ISSN 2278-3091

Volume 10, No.3, May - June 2021

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse901032021.pdf

https://doi.org/10.30534/ijatcse/2021/911032021

# Analysis of authorship attribution technique on Urdu tweets empowered by machine learning

**Zain Ali[1], Arfan Ali Nagra[2], Zufishan Hameed[3], Muhammad Asif[4]**

[1]Department of Computer Science, Lahore Garrison University, Pakistan, zain.ali90890@gmail.com
[2]Faculty of Computer Science, Lahore Garrison University, Pakistan, arfan137nagra@gmail.com
[3]Department of Computer Science, Lahore Garrison University, Pakistan, zufishan14514@gmail.com
[4]Faculty of Computer Science, Lahore Garrison University, Pakistan, drmuhammadasif@lgu.edu.pk

## ABSTRACT

The process of identifying the author of an anonymous document from a group of unknown documents is called authorship attribution. As the world is trending towards shorter communications, the trend of online criminal activities like phishing and bullying are also increasing. The criminal hides their identity behind the screen name and connects anonymously. Which generates difficulty while tracing criminals during the cybercrime investigation process. This paper evaluates current techniques of authorship attribution at the linguistic level and compares the accuracy rate in terms of English and Urdu context, by using the LDA model with n-gram technique and cosine similarity, used to work on Stylometry features to identify the writing style of a specific author. Two datasets are used Urdu_TD and English_TD based on 180 English and Urdu tweets against each author. The overall accuracy that we achieved from Urdu_TD is 84.52% accuracy and 93.17% accuracy on English_TD. The task is done without using any labels for authorship.

**Key words:** K-Nearest Neighbors (KNN), Latent Dirichlet allocation (LDA), Natural Language Toolkit (NLTK), Term Frequency and Inverse Document Frequency (TF-IDF).

## 1. INTRODUCTION

Now a day technology becomes an important aspect of our daily lives. It totally changes the way humans used to live or communicate, and becomes an important tool of communication. It facilitates shorter forms of communication more easily rather than traditionally longer forms, like essays and handwritten letters [1]. The tweeter is an example of a shorter form of online communication. Which provides visual or text-based service, where users post short tweets (limited number of characters), upload videos, pictures and share links, which is then readable by the other users who are following that person, they can retweet it and can post comments. There are many other social network based websites, like Internet Relay Chat (IRC) rooms, which provide a text-based real-time ‘chatting’ service between internet-connected computers created in 1988[4]. Where short messages posted by the users to the ‘chatroom’ which is then visible by all users in this chatroom. Facebook provides different forms of communication between users such as inbox messaging, wall posts, and comments, focusing on shorter messages. YouTube is also a big entertainment source over the internet which includes comment sections for the viewers to express their views and thoughts in a precise form. It is clear that short messages become a trend on the internet. It also leaves a great impact on other technologies, like SMS (short message services) have become a common use of mobile phones. But with these benefits science has also flawed like cybercrime activities. Through the use of computer networks and network activities, cybercrime is spreading globally. One type of cybercrime is scattering information on the internet which can cause criminals to steal information and spread it online.

Cybercriminals use many online forums to spread illegal content such as online chat rooms, websites, emails, and newsgroup sites [8]. The basic features of these channels are impersonality. People usually do not share information about their identity in online forums. Compared to other crimes, cybercrimes occur through online activities which can cause law enforcement agencies (LEA) to identify and trace theft. These circumstances emerge LEA into more difficult to trace the actual criminal because online forums have many user activities to identify the unique user. LEA needs a solution on an urgent basis which will allow investigators to catch actual cybercriminals.

Rong Zheng 2003 proposed a framework for authorship analysis to automatically trace the criminals through the messages shared on the internet. The framework includes three types of message features including structural features, style markers, and content-specific features are taken out and influencing learning algorithms to produce feature models to identify authors of illegal messages [9]. Experimental study on the data sets Chinese and English online newsgroup messages to evaluate the accuracy of the framework. The result shows that the proposed method can identify the authors of English and Chinese internet messages with good accuracy. This approach helps to identify and trace the cybercrime investigation context [7].

The purpose of this paper is to analyze the Urdu language author and compute the performance of the authorship framework on a linguistic level, based on the provided Urdu tweet data set. This actually helps the law enforcement agencies (LEA) while dealing with the identity tracing problem without any linguistic barrier, during cybercrime investigation. We extracted three types of features to identify the author of the given text which are writing style, content, and hybrid (a mixture of the writing style and the content) [6]. We are particularly interested in the applicability of the proposed technique in a multilingual context.

The rest of the paper is sorted as follows. Segment 2 surveys the current work on authorship analysis and highlights the major types of text features and techniques. Section 3 our research question that we expect to address. Section 4 describes our proposed authorship attribution technique in detail. Section 5 presents the test study that addresses the research questions raised in Section 3, based on several experimental data sets. We conclude the article in Section 6 by summing up, with our research contributions and pointing out the future work.

## 2. LITERATURE REVIEW

Authorship Attribution is the procedure of author recognition from unknown documents. This analysis can be crumbling towards three alternative areas [7]. Author recognition clarifies the isolated author's probability of writing the document in terms of writing style compared to other author writing styles [5]. Author Characterization condenses attributes of an author and produces the profile of an author on the basis of their work. Recognition differentiates several documents and determines a document that this document is written by one author without recognizing the actual author.

Analysis of authorship is the emergence of features that endure persistently for a variety of writing documents that are created by some authors. Features fall down into three parts [7].Firstly Style markers are also called content free features to include recurrence words function, sentence length, number of punctuation, and vocabulary lavishness. Secondly, Structural features include acknowledgment statements, sayonara statements. Thirdly Content-specific features involve prevalence keywords, special nature for special content.

### 2.1 Approaches for Authorship Analysis

Stylometry is a technique that is used to identify the author of the text on the basis of the writing style of long antecede computers [8, 9]. Initially, statistical methods used for authorship analysis. The reason behind this was that different authors have a different writing style which distinguishes different diffusions of words. Time passes and the variety of authors will increase and the quantity of writing files also increases so identification of new documents becomes a problem and considered as a statistical hypothesis test or categorization problem. So at the beginning statistical methods were used for authorship analysis. Brainerd [10] used statistical distribution methods to act in lexical data analysis. Thisted and Efron [11] established an important statistical test. Farringdon [12] put in the CUSUM technique in authorship analysis. Francis [13] concise the statistical approaches for authorship analysis. Holmes [14] constructed that the CUSUM technique was untrustworthy because it cannot forecast the actual author over multiple texts.

The invention of modern computers machine learning strategies used for authorship analysis. Bayesian model supervised by Mosteller and Wallace [15] to examine the paper, McCallum, and Nigam [16] differentiate two Bayesian models for text classification. This version has structural limitations however a variety of mighty techniques applied in writer identification. The maximum illustrative one is neural networks. Tweedie [17] used an excellent artificial neural community also referred to as a multilayer perceptron to assign authorship for papers. The result was compatible with the previous work on the same topic. Lowe and Matthews [18] use another neural network called radial basis function (RBF). They carried out RBF to discover the scope of Shakespeare's work [19] collaboration with contemporary John Fletcher on numerous plays. Newly Khmelev [20] introduces the strategy on the basis of Markov chain for authorship attribution which uses chances of succeeding letters as features. Diederich [21] initiated the aid vector machine to this problem. This approach acknowledges the spotted authors in 60% to 80% of the cases. A new place of look at is the author's finding of digital messages on the idea of message work. De Vel et al. [22] used an aid vector machine as getting to know a set of rules to categorize one hundred fifty email files from three authors. In this experiment, 80% accuracy is achieved.

Comparing the approaches system learning strategies attain big accuracy as opposed to statistical methods. They can stereotype the prepared personal phrase usage with a big set of features. Based on the review all approaches which were initially used and newly used for authorship attribution are mentioned below:

Manual Analysis uses laboring tests and analysis on a given text document to find the identification of author characteristics. Statistical Analysis makes use of statistical techniques for manipulating record numeric values on a given data set to identify the author. Machine Learning makes use of classification techniques to assume the author of the work phase on statistics sets [23].

In the past, huge numbers of workers for authorship in the English language had been concentrating on several techniques for recognition of capabilities from text and the author of a text. According to our observation and research, there is no criteria and work for authorship in the Urdu language up to now. So we decided to work on the Urdu language for authorship recognition and for this plan we use Urdu tweets as our dataset collected from the twitter website.

First, we make our effort to address author recognition in Urdu language and this method is done by LDA using cosine similarity and the KNN classification component is used for more precise results.

This research is done to extend the sector of authorship analysis closer to figuring out how the authorship set of rules plays on a linguistic level, especially in the Urdu context. For this purpose twitter, the Urdu data set is used. Twitter is a microblogging website, and a post on Twitter is called a tweet. As tweets are considered one of the shortest forms of communication, currently in use. And it's getting popular over the internet in recent years, it is reported that it receives over 500 million tweets per day and around 200 billion tweets per year [2]. The reason for its popularity is the limited characters in a post. It’s far a restriction that posts need to be 280 characters or less in the period, the most common length of a tweet is 33 characters. Historically, the simplest 9% of tweets hit Twitter's 140-individual limit [3]. We are motivated to promote the Urdu language as a very little amount of research is done in it. This studies targets to answer the subsequent questions:

- How effectively existing authorship attribution techniques give attribution to the authors of the tweets written in the Urdu language?
- For accurate author profiling, what a number of tweets per writer are needed? Is there any significant effect on accuracy by increasing or decreasing the number of tweets per author[29]?

## 3. METHODOLOGY

In this section, we discuss the frameworks used for authorship identification, steps to proceed, and the corpora. In the dataset, Urdu and English tweets were used. The tweet datasets are a collection of 8 authors 4 for Urdu (for the sake of privacy let's call them author_U1, author_U2, author_U3, and author_U4) and 4 for English tweets (let's call them author_E1, author_E2, author_E3, and author_E4). All of them were selected through some random function, we selected their most recent 180 tweets from March 2020. The complete data set passed through many steps of preprocessing steps like tokenization, lower casing, n-grams, feature extraction, feature selection, document term matrix preparation, topic extraction, and KNN classification.

We created our own English_TD and Urdu_TD dataset from publicly available twitter tweets. Urdu_TD dataset has a total of 720 tweets written by 4 twitter users and 720 for the English dataset. Urdu dataset contains 31,960 words (tokens) in total and the English dataset contains 35,356 tokens in total. Longest tweets were written by the author_E3 of 7,630 words the shortest written by author_U1 of 5,684words.

### 3.1 Datasets

We used two representations of the document while preparing datasets from English_TD and Urdu_TD. Documents from each data set were divided into 80-20, 80% of the data from each document used for training, and the remaining 20 used for testing.

### 3.2 Document Preprocessing

Authorship identification is purely based on the author's writing style. It is also observed in the writing survey that it is not feasible to clean the data set by removing special characters nor to correct the grammatical error and their preference of word suffixes and prefixes, latter capitalization all provide important information about the author. So, removing or correcting such things will reduce the number of features related to the particular author.

### 3.2.1 Tokenization

It divides sentences or a paragraph into little units like words or characters by ignoring white spaces. And store each token with their frequency of occurrence in the dataset with the help of Natural Language Toolkit (NLTK).

### 3.2.2 Lowercasing

All upper case text converted into the lower case before doing any further processing. As in the Urdu language, there is only one case so no need to apply lowercasing.

### 3.2.3 N-Grams

N-gram group neighboring words or characters of length n, for any language. Such as, n-grams feature works at a linguistic level with the same accuracy. It can easily capture the writing style of the language structure of a writer. The accuracy of n-gram depends on the value of n. For the small value of n important differences may fail to capture, but for large n values long n-grams will produce which can cause limitations and we can stick to some specific cases. In a word-level n-gram, a good rule of thumb is to use n-grams where n ∈ {1, 2, 3...5}. Table 1 shows the n-grams types sentence representation

**Table 1:** N-grams (1–5) for the sentence “I’m the author of this paper”

| N-Grams Types | Sentence Representation |
|---|---|
| Word Unigram | I’m, the, author, of, this, paper |
| Word bigram | I’m_the, the_author, author_of, of_this, this_paper |
| Word Trigram | I’m_the_author, the_author_of, author_of_this, of_this_paper |
| Word Fourgram | I’m_the_author_of, the_author_of_this, author_of_this_paper |
| Word Fivegram | I’m_the_author_of_this, the _author_of_this_paper |

Shows a simple sentence “I’m the author of this paper” and the total arrangements of words unigrams, bigrams, trigrams, four grams, and five grams produced from it. For the ease of understanding, underscores (_) are used to supplant spaces.

In the bag of the word model, most of the contextual information is lost, in order to overcome the limitation n-gram used to catch all the more semantically important data from text. In addition, it has been demonstrated to be successful in recognizing the gender of tweeters [27].

### 3.2.4 Remove Stop Words

In English language “a, are, an, is, the, this gets” are the stop words. Which have a high recurrence in the content of language but stop words have negligible lexical information. So, we prefer to expel these words from the dataset before doing any further processing. But on account of the Urdu language, we don’t have such a stop words list. So, we include limitations that the selected word should not appear in every document. We discount all words occurring in documents of more than 70%.

### 3.2.5 Word Stemming

The process of separating the root word from the given words called stem word. Rule-based stemmer was utilized with the assistance of NLP tools to stem datasets words.

### 3.3 Syntax Analyzer & Feature Extraction

It's a process of Extracting numerical information from the documents and that information is called a feature. For training, a model only best-fit features are selected for better results. TF-IDF is used for this purpose.

### 3.3.1 TF-IDF

It is used to calculate term frequency or raw frequency and inverse document frequency (TF-IDF) of each word. It produces distinct feature vectors of the captured information from the texts which ultimately helps in distinguishing the actual author for a given document. It may very well be gotten distinctly by increasing the ratio of the word in a text document to the reciprocal of the ratio utilized in all documents. Figure 1 shows the proposed methodology of this research paper.

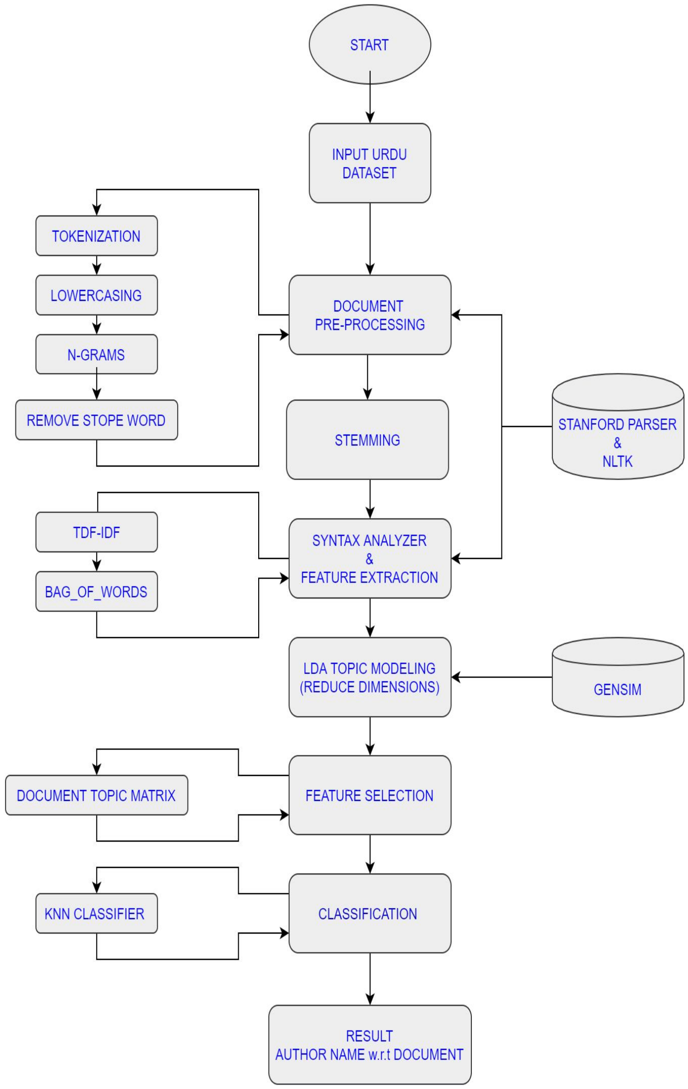

**Figure 1:** Design of text-based author identification analysis approach

### 3.3.2 Bag of Words Extraction

Bag of words is a classic model where the text is considered as a set of words having a repetition of body phenomena. On the other hand, the characteristic of a document is represented as the repetition of text that shows to make a lexicon, and this lexicon comprises character n-grams, word n-grams, and other different features that are taken out from the text. If we utilize all words of lexicon then it can expand the body length which is hard for enumeration. For feature selection, we used a term-document recurrence strategy.

### 3.3.2.1 Term Document Frequency

We examine those words having 10 and more than 10 eventualities and it decreases the Urdu data set lexicon size to 906 terms.

### 3.4 Document Term Matrix

Text documents serve as vectors where each feature shows repetition phenomena. These vectors can be utilized to discover a resemblance between two documents. We build a document term matrix from a training dataset dependent on chosen attributes that were saved in the form of lexicon by the Gensim Dictionary class. LDA model glance for recurrent term trimming in the whole document term matrix.

### 3.5 Feature Selection using LDA

For the information accessing and feature selection, we use topic modeling algorithms. The topic Modeling Algorithm simply like LDA [24] is functional for arranging textual data into assemblies of documents [24, 25]. Table 2 shows the English dataset distinct words and lexicon sizes. Table 3 shows the Urdu dataset distinct words and lexicon sizes.

**Table 2:** ENGLISH_TD datasets used in experiments

| Dataset | Training Documents Words | Distinct Words | Lexicon Size |
|---|---|---|---|
| Author_E 1 | 6,105 | 456 | 223 |
| Author_E 2 | 7,475 | 498 | 255 |
| Author_E 3 | 7,630 | 501 | 276 |
| Author_E 4 | 6,750 | 515 | 233 |

**Table 3:** Urdu_TD datasets used in experiments

| Dataset | Training Documents Words | Distinct Words | Lexicon Size |
|---|---|---|---|
| Author_U1 | 5,684 | 431 | 211 |
| Author_U1 | 6,780 | 424 | 234 |
| Author_U1 | 6,904 | 532 | 240 |
| Author_U1 | 6,200 | 442 | 221 |

LDA is a pliable chance model for amassing distinct information that indicates the files as assemblies of topics with man or woman changes in files and each topic indicates an enumeration of phrases with chances for them to allude to the theme.

### 3.6 LDA + Cosine Similarity

This practice is a prime grant as it profits a country of the art performance in authorship identity with many applicants' authors. The predominant plan is to use the LDA version in such a manner that it gives us dimensionality depletion alongside preserving writer writing fashion and then use cosine in the LDA version topic space to decide the probable writer of the text file. We used n-grams to symbolize the creator's writing fashion. Documents have been proven as a bag of phrases so each file from schooling and take a look at sets modified right into a sparse vector and depicted into LDA topic space to supply vector depiction which can be proven as ui and vi as result [25].

In-text identical measure cosine is one of the most approved ones. It is an extended matrix from calculating lexical to compute closeness between document vectors. In collection to discover cosine similarity between two files u and v, first, we need to conciliate them to at least one in L2 norm:

$$\sum_{i=1}^{k} u_i^2 = 1 \qquad (1)$$

Cosine similarity between two vectors u and v will be dot product of them:

$$\cos(u, v) = \frac{\sum_{i=1}^{k} u_i v_i}{\sqrt{\left(\sum_{i=1}^{k} u_i^2\right)}\ \sqrt{\left(\sum_{i=1}^{k} v_i^2\right)}} \qquad (2)$$

u and v are vectors of n dimension over the documents over the documents set u and v where i=1,2,3,….k. Cosine similarity is easy in implementation complexity as in Gensim[26].

### 3.7 Classification

Text documents are shown as vectors where frequency phenomena are represented by attributes in each document. Vector used to discover the similarity between documents. We apply the KNN set of rules to our data that will discover ways to classify new documents based on their distance to our recognized documents. The algorithm requires distance metrics like Euclidean distance or cosine to determine which regarded articles are near to the brand new one. And in our case we use cosine.

## 4. RESULTS

Through our eight datasets in which four datasets of English tweets and four datasets of Urdu tweets, we authenticate the authorship recognition by applying a test on the dataset. For the construction of low dimensionality, The LDA version collects tokenized text files with n-grams education set without a tag as input and for the analysis, the LDA model gains untagged text documents from the training set. Cosine base classifiers with the manufacturing of the LDA K theme form a bottom measurement portrayal of an education set primarily based on lexicon and estimate the categorization with an experiment set using a bottom measurement depiction. Accuracy Rate (AR) of Authorship recognition is calculated by the following equation:

$$Accuracy\ Rate = \frac{Number\ of\ Correct\ Articles}{Total\ Number\ of\ Test\ Articles} \times 100 \qquad (3)$$

### 4.1 Experimental Setup

All experiments performed in Intel core i5@2.50Ghz working on windows 10 64 bit with 8 GB memory to test the performance and found the accuracy. Using Python 3.7 (python software) and LDA implementation inside the Gensim [26] library used for the increase of the system. To estimate and differentiate LDA for authorship recognition we use table 2 and table 3 datasets having8 documents and 4 authors for the Urdu data set and 4 authors for the English dataset. We used different execution metrics such as recall, precision, and F1 measure through accuracy to illustrate aspects of a self-choice based on KNN classifier on table 6 and table 7 datasets.

### 4.2 Results and Discussion

To verify outcome we appraise LDA related authorship recognition on Urdu and English dataset and we achieve different results on the dataset with different riddles on words to produce lexicon with different features and words of LDA. We set different parameters for Urdu and English datasets as shown in table 4 and table 5.

We set the accuracy by LDA plus cosine similarity in the LDA model, putting the variety of subjects' k between 12, 24, 36... 120 with different lexicon proportions. Results show that by setting different numbers of topics change the accuracy percentages. In defined range accuracy increases and then starts to decrease. In the same dataset with non-identical lexicon proportions, these instructions for parameters are not feasible.

For the appraise, offered LDA technique on four datasets initially we use the same variety of topics with comparable lexicon size, consequences were not adequate in terms of tokens and length in the dataset so for the LDA model we cannot use same lexicon size in the datasets. Then we change the lexicon size differently by keeping the same digit k topics within the LDA model. We reveal that changing the lexicon size performance of data sets increases and each topic matches the writing style of an author so we set the number of topics ranges between 12 and 120 and fixing k=12 is an appropriate option. Although the cost of k can be large than 12 which means any writer may also have two representations of writing style. When making use of the LDA version on the dataset by placing KNN classifier k=7 with lexicon size 255 words and LDA 60 subject then we accomplished 84.13% accuracy. So the accuracy of result appraisal shows that LDA based Authorship attribution model acts on datasets on every k topic selection. Table 4 shows the accuracy of Urdu dataset authors after applying the KNN classifier. Table 5 shows the accuracy of English dataset authors after applying the KNN classifier

**Table 4:** 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% |

**Table 5:** Accuracy of the English_TD datasets.

| **Dataset** | **Parameters** | **Accuracy Rate (%)** |
|---|---|---|
| Author_E1 | Lexicon 223, k=6 | 96.22% |
| Author_E2 | Lexicon 255, k=14 | 84.13% |
| Author_E3 | Lexicon 276, k=16 | 98.20% |
| Author_E4 | Lexicon 233, k=24 | 96.12% |

The proposed LDA approach is applied on English_TD datasets, in order to evaluate the variety of subject's k between 4 and 25 relying upon the number of writers and their writings with several lexicon proportions. Keeping specific k topics of the dataset we adjust the LDA model for different lexicon sizes. As we cannot use each LDA model for the same lexicon size. We adjust the LDA model with several lexicon values with fixed k values. We describe the pleasant execution of every dataset with several lexicon sizes and k value between 3 to 70, within the present conditions we assume that every content material of dataset fit with the writing style of a writer by fixing k value 6 for dataset Author_U1, 14 for Author_U4 and 16 for Author_U2 is a suitable choice. Although k value could be larger than 6, 14, and 16 that will describe that any author may have two or more representations of writing style.

When applying the LDA model on the dataset Author_U1 and Author_U4 by setting KNN classifier k=6 and k=14 with lexicon size 211 and 221 terms then we achieved 83.23% and 86.25% accuracy. On the dataset Author_U2 and Author_U3, we found an accuracy of 85.34% and 93.15% with a lexicon of 234 and 240 terms and k values 16 and 28 severally. These results distinctly show that our method works effectively on a dataset because the LDA version attains satisfactory outcomes while k subjects are equal to training files by assuming every document shows only one topic. The LDA model executes the same method on the Urdu dataset and English dataset with different k topics.

### 4.3 Interpretation of Unclassified Documents

There can be few reasons for unclassified documents. The first reason is that we originate that some authors in their documents indicate paragraphs of different writers and then talk about their opinion on that matter. Due to this reason, they mix their writing fashion with other writers. The second reason is that the author wrote down several domains such as sports, entertainment, politics, and many others which were not domain-specific. The third reason is that the small area of experiment documents may be the reason for unclassified.

## 5. COMPARATIVE STUDY

The analysis of table 4 and table 5 shows our accuracy for both English and Urdu dataset. We attained an overall 98.20% accuracy on the English dataset and 86.25% accuracy on the Urdu dataset. But in a comparative study, it's important to know that, in the previous research the results were quite different, they achieved 84.52% accuracy on English datasets and 93.17% on Urdu news articles [28]. In this study accuracy for the English dataset is better than the previous research. This could be a good reason that accuracy differs because of different datasets, in terms of size and content. They used PAN12 as an English data set, and the Urdu dataset consisted of 4,800 news articles written by 12 different authors. But in this study, we achieved very good accuracy at a very small dataset. Hence, we could state that the accuracy could be increased if we increase the size of our dataset.

## 6. CONCLUSION

In this research, we resolved the problem of authorship recognition for the Urdu and English tweets dataset. For this purpose, we planned a new method usage of latent Dirichlet allocation (LDA). Table 6 shows the execution metrics of Urdu dataset and table 7 shows the execution metrics of English dataset.

**Table 6:** Classification Report of Urdu_TD dataset.

| Authors | Precision | Recall | F1 measure |
|---|---|---|---|
| Author_U1 | 0.87 | 0.82 | 0.85 |
| Author_U2 | 0.81 | 0.88 | 0.84 |
| Author_U3 | 0.83 | 0.99 | 0.98 |
| Author_U4 | 0.94 | 0.89 | 0.86 |
| AVG | 0.863 | 0.895 | 0.8825 |

**Table 7:** Classification Report of English_TD dataset.

| Authors | Precision | Recall | F1 measure |
|---|---|---|---|
| Author_E1 | 0.94 | 0.98 | 0.96 |
| Author_E2 | 0.81 | 0.88 | 0.84 |
| Author_E3 | 0.97 | 0.99 | 0.98 |
| Author_E4 | 0.95 | 0.96 | 0.96 |
| AVG | 0.917 | 0.953 | 0.935 |

Our approach produces satisfactory outcomes of precision, recall, and F1-measure, such as precision measure was vary from 81% to 94%, recall measure was 82% to 99% and F1 measure was 84% to 98% on Urdu datasets and English datasets. It answers our first research question that we discussed in section 3. We achieved good accuracy for both Urdu and English language datasets although the accuracy of the Urdu dataset is low as compared to the English dataset. But that was also the biggest challenge that we have faced because each language needs several testing at every phase. So, the option of appropriate configurations is important. But the accuracy of the results can be improved significantly by improving the quality of tuning the vocabulary size and k topics in Latent Dirichlet allocation (LDA).

In this research, we used 180 tweets per author as a threshold, and from the results, we proved that for this threshold we were able to achieve 84.52 %( for Urdu dataset) and 93.17% (for English dataset) accuracy. So, accuracy can be improved if we increase the size of the dataset, and it answers our second research question. Now law enforcement agencies (LEA) can easily investigate the criminal without any language barrier. For future work, the implementation of the supervised learning model can be a possible improvement in this study, and it will increase accuracy.

## REFERENCES

1. Robert Layton, Paul Watters, and Richard Dazeley: **Authorship Attribution for Twitter in 140 characters or less**. (2010).
2. "#numbers." :**Twitter Usage Statistics**
3. Sarah Perez: **Twitter's doubling of character count from 140 to 280 had little impact on the length of tweets**. (2018).
4. DDoS: **Internet Relay Chat**
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. 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 Identification Model**", (2019).

29 R. Alroobaea, "An empirical combination of machine learning models to enhance author profiling performance," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 2, pp. 2130–2137, 2020, doi: 10.30534/ijatcse/2020/187922020.