JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 06 Issue 01
2023

Research Article
https://doi.org/10.56979/601/2023

A Hybrid Approach for Analysis of Urdu Tweets Authorship Empowered by Genetic Algorithm and K-Nearest Neighbors

Zain Ali[1*], Arfan Ali Nagra[1], Khalid Masood[1], Muhammad Abubakar[1], and Muhammad Mudassar[2]

[1]Department of Computer Science, Lahore Garrison University, Lahore, 54000, Pakistan.
[2]Department of Technology, The University of Lahore, Lahore, 54000, Pakistan.
Corresponding Author: Zain Ali. Email: zain.ali90890@gmail.com

Received: November 11, 2023 **Accepted:** November 27, 2023 **Published:** December 05, 2023

Abstract: Authorship attribution is the process of identifying the author of a puzzling report from a jumble of unclear material. As the world moves toward more constrained exchanges, Internet crimes such as phishing and harassment are becoming more common. Consequently, locating culprits during cybercrime investigative processes is a challenge. This research evaluates current authorship attribution algorithms on a semantic level, as well as the accuracy rate in Urdu situations. Urdu language datasets were used as Urdu TD, which is based on 600 Urdu tweets per author. The LDA model was used to chip away at stylometry elements to distinguish the composing style of specific authors using the n-gram method and cosine similarity. After applying the LDA model for feature selection, we used a genetic algorithm. After obtaining the features we applied the KNN classifier. The idea of combining the genetic algorithm and KNN classifier is to create a hybrid model that outperforms each classifier in terms of classification accuracy. In this study, the proposed authorship attribution model had an excellent ability to classify simple and different Urdu languages, with the highest accuracy of 98.20%, recall of 99%, precision of 97%, and f1 measure of 98%. The task was managed without utilizing any labels for authorship. This system should help improve standards for authorship attribution and classification methods.

Keywords: K-nearest neighbors (KNN); latent dirichlet allocation (LDA); Natural language toolkit (NLTK); term frequency; Inverse document frequency (TF-IDF).

1. Introduction

The manual analysis uses multiple tests and analyses on a given text material to determine author qualities. Statistical analysis uses statistical methodologies for adjusting record numeric values on a particular data collection to identify the author. A suggested method is tested utilizing shorter messages such as micro messages which are composed of 140-character blocks of text [1]. Shorter communications consisting of blocks of text of 200, 400, and 800 characters from Twitter feeds are used to assess the suggested technique. This research intends to authorship analysis by elucidating how the authorship set of rules operates on a linguistic level, particularly in the Urdu context. On Twitter, Urdu data gathering is used for this purpose. A tweet is a single post on Twitter, which is a microblogging website. Tweets are a fantastic illustration of one of the fastest forms of communication now in use. In recent years, it has grown in popularity on the internet [2]. The restricted number of characters in each post is one of the reasons for its popularity. The limitation is that posts be no more than 280 characters long. It is not a necessity that posts be 280 characters or less in length; the most common length of a tweet is 33 characters. Only approximately 9% of tweets have ever gone beyond the 140-character limit on Twitter. We are motivated to promote Urdu since it is a language that receives little attention. On the internet, short communications have become fashionable. Users upload brief messages to the 'chat room,' which are subsequently available to all users in this chat room [4]. Facebook offers a variety of ways for users to communicate with one another,

including inbox messaging, wall postings, and comments, with a focus on brief communications. YouTube is another popular online entertainment platform that provides comment areas for users to share their opinions and thoughts concisely. However, as a result of these advantages, science has flaws, such as cybercrime.

The distribution of illegal content through the internet is a widespread type of cybercrime. Such items include pirated software, child pornography, and stolen property. Cybercriminals have exploited a variety of Web-based methods to disseminate illicit information. People normally do not need to reveal their true identifying information, such as their name, age, gender, or residence. Cybercrime carried out through such different channels offers particular hurdles for law enforcement organizations in terms of criminal identity tracking as compared to traditional crimes. Cybercrime is expanding over the world due to the usage of computer networks and network activities. Spreading information on the internet is one sort of cybercrime that can lead to criminals stealing information and spreading it online. Many online forums are used by cybercriminals to propagate unlawful information. These channels are characterized by their impersonality. In most cases, people do not reveal that they are true identities in internet forums. Cybercrime differs from other types of crime in that it occurs through online activities. Online forums circumstances make it more difficult for law enforcement authorities (LEA) to track down the actual offender. The LEA urgently requires a solution that will help investigators to detect and track actual theft.

Rong Zheng [7] suggested a methodology that might be used to automatically track criminals using social media posts. Message attributes and content-specific information are removed from the framework and used to influence learning algorithms to create feature models that may be used to identify the authors of unlawful communications. To test the framework's accuracy, researchers used data sets of Chinese and English internet newsgroup messages. The results reveal that the suggested technique is capable of accurately identifying the authors of English and Chinese online communications [8]. This method aids in the identification and tracking of the cybercrime investigation setting.

The emergence of qualities that remain for a range of writing texts written by some writers is referred to as authorship analysis. The features are divided into three categories [7]. Recurrence terms, sentence length, punctuation quantity, and vocabulary lavishness are examples of style indicators, which are also known as content-free qualities. Second, the phrases "thank you" and "goodbye" are structural parts. Finally, for original material, keyword prevalence and unique nature are content-specific criteria.

Based on the supplied Urdu Twitter data set, we are particularly interested in assessing the author and overall performance as well as the implications of different sorts of characteristics in the context of cybercrime investigation. We are especially interested in analyzing the suggested framework's usefulness in a multilingual setting, given the transnational character of cybercrime. This aids law enforcement authorities (LEA) in dealing with the identity tracking challenge in cybercrime investigations without a linguistic barrier. To identify the author of a particular text, we used three types of features from online unlawful communications that have been reported in authorship analysis research and identification [6, 9]. The applicability of the suggested approach in a multilingual situation is of special interest to us. The remaining research is organized as follows. We will talk about some previous research on authorship attribution and several types of text features and methodologies in the next section. The approach was the main focus of the third segment. The fourth portion delves into the experimental outcomes of our model for attribution of authorship to Urdu tweets, as well as a comparison to prior results covered in this segment. This work will be completed in the last portion.

2. Literature Review

The Urdu authorship attribution is used to indicate the authorship of a work. To identify a certain author, a text can be analyzed using a variety of ways. Character and byte-level matching approaches are available, as well as letter and word-level stochastic systems. These include ways that use Parts of Speech (POS) tagging to build a probabilistic model based on word classes and compare the given text to a previously tagged corpus of the author in question. Suggest author profiles, a character-based recognition system that uses n-gram frequencies at the byte level. Because their technology is based on character-level processing, it is essentially language agnostic. The author profiles are created based on the author's text. For the test data, similar profiles are created as well. Their approach differs from the traditional perplexity and cross-entropy based n-gram matching technique that is used in speech pattern matching for voice

recognition and is relatively common. Techniques based on maximum likelihood estimation (MLE) can also be applied.

2.1. Approaches For Authorship Analysis

In terms of writing style, author recognition highlights the solitary author's likelihood of authoring the document in comparison to other authors writing styles. Stylometry research has mostly focused on forensic authorship analysis, which has aroused a lot of interest and resulted in a lot of research over the years [8, 9]. Initially, authorship analysis was done using statistical approaches. The reason for this is that various writers write in different styles, which distinguishes distinct word diffusions. As the number of writers and writing files grows, it becomes more difficult to identify new papers, which are referred to as a statistical hypothesis test or classification difficulty. As a result, at first, authorship analysis was done using statistical approaches. In lexical data analysis, Brainerd [10] employed statistical distribution approaches. Thirsted [11] developed a crucial statistical test. In authorship analysis, Farringdon [12] introduced the CUSUM approach. The statistical methodologies for authorship analysis are summarized by Francis [13]. The CUSUM approach, according to Holmes [14], is unreliable since it cannot predict the true author across several texts.

Authorship analysis was made possible by the introduction of contemporary machine learning algorithms. Mosteller and Wallace [15] oversee a Bayesian model to assess the paper, McCallum, and Nigam [16] distinguish two Bayesian models for text categorization. Although this version has structural restrictions, many powerful approaches are used to identify the writers. Neural networks are the most illustrative. To assign authorship for works, Tweedie [17] employed an effective artificial neural network known as a multilayer perceptron. The outcome was consistent with earlier research on the same subject. Lowe and Matthews [18] employ a radial basis function neural network (RBF). They used RBF to learn more about Shakespeare's work [19] and his cooperation with contemporary John Fletcher on several plays. Khmelev [20] offers a new technique for authorship identification based on the Markov chain, which employs the odds of succeeding letters as characteristics. Diederich [21] was the first to propose the assistance vector machine as a solution to this problem. In 60 percent to 80 percent of cases, our method recognizes the spotted authors. The author's discovery of digital communications on the concept of message labor is a fresh area to look at. De vel [22] used an assist vector machine to categorize 150 email files from three writers using a set of criteria he learned.

When comparing the two methodologies, system learning strategies outperform statistical methods in terms of accuracy. They can categorize the use of prepared personal phrases using a variety of characteristics. Based on the review, the following are all of the methods for authorship attribution that were previously employed and are now being used. The manual analysis uses multiple tests and analyses on a given text material to determine author qualities. Statistical analysis uses statistical methodologies for adjusting record numeric values on a particular data collection to identify the author. On a collection of statistics, Machine Learning uses classification algorithms to determine the author of a work phase.

Current semantic analysis approaches in natural language processing (NLP) are inadequate [29]. These characteristics reflect how an author organizes a text. They were discovered when authorship attribution was applied to emails and online forum postings. Paragraph length, signature use, font color, and font size are all examples of these metrics. Because it is difficult to capture stylistic aspects of brief texts, structural features are important when crediting them. These characteristics are unique to a given language. These characteristics' measurements must be specified manually as the primary difficulties to consider while handling an authorship attribution problem. Every word and sentence represents a feature. And optimize the cataloging process, feature selection is critical. Wrappers and filters are the two basic categories of feature selection techniques. In our case, we achieve feature selection using a genetic algorithm with the KNN classification component providing more exact results.

This study aims to further the field of authorship analysis by demonstrating how the authorship set of rules works on a linguistic level, specifically in the setting of Urdu. This is accomplished by the usage of Urdu data collection on Twitter. A tweet is a single post on the microblogging website Twitter. Tweets are an excellent example of one of the quickest ways of communication now in use. It has gained in popularity on the internet in recent years, with one of the reasons being the limited number of characters allowed in each post. The maximum character length for posts is 280 characters. The length of a tweet does not have to be 280 characters or fewer; the most common length is 33 characters. Only about 9% of tweets on Twitter

have ever gone over the restriction of 140 characters [3]. We are motivated to promote Urdu since it is a language that is underappreciated. The following questions will be addressed by this study. How effective are existing authorship attribution methods for attributing attribution to Urdu-language tweet writers? For an accurate author profile, how many tweets per writer are required? Is there a difference in accuracy between increasing and decreasing the number of tweets per author?

3. Methodology

The frameworks for authorship identification, the actions to take, and the corpora are all discussed in this part, which are also represented in Figure 1. Tokenization, lower casing, n-grams, feature extraction, feature selection via a genetic algorithm, document word matrix building, topic extraction, and KNN classification were applied to the entire data set.

3.1. Document Preprocessing

Authorship is determined only by the author's writing style [5]. In the writing survey, it was also discovered that cleaning the data set by removing special characters or correcting grammatical errors is not possible and that the author's preference for word suffixes and prefixes, as well as later capitalization, all provide important information about the author. As a result, eliminating or correcting such items will reduce the number of characteristics associated with a certain author.

3.1.1. Tokenization

By eliminating white spaces, it splits phrases or paragraphs into little units such as words or characters. With the aid of the Natural Language Toolkit, save each token in the dataset together with their frequency of occurrence (NLTK).

3.1.2. Lowercasing

Before any further processing, all upper case text is transformed to lower case. Because the Urdu language only has one case, there is no need to lowercase it.

3.1.3. N-Grams

For every language, an N-gram combines contiguous words or letters of length n. For example, the n-grams include works that are equally accurate on a language level. It can readily capture a writer's linguistic structure and writing style. The n-precision grams are determined by the value of n. important distinctions may not be captured for small n values, however for big n values, lengthy n-grams will be produced, causing limits, and we can stick to a few specific examples. In a word-level n-gram, a good approach is to use n-grams where n ∈ {1, 2, 3,..5}.

Table 1. N-grams (1–5) for the sentence "میں اس مضمون کا مصنف ہوں"

N-Gram Types	Sentence Representation
Word Unigram	میں، اس، کاغذ کا، مصنف، ہوں۔
Word bigram	میں_ہوں، کاغذ کا، اس_کا_مصنف، اس_کا_مصنف
Word Trigram	میں_اس_کاغذکا_مصنف، اس_کا_مصنف، اس_کا_مصنف ہوں
Word Fourgram	میں_کا_مصنف ہوں، اس_کا_مصنف، اس_کاغذ کا_مصنف ہوں
Word Fivegram	میں_اس_کا_مصنف ہوں، اس_کاغذ کا_مصنف ہوں

Table 1 shows the total arrangements of words Unigram, Bigram, Trigram, Fourgram, and Fivegram formed from a simple statement "میں اس مضمون کا مصنف ہوں" To make things easier to grasp, underscores (_) are utilized to replace spaces. Because relevant content is lost in the bag of words approach, an n-gram is utilized to capture all of the more semantically useful data from text. It also correctly identifies the gender of tweeters [27].

3.1.4. Remove Stop Words

"A, are, and is, the, this gets" are stop words in the English language. Which words have a high frequency in language content but very little lexical information? As a result, we decided to eliminate these terms from the dataset before proceeding with any further processing. However, due to the Urdu language, such a list of stop words does not exist. All terms contained in papers receive a 70 percent discount.

3.1.5. Words Stemming

Separating the root word from the provided words yields the stem word. With the help of the NLP tool, a rule-based stemmer was used to stem dataset words.

3.2. Syntax Analyzer And Feature Extraction

It is a feature, which is a means of extracting numerical information from documents. A model with only the best-fit attributes is chosen for training to improve results. This is accomplished through the usage of TF-IDF.

3.2.1. TF-IDF

It makes it possible to determine the frequency of the terms of each word, the raw frequency, and the inverse frequency of the document (TF-IDF). Generates vectors of different characteristics from the information recorded by the texts, making it possible to distinguish the author of the material. Figure 1 shows the proposed authorship attribution methodology.

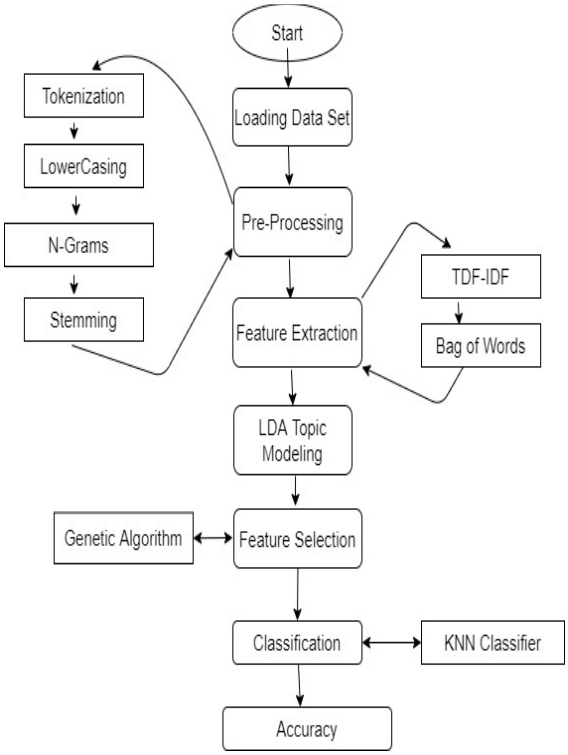

Figure 1. Design of text-based author identification analysis approach

3.2.1.1. Term Document Frequency

The vocabulary size of the Urdu data set is reduced to 906 items when we look at words having ten or more alternatives.

3.2.1.2. Document Term Matrix

Text documents are vectors, with each feature exhibiting a recurring pattern. These vectors may be used to compare two documents and find similarities. From a training dataset, we create a document word matrix based on selected features that the Gensim Dictionary class stores as a lexicon. For recurrent term trimming throughout the entire document term matrix, look at the LDA model.

3.2.2. Bag of Words Extraction

The repetition of text that appears to build a lexicon, which contains character n-grams, word n-grams, and other qualities retrieved from the text, is used to represent the characteristics of a document. We may expand the body length by using all of the terms in the lexicon, which is tough to quantify. We utilized a term-document recurrence technique to find characteristics.

3.3. LDA Model

The topic Modeling Algorithm, often known as latent Dirichlet allocation (LDA) is useful for organizing textual material into document assemblies [24, 25]. LDA is a flexible chance model for collecting unique data in which files are represented as collections of topics with individual changes in files, and each topic

is represented by a list of words that could refer to the subject. This prize is fantastic since it helps a state of overall performance execution in authorship identification with a large number of candidates' writers. The most common practice is to apply the LDA rendition in such a way that it provides us with dimensionality exhaustion near a safe essayist creating style and to occasionally use cosine in the LDA variation subject matter region to select the manageable creator of the textual content document. We used n-grams to represent the designers who were creating the design. Documents were tested as a set of expressions, which meant that every report from tutoring and inspection units was immediately converted into a sparse vector and portrayed in the LDA subject matter area, resulting in vector portrayal that could be examined as mi and ni as result.

An identical measure in the text one of the most well-liked is cosine. It's a lexical matrix that's been enhanced to calculate document vector proximity. We must first reconcile two files m and n to at least one in the L2 norm to find cosine similarity between them. As m shown in equation 1.

$$\sum_{i=1}^{k} m_i^2 = 1 \tag{1}$$

The dot product of two vectors m and n is their cosine similarity are shown in equation 2.

$$\cos(m,n) = \frac{\sum_{i=1}^{k} m_i n_i}{\sqrt{\left(\sum_{i=1}^{k} m_i^2\right)} \sqrt{\left(\sum_{i=1}^{k} n_i^2\right)}} \tag{2}$$

m and n are n-dimensional vectors over the documents set m and n, where i=1,2,3,....k. As seen in Gensim [26], cosine similarity is simple to implement.

3.4. Feature Selection

We employ a genetic algorithm to access information and pick features.

3.4.1. Genetic Algorithm

We apply a genetic algorithm that uses three operators to provide the best solution for information access and extraction of features: selection, cross across, and mutation. In the selection process, the best-fitting parts of the given information are chosen. In a cross-over, all of the best-fitting qualities are set. In the mutation, the values of the best-fitted information are swapped. The model's accuracy is calculated as a fitness of the solution using a genetic technique for feature selection, and the accuracy is calculated using KNN classification.

3.5. Classification

Text documents are represented as vectors, with properties in each document representing frequency phenomena. The similarity between documents is discovered using a vector. To figure out which well-liked articles are similar to the most recent one, the algorithm requires distance measures such as Euclidean distance or cosine, which we utilize in our situation.

3.5.1. K-Nearest Neighbor

The K-nearest neighbor (KNN) algorithm is a non-parametric machine learning technique for assigning an unclassified typical point to a group of recently reported points [23]. Text documents are represented as vectors, with properties in each document representing frequency phenomena. The similarity between documents is discovered using a vector. We apply the KNN classifier to our data collection, it sorts articles into categories based on how close they are to the papers we have discovered.

3.6. Evaluation

To evaluate the model, a confusion matrix is used. Accuracy, precision, recall, and f-measure are described in Equations 3, 4, 5, and 6.

$$Accuracy\ Rate = \frac{Number\ of\ Correct\ Classified\ Tweets}{Total\ Number\ of\ Tweets} \times 100 \tag{3}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{4}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{5}$$

$$F - Measure = \frac{(2 * Precision * Recall)}{Precision + Recall} \quad (6)$$

4. Results

By putting our eight datasets of Urdu tweets to the test, we can certify authorship recognition. The LDA model uses tokenized text files from the n-grams education set without a tag as input for low dimensionality construction, and untagged text documents from the training set for analysis. With an experiment set utilizing a bottom measurement depiction with the development of the LDA K theme, cosine base classifiers produce a bottom measurement depiction of an educational environment based on language and estimate categorization.

4.1. Datasets

We used Tweepy, a python tool, to acquire publically available Twitter tweets, and we also constructed Urdu TD datasets from those tweets. Urdu tweets were utilized in the dataset. For Urdu, the tweeter datasets are a group of eight writers. All of them were chosen at random with the most current 600 tweets per author from March 2020 being chosen. The Urdu TD dataset contains 4800 tweets from eight Twitter users. The total number of words (tokens) in the Urdu collection is 31,960. Author U7 wrote the longest tweets, with 7,630 words, and author U1 wrote the shortest tweets, with 5,684 words. When producing datasets from Urdu TD, we employ a representation of the document. Each data set's documents were separated into 80-20 groups, with 80 percent of each document's data used for training and the remaining 20% for testing.

4.2. Experimental Setup

To test performance and accuracy, all testing was done on an Intel core i5@2.50GHz PC running Windows 10 64 bit with 8 GB of RAM. The system was expanded using Python 3.7 (python software) and the LDA implementation in the Gensim [26] module. We utilized table 2 with eight author tweets to estimate and differentiate LDA for authorship recognition. Table 4 displays the characteristics of a self-choice based on the KNN classifier, as measured by the recall, precision, and F1 measure through accuracy.

Table 2. Urdu_TD datasets used in experimentation.

Dataset	Training Document Words	Distinct Words	Lexicon Size
Author_U1	5,684	431	211
Author_U2	6,780	424	234
Author_U3	6,904	532	240
Author_U4	6,200	442	221
Author_U5	6105	456	223
Author_U6	7475	498	255
Author_U7	7630	501	276
Author_U8	6750	515	233

4.3. Results and Discussion

We evaluate LDA-related authorship recognition on the Urdu language dataset to validate the result, and we create a lexicon with separate LDA features and words by acquiring varied results on the dataset with different puzzles on words. Tables 2 indicate the parameters we used for the Urdu language dataset respectively.

With a range of individuals' k between 12, 24, 36, and 120 and varying lexical proportions, we employed LDA + cosine similarity to measure accuracy in the LDA model. The findings show that the percentages of correctness vary depending on the number of subjects. Within a certain range, accuracy improves before starting to deteriorate. These parameter instructions are incompatible with non-identical lexicon proportions in the same dataset.

On a dataset, we performed the evaluation using the LDA approach [25]. Initially, we used the same collection of themes with similar lexicon sizes, but the results were insufficient in terms of tokens and dataset length, so we couldn't use the same lexicon size in the datasets for the LDA model. The lexicon size is then modified within the LDA model by keeping the same digit k topics. We noticed that raising the

lexicon size enhances data set performance, and because each subject correlates to an author's writing style, we decided on several subjects of 12 to 120, with k=12 being a good choice. Even though the cost of k may be more than 12, each writer may have two separate writing styles. We achieved 84.13 percent accuracy using the LDA version of the dataset and the KNN classifier k=14 with a lexicon size of 255 words and LDA of 60 subjects. As a result, the LDA-based Authorship Attribution Model works on datasets for each of the k topic selections, indicating that the outcome evaluation is correct. The accuracy of Urdu dataset authors after applying the KNN classifier is shown in Table 3.

Table 3. Accuracy of the Urdu_TD datasets.

Dataset	Parameters	Accuracy Rate (%)
Author_U1	Lexicon 211, k=6	83.23%
Author_U2	Lexicon 234, k=16	85.34%
Author_U3	Lexicon 240, k=28	83.15%
Author_U4	Lexicon 221, k=14	86.25%
Author_U5	Lexicon 223, k=6	96.22%
Author_U6	Lexicon 255, k=14	84.13%
Author_U7	Lexicon 276, k=16	98.20%
Author_U8	Lexicon 233, k=24	96.12%

On Urdu TD datasets, the suggested LDA technique is used to assess the diversity of subject k between 6 and 28 based on the number of writers and their works with various lexicon proportions. We adapt the LDA model for different lexicon sizes by keeping certain k topics from the dataset. Because each LDA model cannot be used for the same lexical size. With multiple lexicon values and fixed k values, we adapt the LDA model. We explain why, under the current settings, every dataset with a variety of lexicon sizes and k values ranging from 3 to 70 runs smoothly. By determining the k number, we presume that all content material in the dataset fits with a writer's writing style. Author U1 should use dataset Author U1, Author U4 should use dataset Author U4, and Author U2 should use dataset Author U2. Although the k value might be greater than 6, 14, or 16, this indicates that every author could have two or more writing styles.

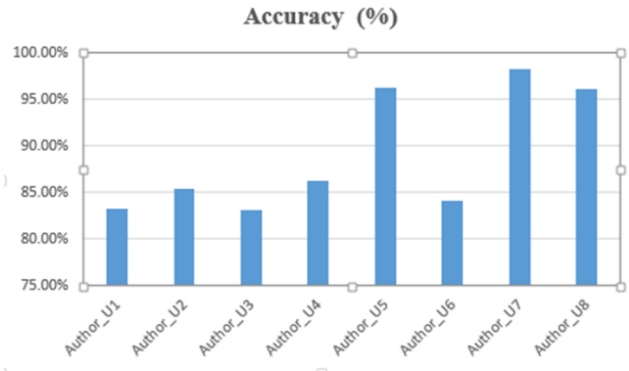

Figure 2. Accuracies of author's text-based tweets

Figure.2 shows the algorithm's accuracy through feature selection using a genetic algorithm and performs classification through the Knn classifier upon author Urdu tweets. On the datasets Author U1 and Author U4, we achieved 83.23 percent and 86.25 percent accuracy using the LDA model with KNN classifiers k=6 and k=14 and lexicon sizes of 211 and 221 words, respectively. On the datasets Author U5 and Author U7, we found an accuracy of 96.22 percent and 98.20 percent, respectively, with lexicons of 223 and 276 words and k values of 6 and 16. These findings demonstrate that our strategy works on a dataset because the LDA version obtains good results even when k subjects are equal to training files by assuming every document shows only one topic. On the Urdu dataset with varying k subjects, the LDA model uses the same technique. Table 4 illustrate the outcome of the authorship attribution proposed model in the form of a percentage.

Table 4. Results of Proposed Model.

Dataset	Accuracy	Precision	Recall	F1
Author_U1	83.23	87	82	85
Author_U2	85.34	81	88	84
Author_U3	83.15	83	99	98
Author_U4	86.25	94	89	86
Author_U5	96.22	94	98	96
Author_U6	84.13	81	88	84
Author_U7	98.20	97	99	98
Author_U8	96.12	95	96	96

On Urdu tweet datasets, our approach produces satisfactory accuracy, precision, recall, and F1-measure results, such as the highest accuracy of 98.20 percent, precision measures ranging from 81 to 97 percent, recall measures ranging from 82 to 99 percent, and F1 measures ranging from 84 to 98 percent. The results of the proposed model evaluation metrics are shown in Figure 3.

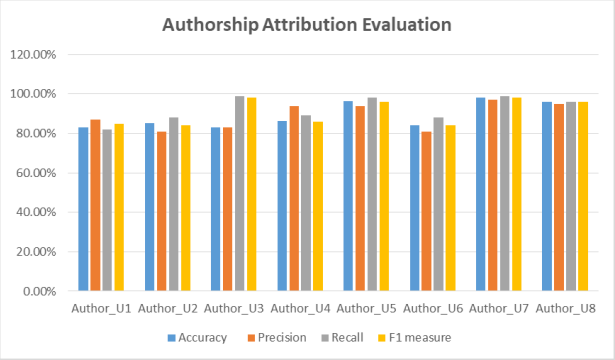

Figure 3. Proposed Model Evaluation Measures

4.4. Comparative Study

The accuracy of our Urdu datasets is seen in table 3. On the Urdu dataset, we achieved an overall accuracy of 98.20 percent. However, it's crucial to note that the findings of a prior study were quite different, with a score of 93.17 percent on Urdu news items [28]. The accuracy of the Urdu dataset in our analysis is higher than in earlier studies. This could be one of the reasons why accuracy varies depending on the size and substance of the datasets. PAN12 was utilized as an Urdu dataset, which had 1800 news articles produced by four different authors. However, in our research, we were able to reach a high level of accuracy using 4800 tweets posted by 8 authors. As a result, we may conclude that increasing the dataset size will improve accuracy. Table 5 compares the results of the proposed model to those of a previous study model.

Table 5. Comparison of Proposed Model with previous Research Model.

Methods	Accuracy	Precision	Recall	F1
KNN [28]	93.17%	93%	93%	93%
Proposed Model	**98.20%**	**97%**	**99%**	**98%**

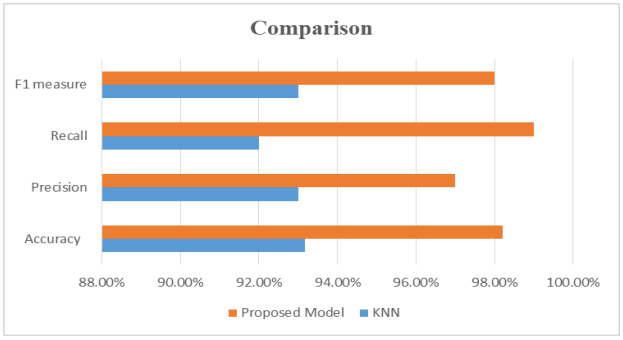

Figure 4. Comparison of Proposed Model with previous Research Models.

5. Conclusion

For the Urdu tweeter datasets, we solved the challenge of authorship recognition. We devised a new approach based on latent Dirichlet allocation (LDA) for this aim. On Urdu datasets, our technique produces satisfactory precision, recall, and F1-measure results, with precision ranging from 81 percent to 97 percent, recall from 82 percent to 99 percent, and F1 from 84 percent to 98 percent. It provides an answer to our initial research inquiry. For Urdu language datasets, we achieved high accuracy. But that was also the most difficult difficulty we faced because each language requires multiple tests at each stage. As a result, having the option of proper configurations is critical. However, by enhancing the quality of tweaking the vocabulary size and k topics in Latent Dirichlet allocation, the accuracy of the findings can be greatly enhanced.

We utilized 600 tweets per author as a threshold in this study, and the findings showed that we were able to reach 98.20 percent accuracy for the Urdu dataset using this threshold. As a result, increasing the dataset size improves accuracy, which satisfies our second study question. Without a language barrier, law enforcement authorities (LEA) may now readily investigate the offender.

6. Future Work

The use of the supervised learning model in future research could be a useful addition to this study, as it would improve accuracy. The classification method still has space for improvement. Additional functions should be introduced to address mistakes that occur in certain conditions. On the other hand, by increasing the dataset and using deep learning classification we can improve the results as well as accuracy.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Robert Layton, Paul Watters, and Richard Dazeley: Authorship Attribution for Twitter in 140 characters or less. (2010).
2. Twitter Usage Statistics. [Online]. Available: https://www.statista.com/topics/737/twitter/#topicHeader__wrapper/
3. Sarah Perez: Twitter's doubling of character count from 140 to 280 had little impact on the length of tweets. (2018).
4. Alomari, Esraa & Manickam, Selvakumar & Gupta, B B & Karuppayah, Shankar & Alfaris, Rafeef. (2013). Botnet-based Distributed Denial of Service (DDoS) Attacks on Web Servers Classification and Arts.
5. Sara El Manar El Bouanani, Ismail Kassou," Authorship Analysis Studies: A Survey" in International Journal of Computer Applications (0975 – 8887) Volume 86 – No 12. (2014).
6. Yunita Sari, Mark Stevenson, and Andreas Vlachos: Topic or Style? Exploring the Most Useful Features for Authorship Attribution.
7. Rong Zheng, Yi Qin, Zan Huang, and Hsinchun Chen: Authorship Analysis in Cybercrime Investigation
8. Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, "On the Feasibility of Internet-Scale Author Identification" in IEEE, 2012.
9. Yunita Sari, Mark Stevenson, Andreas Vlachos: Topic or Style? Exploring the Most Useful Features for Authorship Attribution.
10. B. Brainerd, Statistical analysis of Lexical data using Chi-squared and related distributions Computers and the Humanities, 9, 161–178. (1975).
11. R. Thisted, and B. Efron, Did Shakespeare Write a Newly Discovered Poem? Biometrika, 74, 445–455. (1987).
12. J. M. Farringdon, Analyzing for Authorship A Guide to the Cusum Technique. Cardiff: University of Wales Press. (1996).
13. I. S. Francis, An Exposition of a Statistical Approach to the Federalist Dispute. In J. Leed (Ed.), The Computer and Literary Style (pp. 38–79). Kent, Ohio: Kent State University Press. (1966).
14. D. I. Holmes, The Evolution of Stylometry in Humanities. Literary and Linguistic Computing, 13, 3. (1998).
15. CF. Mosteller, Frederick, and D. L. Wallace Applied Bayesian and Classical Inference: The Case of the Federalist Papers, in the 2nd edition of Inference and Disputed Authorship, The Federalist, Springer-Verlag, (1964).
16. A. McCallum and K. Nigam, A Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 Workshop on "Learning for Text Categorization", (1998).
17. F. J. Tweedie, S. Singh, and D. I. Holmes, Neural Network Applications in Stylometry: The Federalist Papers. Computers and the Humanities, 30(1), 1–10 (1996).
18. D. Lowe, and R. Matthews, Shakespeare vs. Fletcher: A Stylometric Analysis by Radial Basis Functions. Computers and the Humanities, 29, 449–461 (1995).
19. W. Elliot and R. Valenza, Was the Earl of Oxford The True Shakespeare? Notes and Queries, 38:501–506, (1991).
20. D.V. Khmelev and F. J. Tweedir, Using Markov Chains for Identification of Writers, Literary and Linguistic Computing, vol.16, no.4, pp.299–307, (2001).
21. J. Diederich, J. Kindermann, E. Leopold, and G. Pass, Authorship Attribution with Support Vector Machines, Applied Intelligence, (2000).
22. O. de Vel, A. Anderson, M. Corney, and G. Mohay, Mining E-mail Content for Author Identification Forensics, SIGMOD Record, 30(4): 55–64, (2001).
23. KNN Algorithm. [Online]. Available: https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algorithm_finding_nearest_neighbors.htm
24. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, vol. 3, no. 3, pp. 993–1022, 2003.
25. M. Omar, B.-W. On, I. Lee, and G. S. Choi, "LDA Topics: representation and evaluation," Journal of Information Science, vol. 41, no. 5, pp. 1–14, 2015.
26. R. Rehurek and P. Sojka, "Software framework for topic modeling with large corpora," in Proceedings of the Workshop New Challenges for NLP Frameworks, pp. 45–50, Valletta, Malta, May 2010
27. J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating Gender on Twitter," Association for Computational Linguistics, vol. 146, pp. 1301–1309, 2011.
28. Dr. Waheed Anwar, Imran Sarwar Bajwa, and Shabana Ramzan," Design and Implementation of a Machine Learning Based Authorship analysis Model", (2019).
29. N. Tabassum et al., "Semantic analysis of Urdu English tweets empowered by machine learning," Intell. Autom. Soft Comput., vol. 30, no. 1, pp. 175–186, 2021, doi: 10.32604/iasc.2021.018998.