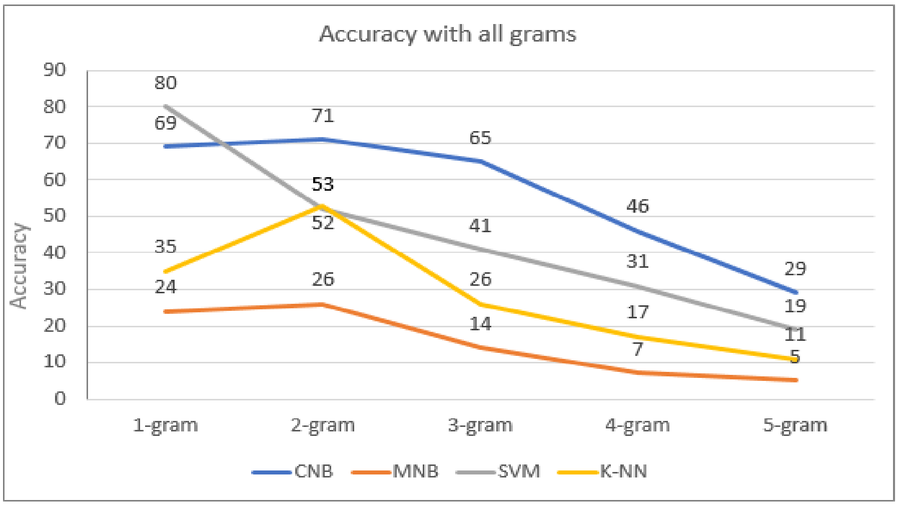

Figure 2: Accuracy throughout the entire range of grams for every single classifier (SVM both technique has same result)

Table 2: Results of evaluation measures in classification process with full dataset for 1 to 5-grams tokens

N-Gram	Measurement		CNB%	MNB%	SVM (O-v-O) %	SVM (O-v-R) %	K-NN %
1-gram	Accuracy		69	24	80	80	35
	Precision	M. A	72	16	83	83	40
		W. A	73	22	84	84	43
	Recall	M. A	65	17	76	76	34
		W. A	69	24	80	80	36
	F1-Score	M. A	64	13	77	77	31
		W. A	66	18	79	79	32
2-gram	Accuracy		71	26	52	52	53
	Precision	M. A	75	22	54	54	59
		W. A	75	29	60	60	62
	Recall	M. A	68	18	44	44	52
		W. A	71	26	52	52	54
	F1-Score	M. A	68	17	45	45	50
		W. A	69	22	51	51	52
3-gram	Accuracy		65	14	26	26	41
	Precision	M. A	68	15	30	30	48
		W. A	68	20	37	37	51
	Recall	M. A	63	10	20	20	40
		W. A	66	14	26	26	41
	F1-Score	M. A	62	10	21	21	39
		W. A	63	13	26	26	41
4-gram	Accuracy		46	07	17	17	31
	Precision	M. A	48	10	23	23	36
		W. A	50	15	28	28	38
	Recall	M. A	45	05	14	14	30
		W. A	46	07	18	18	32
	F1-Score	M. A	43	05	16	16	29
		W. A	44	08	19	19	30
5-gram	Accuracy		29	05	11	11	19
	Precision	M. A	39	09	18	18	30
		W. A	41	13	22	22	32
	Recall	M. A	29	03	09	09	19
		W. A	29	05	11	11	19
	F1-Score	M. A	30	04	10	10	21
		W. A	31	06	12	12	21

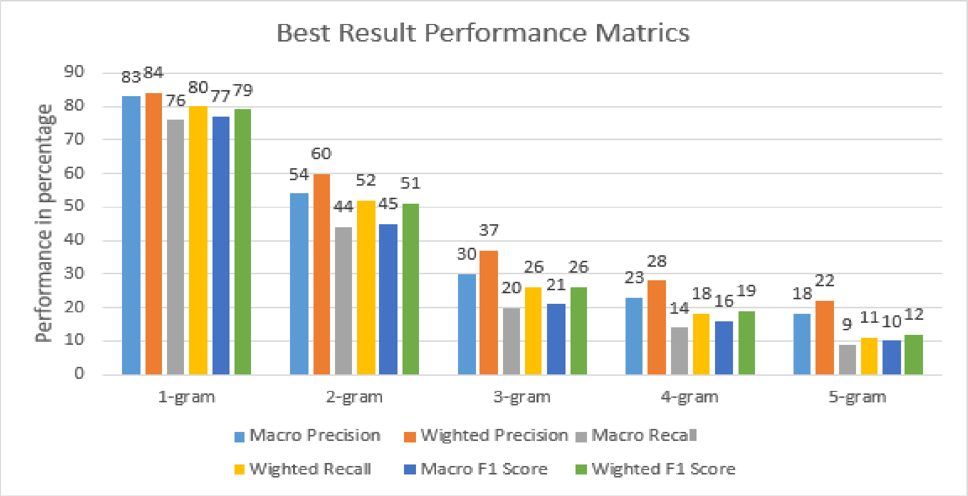

Figure 3: Precision, recall and f1-score of the best result (result of SVM with O-v-O & O-v-R is same)

4.2 Discussion

Luyckx and Daelemans [24] claim that most research that uses statistics or machine learning to figure out who wrote something focuses on two or a small number of authors. This strategy makes it seem like the training data features are more important than they are and have proven to be biased towards these few groups of authors. Also, most studies also exaggerate the accuracy of their method by using training data volumes that are unreasonable for the context in which stylometry is used, such as forensics [24]. Using more author's data that acts as a negative instance can make this task more realistic. Moreover, most of the time, when authorship is being attributed, it can cause some authors to have very little data and others to have more [34]. Therefore, to be more realistic, we utilized the imbalance dataset and 150 writers. As indicated in the section above, CNB outperformed MNB for for all n-grams. The specific cause of the situation is that traditional MNB's two issues—skew bias data and assuming that characteristics are independent—were solved by CNB.

In the 1-gram SVM beats CNB due to the reduced dimensionality of the data compared to other grams. According to documents from Sikit-Learn and Hiran [30], SVM is effective in environments having high dimensions. Still efficient in situations where the number of dimensions exceeds the number of samples. But its efficiency declines when the number of features far exceeds the number of samples.

With an accuracy of 80%, word 1-grams provided our best result by the SVM classifier for both (O-v-O and O-v-R) techniques. 84% for weighted precision and 84% for macro precision. 76% for weighted recall and 80% for macro recall. Macro F1 scores were 79%, and weighted F1 scores were 77%. Because of the following three factors, most of the studies we cited in the literature review were more accurate than we did.

- There were fewer classes, but each class had many instances. Most authorship attribution studies focus on a minimal group of authors and report positive findings. However, performance drastically decreases when the number of authors being investigated gradually increases.
- Their dataset is balanced. Training dataset imbalance was a problem faced in this research. The issue of class imbalance affects various text classification tasks, one of which is the detection of authors. There is a limited number of training texts available for specific authors, but many training texts are available for other authors.

- Very effective pre-processing tools or various feature types andcombinations were utilized. A mix of lexical, syntactic, structural, and content-specific features was used by employing POS tagging to increaseaccuracy. However, because of the limited availability of such technologyat the moment, especially POS tagger, which is unavailable in Kurdish butavailable in other languages, we cannot make use of it now.

Each relied on one or more of those three factors to achieve such remarkable results.

Our work is similar to that of Luyckx and Daelemans [24], who obtained poor accuracy by using 145 authors and very little data for each author. However, However, the combination of features leads to their results being more accurate. On the other hand, a comparison with others studies showed that their results were still considered as not very accurate. Another example is the work of Nazir et al. [25], who utilized the Imbalance dataset with 94 authors and experienced poor accuracy. Ramezani et al. [26] used forty different authors for their investigation; however, they only used a small amount of data from each author, which led to low accuracy. Table 3 shows the comparison between our work and other work that resemble ours in certain respects.

Table 3: Accuracy comparison of our results with other studies that resemble ours in certainrespects.

Research	Dataset	Number of document per author	Number of Author	Accuracy
Our Research	2798 documents	unbalanced	150	80%
[24]	Not available	1400 word	145	50%
[25]	21938 documents	unbalanced	94	74%
[26]	200 documents	5	40	43%

On the other hand, partitioning the dataset has an effect not only on the performance of each class but also on the performance of the entire dataset.

If we wish to boost accuracy and other measurements, we must implement some or all of the following procedures.

- Creating a dataset balanced by either providing more data to classes withfewer documents or employing resampling strategies.
- Using other types of features or language models and combining features.

5. Conclusion and Future Work

5.1 Conclusion

We studied the use of machine learning to detect authorship in the Sorani dialect of the Kurdish language, which uses a modified form of the Persian-Arabic script. The Kurdish language is regarded as a low-resource language regarding the availability of resources for natural language processing. Especially when it comes to our situation, based on this fact, we gathered data manually, and we produced a dataset consisting of a volume of 2798 documents and 150 authors.

We contributed a model to our dataset that can identify the author of an untitled document. Additionally, we did not make use of any heavy NLP pre-processing tools. As a language model, we used n-grams, and the frequency of individual tokens, to train several classifiers. We estimated token weights using TF-IDF. The suggested approach compares the unknown document's essential tokens against those of known documents to determine its author. In this problem, common tokens in one document but rare in others differentiate writing styles.

We used two different versions of the Naive Bayes classifier, The Support Vector Machine classifier and the K-nearest neighbour classifier. Using Support Vector Machine classifier, we reached an accuracy of 80% in word 1-grams. During our investigation, we also explained the successful outcomes of several studies conducted in this area.

5.2 Recommendation for Further Research

Additional research could use the POS tagger tool to extract extra features and combine them with the features of this study to gain more performance. Alternatively, they could perform some additional pre-processing steps, such as stemming (elimination of suffixes from words). Lemmatization (grouping many inflected forms of an expression to be studied as a single unit). Both lemmatization and stemming are outside the scope of this study due to time constraints. One other step that other researchers may take to improve classification accuracy is resampling the dataset to make it more balanced. Also, other researchers can utilize alternative models, such as neural networks, to compare our findings to their own. Investigating other Kurdish dialects and scripts. Examining different forms of text, such as novels, poems, and tweets.

6. Conflict of interest

There is no conflict of interest for this paper

References

[1] Chowdhury GG. Natural language processing. Annual Review of Information Science and Technology. 2003; 37(1): 51–89. https://doi.org/10.1002/aris.1440370103

[2] Roy N. Authorship Analysis as a Text Classification or Clustering Problem. 2019.

[3] Iqbal F, Debbabi M, Fung BC. Machine learning for authorship attribution and cyber forensics, Heidelberg: Springer. 2020; 52-55.

[4] Stamatatos E, Daelemans W, Verhoeven B, Potthast M, Stein B, Juola P, Sanchez-Perez MA, Barrón-Cedeño A. Overview of the author identification task at PAN 2014. In CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK. 2014; 1-21.

[5] Sikos J, David P, Habash N, Faraj R. Authorship analysis of inspire magazine through stylometric and psychological features. In 2014 IEEE Joint Intelligence and Security Informatics Conference, IEEE. 2014; 33-40. https://doi.org/10.1109/JISIC.2014.15

[6] Ahmed H. The role of linguistic feature categories in authorship verification. Procedia computer science. 2018; 142: 214-221. https://doi.org/10.1016/j.procs.2018.10.478

[7] Juola P. Authorship attribution. Foundations and Trends® in Information Retrieval. 2006; 1(3): 233-334.

[8] Tamboli MS, Prasad RS. Authorship analysis and identification techniques: A review. International Journal of Computer Applications. 2013 Jan 1; 77(16).

[9] Hriez S, Awajan A. Authorship Identification for Arabic Texts Using Logistic Model Tree Classification. InIntelligent Computing: Proceedings of the 2020 Computing Conference, Volume 2 2020 (pp. 656-666). Springer International Publishing. https://doi.org/10.1007/978-3-030-52246-9_48

[10] Farahmandpour Z, Nikmehr H. A Study on Intelligent Authorship Methods in Persian Language. Journal of Computing and Security. 2015 Jan 1; 2(1): 63-76.

[11] Daelemans W. Explanation in computational stylometry. InComputational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II 14 2013; 451-462. https://doi.org/10.1007/978-3-642-37256-8_37

[12] Payer M, Huang L, Gong NZ, Borgolte K, Frank M. What you submit is who you are: A multimodal approach for deanonymizing scientific publications. IEEE Transactions on Information Forensics and Security. 2014 Nov 6; 10(1): 200-12. https://doi.org/10.1109/TIFS.2014.2368355

[13] Pasdar Y, Najafi F, Moradinazar M, Shakiba E, Karim H, Hamzeh B, Nelson M, Dobson A. Cohort profile: Ravansar Non-Communicable Disease cohort study: the first cohort study in a Kurdish population. International journal of epidemiology. 2019 Jun 1; 48(3): 682-3f. https://doi.org/10.1093/ije/dyy296

[14] Hassanpour A, Sheyholislami J, Skutnabb-Kangas T. Introduction. Kurdish: Linguicide, resistance and hope. International Journal of the Sociology of Language. 2012 Sep 13; 2012(217): 1-8. https://doi.org/10.1515/ijsl-2012-0047

[15] Windfuhr G. ed. The Iranian languages. 1st Edition ed. London: Routledge. 2009; 418.

[16] BLAU J. The Kurdish Language and Literature. Fondation-Institut kurde de Paris, Available at: https://www.institutkurde.org/en/language/

[17] Hassani H, Medjedovic D. Automatic Kurdish dialects identification. Computer Science & Information Technology. 2016 Feb 6; 6(2): 61-78. https://doi.org/10.5121/CSIT.2016.60307

[18] Khalid HS. Kurdish dialect continuum, as a standardization solution. International Journal of Kurdish Studies. 2015;1(1):27-39. https://doi.org/10.21600/ijks.95271

[19] Esmaili KS, Salavati S. Sorani Kurdish versus Kurmanji Kurdish: an empirical comparison. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013 Aug; 300-305.

[20] Taucher W, Vogl M, Webinger P. Refworld | The Kurds: History –Religion – Language - Politics. [online] Refworld. 2015. Available at: https://www.refworld.org/docid/568cf9924.html.

[21] Esmaili KS. Challenges in Kurdish text processing. arXiv preprint arXiv:1212.0074. 2012 Dec 1. https://doi.org/10.48550/arXiv.1212.0074

[22] Ahmadi S. A Formal Description of Sorani Kurdish Morphology. arXiv preprint arXiv:2109.03942. 2021 Sep 8. https://doi.org/10.48550/arXiv.2109.03942

[23] Ramezani R. A language-independent authorship attribution approach for author identification of text documents. Expert Systems with Applications. 2021 Oct 15; 180: 115139. https://doi.org/10.1016/j.eswa.2021.115139

[24] Luyckx K, Daelemans W. Authorship attribution and verification with many authors and limited data. InProceedings of the 22nd international conference on computational linguistics (COLING 2008) 2008 Aug; 513-520.

[25] Nazir Z, Shahzad K, Malik MK, Anwar W, Bajwa IS, Mehmood K. Authorship Attribution for a Resource Poor Language—Urdu. Transactions on Asian and Low-Resource Language Information Processing. 2021 Dec 14; 21(3): 1-23. https://doi.org/10.1145/3487061

[26] Ramezani R, Sheydaei N, Kahani M. Evaluating the effects of textual features on authorship attribution accuracy. InICCKE 2013 2013 Oct 31; 108-113. https://doi.org/10.1109/ICCKE.2013.6682828

[27] Abbasi A, Chen H. Applying authorship analysis to Arabic web content. InIntelligence and Security Informatics: IEEE International Conference on Intelligence and Security Informatics, ISI 2005, Atlanta, GA, USA, May 19-20, 2005. Proceedings 3 2005; 183-197. Springer Berlin Heidelberg. https://doi.org/10.1007/11427995_15

[28] Ahmadi S. KLPT–Kurdish language processing toolkit. InProceedings of second workshop for NLP open source software (NLP-OSS) 2020 Nov; 72-84. https://doi.org/10.18653/v1/2020.nlposs-1.11

[29] Anwar W, Bajwa IS, Ramzan S. Design and implementation of a machine learning-based authorship identification model. Scientific Programming. 2019 Jan 16; 2019. https://doi.org/10.1155/2019/9431073

[30] Hiran KK, Jain RK, Lakhwani K, Doshi R. Machine Learning: Master Supervised and Unsupervised Learning Algorithms with Real Examples (English Edition). BPB Publications; 2021 Sep 16.

[31] Brownlee J. Machine learning algorithms from scratch with Python. Machine Learning Mastery; 2016 Nov 16.

[32] Tan RHR, Tsai FS. Authorship identification for online text. In 2010 International Conference on Cyberworlds, IEEE. 2010; 155-162. https://doi.org/10.1109/CW.2010.50

[33] Seref B, Bostanci E. Performance comparison of Naïve Bayes and complement Naïve Bayes algorithms. In2019 6th international conference on electrical and electronics engineering (ICEEE) 2019 Apr 16; 131-138. https://doi.org/10.1109/ICEEE2019.2019.00033

[34] Stamatatos E. Author identification using imbalanced and limited training texts. In 18th International Workshop on Database and Expert Systems Applications (DEXA 2007) 2007 Sep 3; 237-241. https://doi.org/10.1109/DEXA.2007.5