Investigating Predictive Features for Authorship Verification of Arabic Tweets

Fatimah Alqahtani[†], **Mischa Dohler**[††]
fatimah.alqahtani@kcl.ac.uk mischa.dohler@kcl.ac.uk
[†] Dept. of Informatics, King's College London, [††] Dept. of Engineering, King's College London

Summary
The goal of this research is to look into different techniques to solve the problem of authorship verification for Arabic short writings. Despite the widespread usage of Twitter among Arabs, short text research has so far focused on authorship verification in languages other than Arabic, such as English, Spanish, and Greek. To the best of the researcher's knowledge, no study has looked into the task of verifying Arabic-language Twitter texts. The impact of Stylometric and TF-IDF features of very brief texts (Arabic Twitter postings) on user verification was explored in this study. In addition, an analytical analysis was done to see how meta-data from Twitter tweets, such as time and source, can help to verify users perform better. This research is significant on the subject of cyber security in Arabic countries.

Keywords:
Authorship verification, Stylometry, Arabic, short texts.

1. Introduction

Authorship analysis is the process of extracting and analysing writing style to identify the authorship. It is an interesting field as it is a mixture of linguistics, psychology, and machine learning science [1]. It is an important field helpful for forensic and digital invaginations. Authorship analysis studies include different categories such as Authorship Attribution, Authorship verification, Author Profiling, Authorship Obfuscation, Profiling Hate Speech, Profiling Fake New, Celebrity Profiling, and many more.

Recent attention has been focused on researchers who work on different areas such as psychology, forensics, and computer science for solving authorship analysis problems [2]. However, authorship verification has been proven to be the most complicated and challenging among the other categories of authorship analysis [2]–[5], especially for short texts.

Authorship verification verifies if an unknown document has been written by a known author or not, which results in a binary value [6]. It is instrumental in the field of forensic authorship analysis because it can determine if two texts were written by the same author. The issue of verifying authorship has been a controversial and much-disputed subject within the field of digital forensics and cyber investigations. The authorship verification task has been addressed by researchers to verify the authorship of e-mail messages and many other forms of data.

For authorship analysis tasks and particularly authorship verification, it is more challenging to solve an authorship problem for a short text than for a long one, because the number of words contributes to better learning for the algorithm and therefore a better analysis. Yet, this field has a limited number of studies – in comparison with other fields. Moreover, although number of research has been carried out on authorship verification in different languages, only a few have focused on the Arabic language. The following subsection will review the studies conducted on authorship verification of Arabic texts.

2. Related work

A dataset of four novel writers was used for authorship verification, consisting of 929,233 words divided into training and testing sets [7]. The idea was based on building an author profile of a specific length (50 to 700) that consists of the studied features. First, author profiles were built based on character bigrams and trigrams, which may belong to initial, medial, or final n-gram characters.

Next, the input text is converted to a profile of the same length. Then, a dissimilarity measure [8] is used to compare the degree of similarity between the training and test author profiles. If the dissimilarity sum is less than the author threshold, the author is matched; otherwise, the input document belongs to another author.

Manuscript received June 5, 2022
Manuscript revised June 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.6.18

Results revealed that larger profiles present better accuracy than smaller profiles. For example, the reported accuracy for the initial bigram was 84.61 for a profile length of 200 and 94.87 for a profile link of 700; however, the accuracies for initial and final trigrams were always 100% for profile lengths (200, 500, 700). Although the reported accuracy was high, it was unclear what number of author documents could impact overall accuracy over a simple set of features (characters).

An extensive set of documents was collected from Dar Al-ifta al Misriyyah, consisting of 3000 balanced datasets and 4,686 documents from unbalanced datasets [9]. The method is based on the frequency-based features of unigrams, bigrams, and trigrams and on style-based features (character, lexical, syntactic, semantic, content-specific, structural, and language-specific). First, the data were filtered, and TFIDF vectors were created. A bootstrap aggregating learner was then used to estimate the classification based on a maximum number of votes technique. Several stylometric and frequency-based features were used, showing that combining the bigram model with style-based features achieved the highest accuracy. However, it was unclear whether the author's documents were used in training or chunking in such lengthy article datasets.

Linguistic features have been shown to increase the accuracy of author verification models. The work of [10] relied on a set of Arabic books and showed that the accuracy of the Manhattan distance function score was highest (with comparative measures) when they pre-processed the text with the stem bigrams. Using 19 works attributed to Al-Ghazali, the Manhattan distance measure of Arabic authorship verification provided an increased accuracy of approximately 87%. However, the dataset was a large set of books by the Islamic scholar Al-Gazali, which means that most of the text is similar because it comes from the same author. Therefore, negative examples were not fully present in their experiment. Hence, with one relative similarity measure (the Manhattan distance), the results could not be generalised.

One of the problems of authorship verification is the lack of negative examples (documents that are not written by the author); therefore, if negative data are diverse in features compared to positive examples, any classifier could be subject to overfitting (with low accuracy on real examples). Therefore, a proper feature selection method could provide acceptable results. To resolve this problem, [11] used a dynamic similarity threshold method for authorship verification based on leave-out feature selection. The objective of the work was to rely only on positive examples to train AV classifiers. Their work was based on 19 books by Al-Ghazali for testing positive results using the leave-one-out method and 12 books from the same genres. After text cleaning and tokenisation, the similarity among documents and with the corpus was calculated using the Manhattan distance function based on the cut-off threshold, θ. This threshold, which is calculated based on the equal error rate (EER), is applied to accept the correct classification of a document. It is assumed that false and true authors are normally distributed, where the rate of false positives equals the rate of false negatives. The author reported that with three to nine percent tokens, an accuracy of 70.97% was possible, contrary to previous works where an increase in features caused an increase in classifier accuracy.

Character and word-level lemmas and part of speech linguistics form part of the authorship verification task [12]. However, the minimum text size that affects the task depends on the feature set and the classification method [13]. The work of [14] conducted two experiments; the first was to find the best feature ensemble, and they used the features of tokens, stems, root, diacritics, and POS tags of n-grams (1 to 4) as features for Arabic author verification. The author used a dataset consisting of 253 documents written by different authors from five domains. The average document sizes for each domain were 802, 820, 1,159, 1,108, and 850 words. The accuracy for each domain varied from 84.53% to 80%. It is important to note that the author found that domains with the smallest sample size achieved the worst results. The second experiment was to find the effect of training or testing sample size. The author found that training dataset size did not correlate with improved accuracy of the authorship verification method. In other words, a training set with a smaller number of documents outperformed one with a larger number of documents.

One hundred twenty-five documents from five common genres in Modern Standard Arabic of opinion columns, economics, fiction, nonfiction, and politics were used for Arabic author verification [15]. The authors evaluated the SVM-calculated distance metrics of the Canberra, Manhattan, Cosine, and Jaccard measures using tokens, stems, and POS tags as features. They found that the Canberra distance measure was the best-performing distance measure in most genres, with an accuracy rate as high as 97.8%. However, the method omits digits,

punctuation marks, and special characters in pre-processing, which limits the applicability of these findings to short texts.

In regard to the corpus of very short texts (social media texts), a comprehensive investigation into existing literature revealed a lack of research in Arabic-language authorship verification. To date, very little literature has confronted the authorship verification problem, and existing studies were conducted on long Arabic texts such as books and poems. Although relatively satisfying results have been reported on online messages such as emails or online articles, but no study tackled the social media texts.

To sum up, the main issue with many approaches is the dataset, which consists of either tiny or large textual documents, possibly in a private dataset. Above all, there is no agreement between authors on the best features that could benefit authorship verification. Notably, many studies have used the authors' own datasets, whereas others have not compared related approaches of other authors on the same dataset. Finally, it should be noted that many publications come from the grey literature due to the unavailability of specialised, high-impact journals for Arabic-language research and the difficulty of finding international reviewers.

Unlike most NLP tasks such as sentiment analysis and text classification, AV problems must not conduct much data pre-processing. Stemming, normalisation, diacritics removal, and other data pre-processing techniques would hide the author's style of writing and therefore raise more challenges. Another important finding is that decreasing the sample size is more challenging in the AV tasks because a larger sample size increases the model's ability to train on data and therefore to verify authorship.

3. Methodology

The methodological approach taken in this study is a data-driven based on a combination of machine learning and different set of features to verify users on Twitter and whether a given user has written a given tweet.

3.1 Data collection

Data plays an important role in the authorship identification tasks [16], hence that applies to all authorship analysis problems. Due to lack of research on Arabic authorship verification on short texts, this work will rely on a new collected dataset of Arabic short texts from Twitter. The dataset contains 100 users with a 268,433 total number of tweets (maximum of 3,000 and minimum of 1,000 tweets on average).

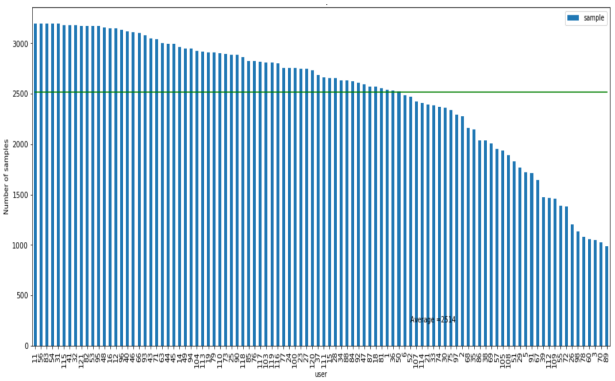

Fig 1: Author's sample count

The users were collected in this study has met a number of requirements to ensure that the collected corpus includes generalisable users tweeting in variety of domains (sport, politics, education, social, etc.). These requirements are as follow:

1. Publicly accessible accounts.
2. Users who tweet in the Arabic language.
3. None of the posted tweets could be identical to any of the other accounts.
4. All users must be genuine and not posted as spam tweets.
5. Personal accounts (not bossiness, marketing, or organisation representatives).

To ensure that the data will be representative, generalised and to avoid any bias, a query was generated to collect 1,000 public accounts based on the previous requirements. Then, the data were revised manually and certain standards were considered to ensure the quality of the data. Experts of any field were excluded as they would mostly tweet using specific words related to their field, so the verification of the authorship would be easy when using (content-based) features. Also, on manual inspection of the nature of the collected tweets, we noticed that some users are writing their tweets in a mixture of the Modern Standard Arabic (MSA) and Colloquial Arabic, so we kept users only with the majority of the tweets being in Colloquial Arabic. Lastly, after filtering and excluding the unsuitable accounts –some has sensitive content-, a random selection of 100

users was made. Tweets were fetched for each user using a data crawler from the Twitter API by entering the username as an input and getting the account data of each user. That included profile information, followers and following, lists, likes and other metadata. However, this work focuses on the tweets' content only, including some metadata which will be explained in the next section.

3.2 Data pre-processing

In order to guarantee user confidentiality, usernames were replaced with a unique ID that could identify the users without exposing their Twitter ID in the results. In addition, Twitter used to allow users to write up to only 140 characters each post, but since September 2017, the number of characters has increased to 280 characters. For that, only tweets published since then with a block size of 280 characters were included in this work to maintain the level of accuracy and to guarantee consistency in user behaviour. This allowed us to avert the problems of text-length dependency in the pre-processing stage. All retweets content were removed, as it do not reflect the user's writing style. Replies content were also removed because it is usually very short, and it could have provided clues about the user based on most-contacted people rather than textual content.

Before doing any NLP task on the text, there are pre-processing steps need to be applied in order to remove any noise in the data such as normalision. In the Arabic language, there are some common pre-processing techniques usually preformed on Arabic texts. These steps include removing diacritics that comes on the words, for example: the word إسْتِناداً à إستنادا, as in most cases removing diacritics doesn't change the word's meaning. Also, normalize the inconsistently typed characters into a constant character such as the letter ة (ta marbotah) à **ه/h** and drop Hamza آ إ أ à ا. Finally, remove the repeated characters صححح à صح and kashida (tadweel) مستحيـــــــــل à مستحيل [17].

However, unlike other NLP tasks such as sentiment analysis or text classification that require the maximum amount of pre-processing. This study is concerned with extracting any information that would lead to the author's identity. So, any extra pre-processing for the texts will strip the content from any distinguishable features of the user

In this work, pre-processing consisted of removing all hashtag symbols (#) and the handles (@). Tweets that contained URLs were also removed, as they were usually accompanied by a general comment written by another user. Line breaks and multiple white spaces were replaced with single space. Stop words, punctuation, and emoji icons were retained. These helped to keep the text to its original shape, which can assist in distinguishing writer style and, therefore, verify authorship. Lastly, the data were scrubbed, and all null and duplicated values were removed.

3.3 Feature extraction

When the data were crawled from Twitter, it contains different information which are: Text ID, Text, Name, Username, Created at, Favourites, Retweet, Language, Client source, Tweet type, URLs, Hashtags, and Mentions. This experiment will only focus on the written content and some meta-data which is timestamp. For that purpose, only the (Username, Text, and Created at) columns were selected. Other unnecessary columns were removed, such as whether the tweet was favourited/mentioned, etc.

As we have two different type of information (Text, timestamp), we will explain the feature extraction for each coloum separately.

- Textual features

In long documents such as books and novels are well structured data, and because of their large size, they have many linguistic features. On the contrary, online content (especially on Twitter) usually consists of a few lines written quickly and spontaneously, and it often contains syntactic and grammatical errors such as spelling mistakes and characteristics such abbreviation use. In addition, online social media content is unstructured data; it is usually not written in paragraphs and usually does not contain a greeting or signature. This format reduces the number of features and may also lead to difficulties in feature extraction.

There are different textual features to be extracted from texts which can be feature specific such as bag of words, n-gram, etc or content-free (style-based) such as the stylometry features. The stylometric features refer to "the study of linguistic style, typically described by features such as sentence length, word choice, word count, and syntactic structure" [18]. These features has proven in many studies the effectiveness of verifying the authorship.

For that, they will used in this experiment as the baseline features. The literature has provided many stylometric features that can be applicable on different type

of texts [19]. However, it is important to mention that stylometric features vary from one language to another, the features in the previous table were originally designed for the English language, they may not be applicable to or relevant for other languages. That irrelevance was proved in [20], where authors who conducted experiments in both English and Chinese found that some English stylometric features could not be applied to Chinese, such as word boundaries and other language rules.

The same problems arise when applying these rules to the Arabic language. For example, some feature created by [19] were based on counting upper-case and lower-case characters whereas Arabic language does not have that. So not all the features were applicable on our dataset. Moreover, not all the features are applicable on social media due to the variation of text length, text structure, and the writing habits.

Considering our dataset, a number of stylometric features were selected that could be applied to the Arabic language and were suitable for use in a social media context. Most of the structural features cannot be applied to tweet content due to the short length, so only lexical and syntactic features were chosen. To make the feature more compatible with our dataset, we have applied another set of features that are conducted on Arabic language presented by [21] which include the special features of Arabic language such as diacritics, Arabic punctuations, and Arabic function words. The following table presents our features which were a combination from the studies of [19] and [21] to have the best and most accurate features of our dataset.

Table 1. The extracted Stylometric features

Feature	Description
Lexical (character)	Number of characters per text Number of white-space characters Number of special characters (@, &, etc) Number of emoticons – Unicode Number of diacritics (ó,ǒ,ò,ǒ,ọ,ọ,ǒ,ò)
Lexical (word)	Number of words per text Average word length (in characters) Number of long words (more than 6 characters) Number of short words (less than 4 characters)
Syntactic	Number of Arabic punctuations (from right to left) Number of Arabic function words

- Non-textual features (Timestamp)

Moving to the meta-data with the column (Created at), this column carries 8 important information which are: week day, month, day, hour, minute, second, time-zone, and the year. These information can tell us more about the user's personality and habits. For example, each user has some free days more than other day when he can use the social media and write posts. Moreover, most people has a specific times to work, relax, or have fun. So, we can have more information about the user based on the time/part of the day that he/she write online posts. So, even if the user tries to hide the style of writing, there are some other unconscious habits and practices which would help to identify the users on social media platforms.

In the beginning, as we selected the features of timestamp which comes in the coloumn (Created at), we had to conduct some feature engineering on the data. Firstly, for the data of the date/time that were collected from Twitter came in this format Sat Jan 19 08:24:30 +0000 2019. We splited these data to be able to use them into different column Day of the week, month, day, hour, minute, second, time zone, and year. Firstly, in order to make all values integer values, we encoded the day of the week and the month into numerical values each with a specific value.

Then, after a careful consideration we found some information that wouldn't be very distinguishable to the model which are (day, minute, second, month, and year). These information are variable and we can't consider it to be a constant practice. For example, (day and year) of posting might be a simple coincidence that can't be considered as a user behaviour. For the (minute and second), these are very precise information that the user spontaneously would post in that time. So we wouldn't consider it as a user pattern to avoid adding a misleading data.

However, other data is important and might give more information about the user. For example, (Hour) where the social media users tend to check their social media account and post in their convenient time. This time varies from one user to another and therefore will give a pattern for each user. Yet, the exact hour might be too specific, for that we will conduct one experiment on the exact hour and the other experiment will consider the quarter of the day.

We divided the day into four quarters which are as follow:

- Morning: From 5am – 11am, where people usually get up for school or work.
- Afternoon: From 12pm – 5pm, which is considered to be a busy time for most people.
- Evening: From 6pm – 11pm
- Late night: From 12am – 4 am

We suppose that categorising the hours of the day will allow to make a better pattern than the specific hour, as it will give the model range of values.

The other meta-data that will be included is the day of the week, we selected this feature because through our observation of the social media users we found that some users tend to use social media only on the weekend more frequently than on the weekdays.

4. Experimental setup

After the feature extraction, each feature was extracted and represented as a vector, and then it was ready to enter the model.

4.1 Data splitting

A series of experiments was carried out with various train/test ratios in order to assess how much data the classifier would need for training to give a reliable yet not biased results.

The first set of experiments conducted by each classifier divided the data 50/50 for training and testing, and it gave low performance due to the small number of training data. The percentage of training was then increased to 60% with better results. Finally, the train/test ratio 70/30 was used because the experiments of 80/20 and 90/10 only made increase in the results. However, for more realistic results, the data were splitted into 70/30 for training and testing, respectively. This is because training the model with a bigger part of the data might lead to overfitting. Therefore, 70% would be efficient enough to train the model.

However, that was conducted on the first experiment only (stylometric features). Then, another kind of data splitting was applied to measure the model performance is k-folds validation (5 folds). Where the data will be divided into 5 sets and one of them is the test set, this will be repeated 5 times where the test set is changed in a reciprocal way. That was applied on all experiments to ensure accurate results and to avoid any overfitting.

4.2 Classification algorithms

Many text categorisation methods have been proposed in the previous studies using machine learning and deep learning techniques. These classifiers vary in the adopted approach: decision trees, naive-Bayes, support vector machines, nearest neighbours, and neural networks [22].

Although these algorithms perform well on general text classification tasks, only a few classifiers proposed in the literature provided satisfactory accuracy in verifying authorship. This study was conducted using the following classifiers: GB, RF, SVM, and KNN, which are some of the most well-known tools for verifying authorship.

This study was conducted using the following classifiers: Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM) and the K-nearest neighbour (kNN), which are some of the most well-known tools for verifying authorship.

4.3 Evaluation Metrics

To measure the algorithm efficiency and find to extent the predicted values matched the true values, a number of metrics were used. There are many available classification metrics that can be used to evaluate the models. However, this study will use precision, recall, accuracy, and F1 score, as they are the most commonly used metrics for authorship verification studies [23] add more ref.

Accuracy determines the ratio of the total number of true samples predicted to the total number of samples. In this study, accuracy was defined by the ratio of the number of correctly recognised tweets—which are the number of tweets classified to the real user versus other users'—to the total number of tweets.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is a measure of performance calculated as a result of true and false positives, which leads to the correct identification of tweets by real users. A higher precision is a greater rate of user verification. Likewise, a low precision value means there are many false positive values. In our case, tweets that did not belong to the claimed user were incorrectly stated as belonging to that user.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall is the result of true and false denial that determines the correct recognition of tweets by strangers. High recall indicates a higher rate of correct identification of tweets by specific users, and lower recall indicates a greater existence of false negative values. Lower recall means that tweets belonging to a specific user will be incorrectly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Where TP is = True positive

TN = True negative

FP = True negative

FN = False negative

Finally, the F1 score is a combination of precision and recall matrices; it combines false negative and false positive values. This score is useful for this type of work because of the many actual negative values (tweets that do not belong to the claimed author). It is calculated as follows:

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

It is important to note that these setting (evaluation metrics, and the used classifiers) will be fixed on all the coming experiment. The difference however will be in the used features selection and data splitting for each experiment.

5. Experimental results

As explained earlier, we will conducted different features to find their effect on the verification task for the Arabic texts. Firstly, we will conduct a baseline experiment that use the stylometric features only. Then, we will use the TF-IDF features and explore which is more distinguishable for the user's writing. Lastly, we will investigate if the timestamp features can give hint about the user behaviour and therefore verifying the user from the tweet.

4.1 Experiment 1. Using linguistic features

4.1.1 Stylometric features

This task was investigated using four different classifiers that are expected to give different results because each classifier performs differently. The results that were obtained from the preliminary experiment of verifying authorship on Arabic short texts are summarised in Table

Table 2. Results of the baseline experiment using stylometric features

Classifier	Avg F1	Avg recall	Avg precision	Avg accuracy
GB	**0.75**	**0.76**	**0.75**	**0.75**
RF	0.74	0.75	0.73	0.73
SVM	0.70	0.72	0.71	0.70
KNN	0.67	0.68	0.66	0.66

The results, as shown in the above table, indicate that GB verified the authorship with an average accuracy of 75%. In addition, the RF classifier achieved an average accuracy of 73%, while the accuracy of verifying authorship using the SVM and kNN classifiers achieved 70% and 66%, respectively.

The results of this study indicate that the users with the best/worst accuracy vary for each classifier. This disparity may be ascribed to the fact that each classifier performs differently. For example, the RF classifier approach works in a tree-based method while SVM sets the optimal hyperplane between the maximal margin of two classes. Moreover, the maximum/minimum accuracy of the users' results varies with each classifier (algorithm). A closer inspection of the figure below shows that the GB classifier reached almost 89% accuracy for some users. On the other hand, the kNN's maximum value was around 77% for the best verified users.

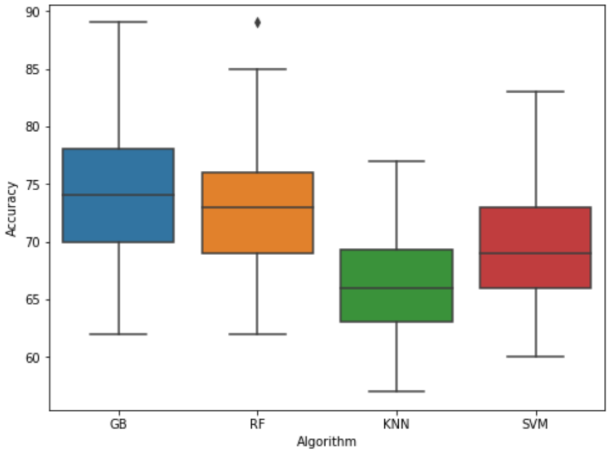

Fig 2. Comparison of algorithms performance

By comparing the results, it can be seen that GB outperformed the other classifiers. In general, it can be said that, when using the ensemble methods (GB and RF), better results are achievable. The ensemble methods allow to consider one sample of the decision tree rather than relying on one entire decision tree. After that, calculating the features that should be used each time and subsequently creating a final predictor based on the combined results of the decision trees' samples.

Although these findings will undoubtedly be scrutinised, there are some immediately dependable conclusions. Overall, these results indicate that this experiment was successful because it was able to verify users on Twitter. In order to ensure that we build a generalisable model that can perform well on the unseen data (test data). We need to evaluate our model performance by applying the Cross Validation strategy.

There are different types of cross validation, however, we will use k-folds with 5 folds. The concept of k-fold is that instead of dividing the data into training and testing, the model will divide the whole data into a number of folds (k folds). After that, the model will train on k-1 folds and test on one-fold only. This process will be repeated with shifting in the training data. Finally, it calculates the average performance of all folds' performance.

Table 2. Results of the stylometric features and applying Cross Validation (5-folds)

	F1	Recall	Precision	Accuracy
Macro avg	0.74	0.75	0.75	0.75

4.1.2 TF-IDF features

In this section, we will use the TFIDF features instead of the stylometic features. In order to compare the performance of the content-specific features like TFIDF and the content-free features like stylometric.

TF-IDF stands for (Term Frequency - Inverse Document Frequency), it is an important technique that is based on the frequency of the words in the same text. It is basically an improvement of BOW technique but with more professional way and by avoiding the flaws that are in the BOW. Where BOW relies only on the presence of the words in the text, regardless of number of repetitions. The TF-IDF is concerned with the frequency of words, it results from multiplying of the two values TF and IDF. Add ref

TF is concerned with the duplication of a word in the text, and it is directly proportional to the word's relation strength with the text. So, if a word repeat more than once that makes it has a strong effect and vice versa. TF is measured by applying the following formula:

$$(1 + \log tf_{t,d})$$

$$\text{Where } tf = \frac{\text{Total number of term existence in documents}}{\text{Total number of all words in the documents}}$$

That was regarding the first part of the term TF-IDF which explain the (Term Frequency), but the concept of (Inverse Document Frequency) is the opposite. Where, the commonness is the word in the text inversely proportional with the strength of the word. If a word is commonly used in the text and among other texts, that weaken the word's effect.

For example, the words (yes, so, very, indeed, etc.) are common used and carry a general meaning which does not make them effect on the text's meaning. That means, the most frequent words in text might have a big effect, but that is in the case where these words are rarely used. Taken

together, the most common words have less weight where the rarely used words have a higher weight.

IDF is calculated through the following formula:

$$Idf_t = \log_{10}(N / df_t)$$

Where *N is* the total number of documents

and df is the number of documents containing the word t.

In this experiment, after processing the text we have entered the pre-processed text to the TFIDF tokenizer.

Table.3: Results of using TF-IDF features only

	F1	Recall	Precision	Accuracy
Macro avg	0.67	0.73	0.64	0.69

As presented in the baseline experiment, Gradient Boosting classifier performed better than the other classifiers at all levels. In addition, GB found to be the best classifier dealing with the kind of data we have. So, we will be using it in this experiment.

Table.4: Comparison of using stylometric features, TF-DF, and combination of both

Features	F1	Recall	Precision	Accuracy
Stylometric features	0.74	0.75	0.75	0.75
TF-IDF features	0.67	0.73	0.64	0.69
Stylometric and TFIDF	0.77	0.75	0.79	0.77

As presented in Table 4, there were a big drop in the results. Comparing the stylometric features with the TF-IDF features, we find that the former exceeds the later until 11%. Which is considered to be a big difference between the two methods. Prove the argument of why the drop in the discussing section (A possible explanation for these results may be)

However, it is TFIDF alone is not enough to verify the author. For that, we have added the Stylometric features to capture the writing style alongside with the TFIDF features. It can be seen that a combination of TF-IDF and stylometric features gave the best possible results.

4.2 Combination of textual and non-textual features

Despite the fact that online social media posts (such as those on Facebook and Twitter) contain a variety of data, including user profiles, content (text), timestamps, location tags, and post responses, studies on authorship identification always have focused solely on the textual content. Other meta-data containing the user's unconscious behaviors, on the other hand, could be useful in enhancing any flaws in the results. The timestamp feature is a behavioural manifestation of underlying circadian cycles of a variety of physical processes that could be used to verify authorship. By choosing a group of users and observing their behavior over time. The distribution of messages for a particular user is thought to remain consistent throughout the day. As a result, non-textual elements should be considered while analyzing data content.

Ttimestamps of tweets can provide valuable information because it provides the time and date of user activity. Therefore, it reflects the temporal patterns of users (their habits and characteristics), which are believed to clarify user behaviour. What is not yet clear in most authorship identification studies is the impact of non-textual features on resolving authorship problems and this feature has not been applied for authorship verification yet.

Further research on other features (non-textual) is, therefore, an interesting next step to test the effect of user behaviour and subconscious activities on verifying authorship. Based on suggestion of a previous study [24], a combination of non-textual features (latitude and longitude of posts, timestamp,..) could increase the model accuracy of verification to results greatly. Although many studies have tried to verify authors using textual features, these have limited applicability on Twitter's short texts. Therefore, enriching textual features with other non-textual features could enhance the authorship verification models.

With the timestamp, it has proven possible to identify users with multiple aliases of the same users on discussion forums where the posting times can be related to the user identity, which gives more accurate results in detection of the users with multiple aliases [25], [26]. It

would be worthwhile to use non-textual as well as textual features to clarify the impact of these features on authorship tasks. For that, further experiments using the timestamp features could shed more light in verifying the users on Twitter.

After selecting the specific time data which are (day of the week, quarter of the day) we added these features to the data used in the baseline experiment. This data contains the user, text, and the stylometric features that were extracted from the texts. We combined these data, and feed them to the model using the GB classifier and the exact same setting of the baseline experiment to ensure finding the actual effect of the features.

4.2.1 Timestamp features

We conducted different experiments and found that the features (day of the week) didn't have any effect on the performance. That due to the number of unique values of days, we found that the number of tweets for each day of the week are very similar and therefore they were not distinguishable and didn't make any effect. Note that this happened by chance in our dataset, and it can't be generalised on other data. So it worth more investigation in other datasets.

```
for i in ['Weekday'] :
  print(data[i].value_counts())
```

```
Tue    55617
Sat    54932
Mon    54899
Sun    54870
Wed    53686
Thu    51235
Fri    50188
Name: Weekday, dtype: int64
```

Fig 3: Number of tweets for each day of the week in the dataset

It is important to note that the day of week feature made minor effect on the user-level but it gave the same average results. The following table presents the results of using the time feature (quarter of the day).

Table 5: Results of combing time features with the stylometric features

Feature	F1	Recall	Precision	Accuracy
Stylometric	0.74	0.75	0.75	0.75
Stylometric + time	0.74	0.74	0.75	0.74

As shown in the table above, the time feature has not improved the performance of the model. That demonstrates that there is no explicit correlation between the social media users and the time/day of posting –at least in our dataset-.

Discussion

As stated earlier, there is a dearth of studies in the literature that are focused on solving the authorship verification problem in Arabic short text. Although they do not seem to be identical to the present study, some work on the verification of the Arabic language will be discussed and compared to this study.

A previous work [27] that aimed to prove the efficiency of using the Arabic function word to verify the author achieved good results. The experiment of [27] and [11] were carried out to verify the authorship of books with large content, which would certainly help the model undergo more training and, thus, achieve more accurate results. This study also used Arabic function words to train the models. Although the list of words is in classical Arabic whereas the content of Twitter is mostly in Modern Standard Arabic, it seems that they were helpful in verifying the authorship.

Meanwhile, the work of [16] have used the stylometry features, and a total of 22 features were extracted. The authors conducted a study of 12 users on Twitter with 3200 tweets per user. Those users were picked while collecting data on specific domains (religious, media, academia, politics, sports and music). Categorising the users certainly assisted the model, as the classification would then become more topic-based than feature-based. They achieved an accuracy of 68.67%, however, the experiments were focused on authorship attribution

(identification), which is considered to be less challenging than authorship verification.

This study extracted different Arabic stylometric features, in addition to using the list of Arabic function words. The users in this study were not from specific domains, but rather, general accounts, so the models verified the authorship based on the linguistic and stylistic features of the author. Moreover, unlike the majority of the work done in this field which used long-text content, this work used the data of Twitter, which has probably the shortest text among all social media platforms.

7. Conclusion

The findings of the experiments confirmed that machine learning algorithms are applicable in verifying authorship to short texts of Arabic language. In addition, using stylometric features in Arabic language assists greatly in verifying authorship.

In previous studies and this study. stylometric features are considered to be the basic and main feature for the authorship analysis tasks as they reflect the style of writing which these tasks stand on, and for that they are kept in all the experiments. The study was conducted on 100 users writing in Arabic language with a maximum of 3000 tweets. In addition, different textual and non-textual features were extracted. A number of experiment has been conducted to identify the best possible features/combination of features and best possible model with using four classifiers applied (GB, RF, SVM, and kNN).

The results showed that Twitter's meta-data don't have much effect on verifying the user as much as the linguistic features. The best performance achieved in this study were a combination of Stylometric and TFIDF features using the GB classifiers with 0.77 accuracy. Which considered to be a satisfactory initial results for researches in the field of verifying short Arabic texts.

Further experimental investigations are needed to find the effect of other linguistic features such as BOW, word embedding, and the state-of-art transformers in verifying the social media texts in general and in the Arabic language in specific.

References

[1] N. Roy, "Authorship Analysis as a Text Classification or Clustering Problem," 2019. [Online]. Available: https://towardsdatascience.com/authorship-analysis-as-a-text-classification-or-clustering-problem-312549d4a4c0.

[2] O. Halvani, C. Winter, and A. Pflug, "Authorship verification for different languages, genres and topics," *DFRWS 2016 EU - Proc. 3rd Annu. DFRWS Eur.*, vol. 16, pp. S33–S43, 2016.

[3] H. Azarbonyad, "Time-Aware Authorship Attribution for Short Text Streams," *ACM*, pp. 727–730, 2015.

[4] A. A. E. Ahmed, I. Traore, P. O. B. Stn, C. S. C. Victoria, and B. C. V. W. Canada, "Detecting Computer Intrusions Using Behavioral Biometrics," *PST*, 2005.

[5] N. Potha and E. Stamatatos, "An improved impostors method for authorship verification," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10456 LNCS, pp. 138–144, 2017.

[6] A. Altamimi, N. Clarke, and S. Furnell, "Multi-Platform Authorship Verification," *Proc. Third Cent. Eur. Cybersecurity Conf. ACM*, p. 13, 2019.

[7] S. Kumar, "Assessment on Stylometry for Multilingual Manuscript," *IOSR J. Eng.*, vol. 02, no. 09, pp. 01–06, 2012.

[8] V. Kešelj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution," in *Proceedings of the conference pacific association for computational linguistics, PACLING*, 2003, vol. 3, pp. 255–264.

[9] M. Al-Sarem, A. H. Emara, W. Cherif, M. Kissi, and A. A. Wahab, "Combination of stylo-based features and frequency-based features for identifying the author of short Arabic text," in *ACM International Conference Proceeding Series*, 2018.

[10] H. Ahmed, "The Role of Linguistic Feature Categories in Authorship Verification," *Procedia Comput. Sci.*, vol. 142, pp. 214–221, 2018.

[11] H. Ahmed, "Dynamic Similarity Threshold in Authorship Verification: Evidence from Classical Arabic," *Procedia Comput. Sci.*, vol. 117, no. 0, pp. 145–152, 2017.

[12] D. C. Castro, Y. A. Arcia, M. P. Brioso, and R. M. Guillena, "Authorship verification, average similarity analysis," *Int. Conf. Recent Adv. Nat. Lang. Process. RANLP*, vol. 2015-Janua, pp. 84–90, 2015.

[13] S. Ouamour, S. Khennouf, S. Bourib, H. Hadjadj, and H. Sayoud, "Effect of the text size on stylometry—application on Arabic religious texts," *Adv. Intell. Syst. Comput.*, vol. 453, pp. 215–228, 2016.

[14] H. Ahmed, "Sample Size in Arabic Authorship Verification," pp. 1–8.

[15] H. Ahmed, "Distance-Based Authorship Verification Across Modern Standard Arabic Genres."

[16] A. Rabab'Ah, M. Al-Ayyoub, Y. Jararweh, and M. Aldwairi, "Authorship attribution of Arabic tweets," *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA*, pp. 1–6, 2017.

[17] O. Obeid *et al.*, "CAMeL tools: An open source python toolkit for arabic natural language processing," *Lr. 2020*

- *12th Int. Conf. Lang. Resour. Eval. Conf. Proc.*, pp. 7022–7032, 2020.
[18] J. S. Li, L. Chen, P. Singh, and C. C. Tappert, "SPECIAL ISSUE PAPER A comparison of classi fi ers and features for authorship authentication of social networking messages," no. August 2016, pp. 1–15, 2017.
[19] R. Zheng, Y. Qin, Z. Huang, and H. Chen, "Authorship analysis in cybercrime investigation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2665, pp. 59–73, 2003.
[20] R. Zheng, J. Li, H. Chen, and Z. Huang, "A Framework for Authorship Identification of Online Messages : Writing-Style Features and," vol. 57, no. 3, pp. 378–393, 2006.
[21] M. Al-Ayyoub, A. Alwajeeh, and I. Hmeidi, "An extensive study of authorship authentication of Arabic articles," *Int. J. Web Inf. Syst.*, vol. 13, no. 1, pp. 85–104, 2017.
[22] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text classification using machine learning techniques," *WSEAS Trans. Comput.*, vol. 4, no. 8, pp. 966–974, 2005.
[23] O. Halvani, L. Graner, R. Regev, and P. Marquardt, "An Improved Topic Masking Technique for Authorship Analysis," pp. 1–20, 2020.
[24] R. Kaur, S. Singh, and H. Kumar, "AuthCom: Authorship verification and compromised account detection in online social networks using AHP-TOPSIS embedded profiling based technique," *Expert Syst. Appl.*, vol. 113, pp. 397–414, 2018.
[25] F. Johansson, L. Kaati, and A. Shrestha, "Timeprints for identifying social media users with multiple aliases," *Secur. Inform.*, vol. 4, no. 1, p. 7, 2015.
[26] I. Mishra, S. Dongre, Y. Kanwar, and J. Prakash, "Detecting Users with Multiple Aliases on Twitter," in *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2018, pp. 560–563.
[27] K. S. Hussein, "Authorship verification in Arabic using function words: A controversial case study of imam Ali's book peak of eloquence," *Int. J. Humanit. Arts Comput.*, vol. 13, no. 1–2, pp. 223–248, 2019.

Fatimah Alqahtani received a master's degree in information systems from the College of Computer and Information Sciences, Portsmouth University, UK, in 2018. She is currently pursuing the Ph.D degree with the Faculty of Natural, Mathematical & Engineering Sciences, Department of informatics, King's College London, U.K. Her research interests include Natural Language Processing, Cybersecurity, and Machine Learning algorithms, with a special focus on verifying authorship of social media users.

Mischa Dohler (IEEE S99M03SM07F14) was full Professor in Wireless Com- munications at King's College London, driving cross-disciplinary research and innovation in technology, sciences, and arts. He is a Fellow of the IEEE, the Royal Academy of Engineering, the Royal Society of Arts (RSA), the Institution of Engineering and Technology (IET), and a Distinguished Member of Harvard Square Leaders Excellence. He is a serial entrepreneur, composer, and pianist with five albums on Spotify/iTunes, and is fluent in six languages. He acts as policy advisor on issues related to digital, skills, and education. He has had ample coverage by national and international press and media. He is a frequent keynote, panel, and tutorial speaker, and has received numerous awards. He has pioneered several research fields, contributed to numerous wireless broadband, IoT/M2M and cyber security standards, holds a dozen patents, organized and chaired numerous conferences, was the editor-in-chief of two journals, has more than 300 highly-cited publications, and authored several books. He was the Director of the Centre for Telecommunications Research at Kings from 2014-2018. He is the cofounder of the Smart Cities pioneering company Worldsensing, where he was the CTO from 2008-2014. He also worked as a Senior Researcher at Orange/France Telecom from 2005-2008.