Towards a Stylometric Authorship Recognition Model for the Social Media Texts in Arabic

Haroon Nasser Alsager

Department of English,
College of Science and Humanities
Prince Sattam Bin Abdulaziz University
Alkharj, Saudi Arabia

Abstract

Numerous studies have been concerned with developing new authorship recognition systems to address the increasing rates of cybercrimes associated with the anonymous nature of social media platforms, which still offer the opportunity for the users not to reveal their true identities. Nevertheless, it is still challenging to identify the real authors of social media's offensive and inappropriate content. These contents are usually very short; therefore, it is challenging for stylometric authorship systems to assign controversial texts to their real authors based on the salient and distinctive linguistic features and patterns within these contents. This research introduces a new stylometric authorship system that considers both the shortness of data and the peculiar linguistic properties of Arabic. A corpus of 20, 357 tweets from 134 Twitter users. A document clustering based on Document Index Graph (DIG) model was used to classify input patterns in the tweets that shared common linguistic features. A comparative analysis using Vector Space Clustering (VSC) model based on the Bag of Words (BOW) model, conventionally used in authorship recognition applications, was used. Results indicate that the proposed system is more accurate than other standard authorship systems mainly based on vector space clustering methods. It was also clear that the model had the advantage of providing complete information about the documents and the degree of overlap between every pair of documents, which was useful in determining the similarity between documents.

Keywords: Authorship recognition, cybercrime, document clustering, Document Index Graph, linguistic stylometry

Cite as: Alsager, H.N. (2020). Towards a Stylometric Authorship Recognition Model for the Social Media Texts in Arabic. *Arab World English Journal*, 11 (4) 490- 507.

DOI: <https://dx.doi.org/10.24093/awej/vol11no4.31>

Introduction

Despite the influential role of social media in enhancing communication and sharing knowledge and information among individuals all over the world due to its increasing popularity, the problem of anonymity and fake accounts still represents a challenge to individuals and societies (Agarwal, Dokoochaki, & Tokdemir 2019; Boyd & Crawford, 2012; Budinger & Budinger, 2006; Schallbruch & Skierka, 2018). Due to the anonymous nature of social media applications such as Facebook, Twitter, Skype, and Tumblr, millions of users can create fake accounts that can be used for unethical practices and illegal actions. Such applications can post negative comments about each other or send abusive messages and inappropriate content to other users. Such practices have, in many cases, negative impacts on the life of individuals. including suicidal thoughts and attempts, stress, defaming, and distorting the reputation of families and businesses (Golbeck, 2018; Görzig & Frumkin, 2013; Kowalski, Limber, & Agatston, 2012; Lowry, Zhang, Wang, & Siponen, 2016).

Recent studies and reports have revealed that most anonymous posts, comments, and messages on different social media platforms, especially Twitter, are not committed to ethical principles and etiquettes (Anderson, 2018; Reader, 2012). These usually have offensive and inappropriate content that violates the sanctity and freedom of others. Many users also use fake accounts to spread fake news, hate speech, and terrorist propaganda. Different social media sites such as Facebook and Twitter have indeed been accused of having a determining role in allowing hate speech and inciting violence against ethnic minorities in different parts of the world (Citron, 2014; Flynn, 2012).

In the face of these problems, social media platforms such as Facebook and Twitter are developing technologies that can detect inappropriate content. They have also been working to delete fake accounts in a way to combat what has come to be known as ‘platform manipulation’. One main problem with this approach, however, is that social media platforms cannot keep up with this flood of fake and suspicious accounts and offensive contents. Additionally, the policies and rules of many social media platforms do not require users to reveal their true identities. The claim thus is that social media platforms are still failing to address the issues of impersonation arising from fake accounts. Parallel to the attempts of social media platforms to combat problems of cyberbullying and impersonation, different machine learning systems and data mining techniques, feature engineering methods, network embedding training, and linguistic stylometry have been developed. These have been widely used in solving problems related to authorship detection of online messages and forensic investigations in general. Approaches of this kind have been concerned with finding the most likely author/s of controversial documents.

Despite the effectiveness of such approaches and techniques in addressing different authorship issues, social media contents still pose real challenges to these approaches and systems. This is attributed to the nature of social media language itself, where users usually post very short texts. To illustrate the argument, it is still difficult for authorship systems to handle the large volume of short texts on social media platforms due to the lack of linguistic information in these posts. In the processing of social media posts, these texts, referred to as documents, have to be mathematically represented in data space to be amenable for computational analysis. In very short posts or documents, the data space is usually sparse or empty as it will be predominately populated by many zeroes due to the shortness of data. The result is that it becomes difficult for authorship

systems to capture the discriminative power of the data or the distinctive features of posts due to what is known in the literature as data sparsity. In a database, sparsity describes the number of cells in a table that are empty, so the matrix (the way data are usually organized in stylometric applications) will be predominately dominated by many zeroes. This can be illustrated in Table One.

Table One. *An example of data sparsity*

User 01	User 02	User 03	User 04	User 05	User 06	User 07
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	3	0	0	0	0
0	0	0	0	0	0	2

Given a data matrix based on the information shown in Table One, it is clear that the datasets are not meaningful. There will be no point in identifying linguistic patterns that can be used as clues for assigning texts to their plausible authors. This problem has negative impacts on the accuracy and therefore reliability of the results of these approaches and systems (Aggarwal & Reddy, 2018; Grčar, Mladenich, Fortuna, & Grobelnik, 2005; MacKay, 2003; Omar, 2020; Wainwright, 2019).

Another problem is that standard authorship systems do not consider the issue of linguistic peculiarities. In the case of Arabic, standard authorship systems do not usually consider the peculiar nature of Arabic. Farghaly and Shaalan (2009) indicate that Arabic linguistic structures are still challenging for Natural Language Processing (NLP) researchers and developers. They explain that despite the peculiar nature of morphology in Arabic being a highly structured and derivational language, standard NLP systems do not consider the critical role of morphology, which has negative impacts on NLP applications. This can be attributed to the fact that different NLP systems are mostly based on English and other European languages with no consideration of other specific-language features. Numerous studies have indicated that different features of Arabic such as diglossia (Farghaly & Shaalan, 2009), rich and complex morphological system (Attia, 2007), the orthographic representation, and morphological and syntactic ambiguities (Attia, 2008) have negative implications on the consistency and reliability of NLP systems.

Despite the development of different Arabic NLP applications that considered the distinctive linguistic properties of Arabic in addressing different NLP applications including tokenization and Part-of-Speech (POS) tagging (Attia, 2007; Diab, 2009; Habash & Rambow, 2005; Nawar, 2014; Roth, Rambow, Habash, Diab, & Rudin, 2008), spell and grammar checking (Kiraz, 2001; McEnery, Hardie, & Younis, 2018), machine translation (Habash, 2010a; Habash & Rambow, 2005; Hamouda, 2014; Mitkov, 2004, 2014; Soudi, Farghaly, Neumann, & Zibib, 2012), and parsing (Bourahma, Mbarki, Mourchid, & Mouloudi, 2017; Habash, 2010b; Khoufi, Aloulou, & Belguith, 2013; Zaki, Hajjar, Hajjar, & Bernard, 2017), very little has been done about authorship recognition in Arabic in general and the language of social media in particular. A need exists therefore to develop reliable authorship recognition methods that take into consideration the peculiar nature of Arabic and social media language. In this regard, this study seeks to address this

gap in the literature by evaluating the effectiveness of linguistic stylometry approaches that considers both the morphological and lexical properties of texts on the effectiveness of authorship recognition in Arabic. The rationale is that morphological patterns and structures of Arabic can be usefully used to supplement the authorship recognition applications in Arabic. As such, this study asks whether the authorship of controversial social media language with a focus on Twitter posts can be recognized via only linguistic stylometry using a combination of morphological and lexical properties.

Literature Review

Stylometry, the investigation of an author's style using quantitative methods to explore the elusive character of his style and the essence of his use of language, has always been central in authorship attribution and recognition studies (Lennon, 2018; Savoy, 2020). These studies have been concerned with exploring the unique stylistic and linguistic properties of authors as a clue for deciding authors of controversial texts. With the development of computational analysis and the expansion of digital texts, computational methods are considered imperative by many critics for their capacities in dealing with large volumes of data (Altintas, Can, & Patton, 2007; Burrows, 2003, 2005, 2007; Holmes, 1998; Holmes & Forsyth, 1995; Jockers, Witten, & Criddle, 2008; Paton & Can, 2004). They argued that linguistic stylometry based on computational methods has proved successful and reliable in developing objective criteria for capturing the style of a text-based on linguistic structures. An earlier attempt of the use of computational methods in identifying the linguistic and stylistic characteristics of authors for deciding authors of controversial texts took place in the 1960s when two American statisticians used computational analysis in their attempt to reveal the real authors of the Federalist Papers (a collection of 85 articles originally published between 1878 and 1888). Mosteller and Wallace (1964) investigated the stylistic use of the function words as discriminators in the disputed texts. The success of the new approach opened the way to the computerized age of stylometry and authorship studies. Over the last three decades, linguistic stylometry based on computational statistical methods has been widely used in investigating the peculiar characteristics of authors for both literary and forensic purposes. This study is only concerned with the applications of linguistic stylometry in forensic applications.

Over the past few decades, stylometry has been extensively used in forensic investigations to reach reliable conclusions regarding the authorship of controversial texts through looking for some internal evidence within the texts using quantitative and statistical methods (Ilseemann, 2019). Given the feasibility of computer-based technology, forensic investigations make use of stylometry to answer authorship questions of disputed and dubious texts (Daultani, & Vijayalakshmi, 2019). Analysts have been concerned with different variables, including lexical features such as frequent words, rare words, token-based word/sentence length; syntactic features such as the use of function words, sentence types, and punctuation marks; and structural features such as indentation for authorship purposes. The assumption is that the identification of such personal distinctive stylistic features makes it possible to detect an author's signature and distinguish the writing of one author from another or others.

In authorship detection applications, forensic linguistics is generally based on the notion of a linguistic fingerprint, which is defined as the process of collecting linguistic data and

features that stamp a speaker/writer as unique. The assumption is that people use language differently and that this difference between people can be observed just as easily and surely as a fingerprint. To do this, forensic linguistics usually adopts quantitative and statistical methods to investigate the linguistic level/s chosen by the researcher (Omar & Deraan, 2019, p. 184).

The chief merit of the quantitative tools adopted in the stylometry approaches is that it is replicable and thus objective. The claim has always been that speculations concerning anonymous and controversial texts can be backed up with objective evidence derived from computational statistical analyses. The replicability of the stylometry tools offers reasonable solutions to the problems of selectivity and subjectivity that have always been attached to traditional stylistic analysis.

Despite its effectiveness in resolving different authorship problems, it is often argued that many traditional fundamental questions concerning authorship recognition remain unresolved (Koppel, Schler, & Argamon, 2013; Rudman, 1997, 2012; Stamatatos, 2009). Stylometric authorship recognition is often blamed because a successful application cannot be appropriately applied to other genres or languages. Many studies, for instance, have indicated that standard stylometric authorship methods are not appropriate for short texts and social media language (Koppel et al., 2013; López-Escobedo, Méndez-Cruz, Sierra, & Solórzano-Soto, 2013). Standard stylometric methods that are used for determining the authors of long texts such as books and articles have been generally unsuccessful in assigning the controversial texts to the known authors correctly. When it comes to very short texts and the language of social media, these usually cannot achieve the same performance due to the sparsity of content in short texts and the peculiar nature of the language of social media. Furthermore, the language of social media is usually unstructured, informal, and ungrammatical as compared to the language of books and articles.

Authorship analysis of online documents is more challenging than analyzing traditional documents due to their special characteristics of size and composition. The traditional literary works such as books and essays are rich sources to learn about the writing style of their authors. Literary works are usually large, ranging from few paragraphs to hundred pages. They are generally well-structured in composition, following definite syntactic and grammatical rules. The study of stylometric features has long been very successful in resolving ownership disputes over literary and conventional writings. Online documents, on the other hand, are short in size, varying from a few words to a few paragraphs, and often they do not follow definite syntactic and/or grammatical rule; making it hard to learn about the learning habits of their authors from such documents" (Fung, Debbabi, & Iqbal, 2020, pp. 29-30)

Concerning the authorship recognition of short texts and the language of social media, different approaches have been developed over the last decades to address the limitations of the literature that traditionally focused on long texts and to find more reliable solutions for the growing authorship detection problems due to the increasing rates of criminal activities as a result of the unprecedented developments of social media channels. These are referred to in the literature as the authorship recognition of micro messages, social media posts, or microblogging messages (Brocardo, Traore, Saad, & Woungang, 2013; Koppel et al., 2013). Researchers have been working

to address the limitations within the stylometric authorship theory and devise new ways that can be applied to very short texts and social media content. They have been using different social media platforms and networks, including Twitter and Facebook to test the developed methods. Raghavan (2010), for instance, introduced the use of both lexical and syntactic features within texts using Probabilistic Context-Free Grammars (PCFG) for maximizing the accuracy of detecting authors of anonymous online messages. Similarly, Bhargava, Mehndiratta, and Asawa (2013) used an integrated method that considers lexical features, syntactic features, and features specific to tweets (such as hash-tags, mentions, and frequency of Emojis) to identify the authors of controversial posts on Twitter. Despite the effectiveness of such methods in improving the accuracy of authorship performance, the results cannot be extended to all other languages, including Arabic. This study attempts to address this gap in the literature by proposing a stylometric authorship model that can be usefully used for determining authors of disputed and controversial online social media texts in Arabic.

Methodology

Methods

Numerous approaches have been developed for stylometric authorship recognition. Document clustering, however, remains one of the most widely used approaches in stylometric authorship studies and applications. The recent decades have witnessed an unprecedented revolution in developing mechanized solutions for organizing the vast quantity of unstructured digital documents and providing powerful tools for turning this unstructured repository into a structured one (Sebastiani, 2006). The literature suggests that document clustering (simply putting similar texts together) is central in almost all authorship applications (Yu, 2008). Document clustering is used as a starting point for many of the authorship systems (Argamon & Olsen, 2006; Horton, Taylor, Yu, & Xiang, 2006; Labbe & Labbe, 2006; Nakamura & Sinclair, 1995; Ramsay, 2005; Tambouratzis & Vassiliou, 2007; Unsworth, 2000; Yu & Unsworth, 2006). The assumption is that documents or texts clustered together are more likely to be written by the same author. Rexha, Kröll, Ziak, and Kern (2018) explain that authorship recognition can be done using document clustering where the author of a disputed or controversial text can be identified from a set of candidate authors.

Theodoridis and Koutroubas (2003) suggest that text clustering is one of the most primitive mental activities of humans. It long preceded the computer age. It was used to handle a large number of information people used to receive. However, only manual clustering was possible where researchers and professionals used their immediate intuitive knowledge of the world in grouping similar texts together. There was no use of quantitative and numerical methods. In other words, text clustering was usually performed in subjective ways that relied heavily on the perception, knowledge, and judgment of the researcher. With more and easier accessibility to electronic digital data in different disciplines and the power of computing data processing on the one hand and the need for maintaining objectivity standards on the other, it has become ever more likely that such procedures must involve automated computational methods (Gordon, 1996) where human intuition and traditional organization methods are replaced by mathematical and computational techniques (Golub, 2006). In this regard, recent years have witnessed the flourishing of automated statistical clustering and classification systems in authorship systems.

In authorship studies, document clustering is used to automatically group natural language texts according to an analysis of their information/ semantic content, using clustering algorithms (Debole & Sebastiani, 2003, 2004). It is a process of grouping similar documents together into distinct sets without labeling them (Maranis & Babenko, 2009). The underlying principle of cluster-based analysis is that closely associated documents tend to be relevant to the same author. In document clustering applications in general, linguistic features of texts (usually lexical and syntactic properties) are extracted to identify the relationship between texts with the purpose of grouping texts that have common linguistic features together (Justo & Torres, 2005; Srivastava & Sahami, 2009). This is usually referred to as content clustering, where clustering of documents is performed based on the words they contain. Content clustering is carried out using computing linguistic the similarity/ distance or what can be called measuring proximity within documents. It has always been argued that linguistic information within documents is key to understanding and determining the content of such documents. The failure of content clustering methods to address different authorship problems has raised many doubts about its reliability. Perhaps the most serious disadvantage with ATC applications is that in almost all text classification schemes, semantic relatedness is merely judged at the level of lexical semantics without taking compositional semantics into account (Gabrilovich & Markovitch, 2007). Another major criticism of ATC applications is that many of the algorithms used for computing linguistic relatedness represent documents as just bags of words where context is not considered.

Later, clustering by context has been introduced as a working approach to evade the problems caused in clustering by content (Attardi, Di Marco, & Salvi, 1998; Attardi, Gulli, & Sebastiani, 1999; Flanagan, 2005; Kovacs, Repasi, Baksa-Varga, & Barabas, 2008; Mirkin, 2005; Pedrycz, 2005). Clustering by context is based on grouping web pages whereby the context surrounding a link is used for categorizing the document referred by the link. The conception is based on the assumption that a web page that refers to a document must necessarily involve enough hints about its content which themselves are sufficient to classify the document (Attardi et al., 1998; Pedrycz, 2005). Many software programs have been devised to execute such tasks, including SenseClusters (Purandare & Pedersen, 2004). These programs make it possible for users to cluster similar contexts, such as emails and web pages (Pedersen, 2008). The working principle of such programs is that data documents can be grouped based on their mutual contextual similarities (Purandare & Pedersen, 2004). Programs of this kind have indeed proven a successful clustering method when applied to web pages and its merits are more tangible with multimedia material. Nevertheless, an approach of this kind carries with it some limitations. The most serious shortcoming is that it is not concerned with the analysis of the content of documents. One more drawback is that in almost all context classification applications “identical replications of controlled experiments result in different conclusions” (Martin, Claes, & Thomas, 2005, p. 470). In this regard, clustering by context is not appropriate for stylometric studies and the study.

Content clustering is used. The rationale is that stylometric authorship is based on inferring or detecting the author of a document through extracting the stylometric features from the document contents, which are strong predictors for authorship determinism. Experimental results of document classification indicate that content word representation has been shown to give promising results in identifying the content of a document. Equally important, most studies agree

that content word representation gives better results than other more sophisticated approaches to clustering (Brocardo et al., 2013; Dhillon, Kogan, & Nicholas, 2004; Frigui & Nasraoui, 2004).

Procedures

Document clustering based on Document Index Graph (DIG) model was used. The model was first introduced by Hammouda and Kamel (2002). The model is based on building a directed graph where each node represents the unique words and edges represent a complete sentence in any document. This model treats a document as a set of sentences rather than a set of words (Hammouda & Kamel, 2002). Unlike the traditional vector space model (VSM), DIG captures the structure of sentences in the document set, rather than single words only (Castillo, Cervantes, Vilarino, & Baez, 2015; Hammouda & Kamel, 2002).

The DIG has some main advantages. These can be summarized as follows. First, it will enable different features or variables to be represented simultaneously. This means that clustering results are based on both common phrases and single words. Second, it allows incremental phrase-based encoding of documents for detecting the style patterns of documents and efficient phrase matching (Lukka & Shaik, 2016; Momina, Kulkarnia, & Chaudharia, 2007). In this way, the DIG is appropriate for representing the letter combinations, single words, and phrases simultaneously. Third, the DIG model finds the matching patterns while building the graph. Finally, DIG provides complete information about the documents and the degree of overlap between every pair of documents, which is useful in determining the similarity between documents. In the present study, each of the texts also referred to as documents, is represented in a DIG matrix where all words and sentences in all the documents are initially included.

Data

A corpus of real-world data derived from Twitter was created. 20,357 tweets by 134 users on the American rapper Nicki Minaj's concert in July 2019 in Saudi Arabia were selected. The data were randomly selected following the hashtag in Arabic #نيكي_ميناج (Nicki Minaj). The announcement of a concert by Minaj in Saudi Arabia triggered a social media storm in July 2019. Many conservative groups from different Arab countries expressed their disappointment considering it against the Saudi customs and values. On the other hand, many advocates considered the concert as one way towards the modernization process the Kingdom is undergoing. Nicki Minaj herself was attacked by many groups for appearing in Saudi Arabia. Although the concert was finally canceled, the posts included numerous offensive words which make it appropriate for the study. Only tweets written in Arabic were selected. It was also decided that posts are written in the Franco-Arabic alphabet (also known as the Arabic chat alphabet or 'Arabizi') to be disregarded. The rationale is that the orthography of Arabic is utterly different from that of Western languages including English. The Franco-Arabic or so-called Arabizi does not follow the rules of the Arabic alphabet. As the current study is concerned with developing a stylometric model of Arabic, the integration of Arabizi or any other writing format with the Arabic alphabet would have a negative bearing on the reliability of the results.

A profile of the linguistic variables of each user was first created. This resulted in 134 profiles. Each profile included all the selected posts of each user. An example is given in Appendix One. All morphological patterns and lexical types were then extracted and mathematically represented so that they are amenable for processing and analysis. In other words, posts were

converted into a string of morphological patterns and lexicons where the frequencies of all morphological patterns and lexical types were calculated.

One main problem with the corpus, however, is that too many unimportant variables were included. These are described as noisy variables or the curse of high dimensionality data. In stylometric applications, the extremely high dimensionality of text data is a significant issue that has a real bearing on the reliability and accuracy of almost all applications based on such data. Although there must be sufficient information (represented in our case in the number of linguistic variables), having too many variables can be difficult or even intractable. The idea is that the larger the data dimensionality, the more difficult it becomes to define the manifold sufficiently well to achieve reliable analytical results. Only the most important variables thus should be selected because irrelevant and redundant variables often degrade the performance of algorithms both in speed and accuracy. To address this issue, principal component analysis (PCA) was used to define only and all the most distinctive morphological and lexical types.

PCA is one of the basic geometric tools that are used to produce a lower-dimensional description of the variables in a data matrix. The primary function of PCA is to find the most informative and distinctive variables within a database or matrix, hence it is possible to reduce the dimensionality of datasets consisting of a large number of interrelated variables while retaining as much as possible of the variation present in the data sets. The underlying principle is that a data matrix in stylometric applications with large data sets can be reduced so that the most distinctive variables are identified with the purpose of best expressing the data and revealing hidden structures and patterns. PCA works to perform a 'good' dimensionality reduction with no significant loss of information as it removes correlated variables within datasets so that it describes the covariance relationships among these variables. The process is done by computing the principal components scores by measuring all the variables in the data set. In so doing the variables that have the highest loading or weight are identified as principal components and other variables are discarded.

Figure 1. A PCA of the morphological types

Based on the PCA of the morphological types and letter combinations, the highest variables were retained. These were decided to be the highest 60 variables. Similarly, a PCA of the lexical types was carried out, as shown in Figure Two.

Figure 2.A PCA of the lexical types

It was decided that only the first 50 variables should be retained. All variables from 51-1667 were removed. The data now is reduced to only 60 morphological types and 50 lexical ones. These are assumed to be the most distinctive features that can help determine the similarity between the texts.

Results & Discussion

As an initial step in assigning the posts to their authors, a DIG model was designed to group similar posts. The hypothesis is that posts grouped or clustered together are more likely to be written by the same author. Generated clusters were compared to author profiles for evaluating the performance of the suggested technique. This is technically known as the profile-based method. Results indicate that the accuracy rate is around 83%. For comparison results, a VSC model was built where only the most distinctive lexical frequencies of the posts were included. The rationale is that word-distribution-based document representations have been widely used in assessing the relatedness of documents.

VSC is based on measuring the relative distances between the row vectors. The distance between any two vectors in space is jointly determined by the size of the angle between the lines joining them to the origin of space's coordinate system, and by the lengths of those lines. VSM is simply a technique where documents are compared with each other than indexed or classified in terms of their similarity or distance based on the words they contain. It can be defined as the organization of a collection of documents usually represented by a vector space model into distinct clusters based on similarity. The theory was first developed by Salton (1971) for IR purposes four decades ago and since then it has become a standard tool in IR systems. The underlying formula of VSM is initially to extract all useful information within a document collection and record it in an index known as a vector space. Then, a proximity measurement is used to compute the semantic similarity among the documents with the purpose of grouping similar documents together.

VSM is the most widely used document representation method in document clustering and classification applications. This is the representation of a set of documents as vectors in a common

vector space with one real-valued component for each term (Salton, 1971; Salton, Wong, & Yang, 1975). The basic hypothesis of VSM in document clustering is the contiguity hypothesis: “documents in the same class form a contiguous region and regions of different classes do not overlap” (Manning, Raghavan, & Schütze, 2008, p. 223). The default method of VSM is a clustered document space where the documents are grouped into classes, each being represented by a class centroid. Salton, Wong, et al. (1975) introduced a simple process for VSM where “word stems extracted from documents or document abstracts (are) weighted per the frequency of occurrence of each term K in the document” (p. 615). This method is usually referred to as a ‘bag of words’. In VSM applications, individual words are used as indexing terms. Vector space representations ignore the word order and the context where words are used. Each document is represented by the number of occurrences of each word in the document in Euclidean vector space where each token in the vector corresponds to a unique/given word in the matrix (Joachims, 2002; Ozgur, 2006).

Results indicate that up to 46% of this accuracy was lost, however, when only lexical features were used, and 27% was lost when the classification was based on the morphological patterns only. The implication here is that standard stylometric authorship methods are not appropriate for the authorship applications in Arabic. It can be concluded that the integration of morphological and lexical features enhance the authorship performance in Arabic texts in general and very short texts in particular. The results reported here agree with the previous studies in the sense that authorship recognition systems and techniques have to address the specific features of the language of social media (Fung, et. al. 2020). They also support the arguments that the Arabic morphological system carries distinctive stylistic features that can be usefully used in authorship recognition applications (Alghamdi & Selamat, 2019; Omar, Elghayesh, & Kassem, 2019; Omar & Hamouda, 2020)

Conclusion

So far, authorship recognition of social media texts in Arabic poses different challenges for researchers, professionals, and program developers. This can be attributed to various reasons, including the linguistic specific rules that make Arabic different from other Western languages and the peculiar nature of the social media language. Thus, this study asked whether it is possible to develop a reliable authorship recognition technique based on integrating the morphological patterns and letter-pair combinations along with the lexical features of texts using the DIG model. Results indicate that the accuracy rate of the proposed technique is around 83% that surpassed the performance of the traditional stylometric authorship systems based on vector space clustering of only the lexical properties of texts. It was also clear that the use of the DIG model had positive impacts on the clustering performance with its capacity to include different linguistic variables for each author profile. Given the anonymous nature of social media and the increasing rates of cybercrimes in recent years, the findings of the study can be usefully used in forensic investigations and detecting authors of illegal and inappropriate posts and messages.

Acknowledgments

I take this opportunity to thank Prince Sattam Bin Abdulaziz University in Saudi Arabia alongside its Deanship of Scientific Research, for all technical support it has unstintingly provided towards the fulfillment of the current research project.

About the authors:

Haroon Alsager finished his Ph.D. in linguistics at Arizona State University in 2017. His research interests include syntax, historical linguistics, and computational linguistics. Currently, he is an assistant professor of linguistics at Prince Sattam Bin Abdulaziz University.

ORCID: <http://orcid.org/0000-0003-3778-5801>

References

- Agarwal, N., Dokoochaki, N., & Tokdemir, S. (Eds.). (2019). *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*. Cham: Switzerland: Springer.
- Aggarwal, C. C., & Reddy, C. K. (2018). *Data Clustering: Algorithms and Applications*. New York: Chapman and Hall/CRC.
- Alghamdi, H. & Selamat, A. (2019). Arabic Web Page Clustering: A Review. *Journal of King Saud University-Computer and Information Sciences*. 31 (1), 1-14
- Altintas, K., Chan, F., & Patton, J. M. (2007). Language Change Quantification Using Time-separated Parallel Translations. *Literary and Linguistic Computing*, 22(4), 375-393. doi:10.1093/LLC/fqm026
- Anderson, M. (2018). *A Majority of Teens Have Experienced Some Form of Cyberbullying*. Washington: Pew Research Center.
- Argamon, S., & Olsen, M. (2006). Toward Meaningful Computing. *Communications of ACM*, 49(4), 33-35. doi:<http://doi.acm.org/10.1145/1121949.1121972>
- Attardi, G., Di Marco, S., & Salvi, D. (1998). Categorization by Context. *Journal of Universal Computer Science*, 4 (9), 719–736.
- Attardi, G., Gulli, A., & Sebastiani, F. (1999). *Automatic Web Page Categorization by Link and Context Analysis*. In C. Hutchison, & G. Lanzaone, (eds.), *THAI 1999*, 105–119
- Attia, M. A. (2007). Arabic Tokenization System. *Proceedings of the 5th Workshop on Important Unresolved Matters, Prague, Czech Republic*, 65–72.
- Attia, M. A. (2008). *Handling Arabic morphological and syntactic ambiguities within the LFG framework with a view to machine translation*, (Unpublished Ph.D. Dissertation). University of Manchester, Manchester.
- Bhargava, M., Mehndiratta, P., & Asawa, K. (2013). Stylometric Analysis for Authorship Attribution on Twitter. V. Bhatnagar & S. Srinivasa, (Eds.), *BDA 2013, LNCS 8302*, 37–47
- Bourahma, S., Mbarki, S., Mourchid, M., & Mouloudi, A. (2017). Syntactic Parsing of Simple Arabic Nominal Sentence Using the NooJ Linguistic Platform. In A. Lachkar, K. Bouzoubaa, A. Mazroui, A. Hamdani, & A. Lekhouaja (Eds.), *Arabic Language Processing: From Theory to Practice* (244-257). London: Springer.
- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5), 662-679. <http://doi.org/10.1080/1369118X.2012.678878>
- Brocardo, M. L., Traore, I., Saad, S., & Woungang, I. (2013). Authorship Verification for Short Messages using Stylometry. *Proceedings of the IEEE International Conference on Computer, Information and Telecommunication System (CITS 2013)*, Piraeus-Athens, Greece.
- Budinger, T. F., & Budinger, M. D. (2006). *Ethics of Emerging Technologies: Scientific Facts and Moral Challenges*. Hoboken: N. J.: John Wiley & Sons.
- Burrows, J. F. (2003). Questions of Authorship: Attribution and Beyond A Lecture Delivered on the Occasion of the Roberto Busa Award ACH-ALLC 2001, New York. *Computers and the Humanities*, 37(1), 5-32. DOI: [10.1023/A:1021814530952](https://doi.org/10.1023/A:1021814530952)

- Burrows, J. F. (2005). Who Wrote Shamela? Verifying the Authorship of a Parodic Text. *Literary and Linguistic Computing*, 20(4), 437-450. doi:10.1093/LLC/fqi049
- Burrows, J. F. (2007). All the Way Through Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing*, 22(1), 27-47. doi:10.1093/LLC/fqi067
- Castillo, E., Cervantes, O., Vilarino, D., & Baez, D. (2015). Author attribution using a graph-based representation. *Proceedings of the International Conference on Electronics, Communications, and Computers (CONIELECOMP)*.
- Citron, D. K. (2014). *Hate Crimes in Cyberspace*. Cambridge, MA: Harvard University Press.
- Debole, F., & Sebastiani, F. (2003). Supervised term weighting for automated text categorization. *Proceedings of the 2003 ACM symposium on Applied computing*, 784-788. doi:http://doi.acm.org/10.1145/952532.952688
- Debole, F., & Sebastiani, F. (2004). Supervised Term Weighting for Automated Text Categorization. *ERCIM News* 56, 55-56.
- Dhillon, I., Kogan, J., & Nicholas, C. (2004). Feature Selection and Document Clustering. In M. W. Berry (Ed.), *Survey of Text Mining: Clustering, Classification, and Retrieval*. New York: Springer.
- Diab, M. (2009). Second Generation AMIRA Tools for Arabic Processing: Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking. *The 2nd International Conference on Arabic Language Resources and Tools*. Cairo, Egypt.
- Doultani, A. & Vijayalakshmi, M. (2019). Data Forensics On Social Media. *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, 1-5.
- Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4), 1-22. doi:http://doi.acm.org/10.1145/1644879.1644881
- Flanagan, J. A. (2005). Unsupervised clustering of context data and learning user requirements for a mobile device. *The Proceedings of the 5th international conference on Modeling and Using Context*, Paris, France.
- Flynn, N. (2012). *The Social Media Handbook: Rules, Policies, and Best Practices to Successfully Manage Your Organization's Social Media Presence, Posts, and Potential*. Hoboken: John Wiley & Sons.
- Frigui, H., & Nasraoui, O. (2004). Simultaneous Clustering and Dynamic Keyword Weighting for Text Documents. In M. W. Berry, (Ed.), *Survey of Text Mining: Clustering, Classification, and Retrieval* (pp. 87-94). New York: Springer.
- Fung, B. Debbabi, M. & Iqbal, F. (2020). *Machine Learning for Authorship Attribution and Cyber Forensics*. London: Springer Nature
- Gabrilovich, E., & Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 1606--1611.
- Golbeck, J. (2018). *Online Harassment*. Cham: Switzerland: Springer.
- Golub, K. (2006). Automated subject classification of textual Web documents. *Journal of Documentation*, 62(3), 350-371. <https://doi.org/10.1108/00220410610666501>
- Gordon, A. D. (1996). Hierarchical Classification. In P. Arabie, L. J. Hubert, & G. d. Soete (Eds.), *Clustering and classification* (pp. ix-490). Singapore; River Edge, NJ.: World Scientific.
- Görzig, A., & Frumkin, L. A. (2013). Cyberbullying experiences on-the-go: When social media can become distressing. *Cyberpsychology*, 7(1). <https://doi.org/10.5817/CP2013-1-4>
- Grčar, M., Mladenič, D., Fortuna, B., & Grobelnik, M. (2005). Data sparsity issues in the collaborative filtering framework. *The International Workshop on Knowledge Discovery on the Web*.
- Habash, N. (2010a). *Arabic Natural Language Processing*. San Rafael, California: Morgan & Claypool Publishers.
- Habash, N. (2010b). *Introduction to Arabic Natural Language Processing*. San Rafael, California: Morgan & Claypool Publishers.

- Habash, N., & Rambow, O. (2005). Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. *Proceedings of the 43rd Annual Meeting of the ACL*, 573–580. DOI: [10.3115/1219840.1219911](https://doi.org/10.3115/1219840.1219911)
- Hammouda, K. M., & Kamel, M. S. (2002). Phrase-based Document Similarity Based on an Index Graph Model. *The Proceedings of the IEEE International Conference on Data Mining*.
- Hamouda, W. (2014). *Anaphora Resolution for Arabic Machine Translation: A Case Study of Nafs*, (Unpublished Ph.D. dissertation). University of Newcastle Upon Tyne, Newcastle. Retrieved from <https://books.google.com.sa/books?id=IOfMoQEACAAJ>
- Holmes, D. I. (1998). The Evolution of Stylometry in Humanities Scholarship. *Lit Linguist Computing*, 13(3), 111-117. doi:10.1093/LLC/13.3.111
- Holmes, D. I., & Forsyth, R. S. (1995). The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing*, 10(2), 111-127. doi:10.1093/LLC/10.2.111
- Horton, T., Taylor, C., Yu, B., & Xiang, X. (2006). 'Quite Right, Dear and Interesting': Seeking the Sentimental in Nineteenth-Century American Fiction. *The Digital Humanities Conference*, Paris-Sorbonne, France.
- Ilseemann, H. (2019). Forensic stylometry. *Digital Scholarship in the Humanities*, 34(2), 335-349.
- Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms*. Amsterdam: Kluwer Academic Publishers.
- Jockers, M. L., Witten, D. M., & Criddle, C. S. (2008). Reassessing authorship of the Book of Mormon using delta and nearest shrunken centroid classification. *Literary and Linguistic Computing*, 23(4), 465-491. doi:10.1093/LLC/fqn040
- Justo, R., & Torres, I. (2005). Statistical and Linguistic Clustering for Language Modeling in ASR. In A. Sanfeliu, & M. L. Cortes (Eds.), *Progress in Pattern Recognition, Image Analysis, and Applications* (Vol. 3773, pp. 556-565). Berlin, Heidelberg: Springer.
- Khoufi, N., Aloulou, C., & Belguith, L. H. (2013). ARSYPAR: A tool for parsing the Arabic language based on supervised learning. *The International Arab Conference on Information Technology*, Zarqa University.
- Kiraz, G. A. (2001). *Computational Nonlinear Morphology: With Emphasis on Semitic Languages*. Cambridge Cambridge University Press.
- Koppel, M., Schler, J., & Argamon, S. (2013). Authorship Attribution: What's Easy and What's Hard? *Journal of Law and Policy*, 21, 317-331.
- Kovacs, L., Repasi, T., Baksa-Varga, E., & Barabas, P. (2008). Clustering Based on Context Similarity. *The First International Conference on Complexity and Intelligence of the Artificial and Natural Complex Systems. Medical Applications of the Complex Systems*. Biomedical Computing, Targu Mures, Mures.
- Kowalski, R. M., Limber, S. P., & Agatston, P. W. (2012). *Cyberbullying: Bullying in the Digital Age*. Chichester, West Sussex, UK; Malden, MA: Wiley-Blackwell.
- Labbe, C., & Labbe, D. (2006). A Tool for Literary Studies: Intertextual Distance and Tree Classification. *Lit Linguist Computing*, 21(3), 311-326. doi:10.1093/LLC/fqi063
- Lennon, B. (2018). *Passwords: Philology, Security, Authentication*. Harvard: Harvard University Press
- López-Escobedo, F., Méndez-Cruz, C.-F., Sierra, G., & Solórzano-Soto, J. (2013). Analysis of Stylometric Variables in Long and Short Texts. *Procedia - Social and Behavioral Sciences*, 95, 604 – 611.
- Lowry, P. B., Zhang, J., Wang, C., & Siponen, M. (2016). Why do adults engage in cyberbullying on social media? Integration of online disinhibition and deindividuation effects with the social structure and social learning model. *Information Systems Research*, 27(4), 962-986. <https://doi.org/10.1287/isre.2016.0671>
- Lukka, S., & Shaik, R. (2016). A Well Organized Phrase-Based Document Clustering Using ASCII Values and Adjacency List. *Proceedings of the Eighth International Conference on Soft Computing and Pattern Recognition (SoCPaR 2016)*, 113-120.

- MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
- Maranis, H., & Babenko, D. (2009). *Algorithms of the Intelligent Web*. Greenwich: Manning Publications Co.
- Martin, H., Claes, W., & Thomas, T. (2005). Experimental context classification: Incentives and experience of subjects. *The Proceedings of the 27th international conference on Software engineering*, St. Louis, MO, USA.
- McEnery, T., Hardie, A., & Younis, N. (Eds.). (2018). *Arabic Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Mirkin, B. (2005). *Clustering for Data Mining: A Data Recovery Approach*. CRC Press.
- Mitkov, R. (2004). *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press.
- Mitkov, R. (2014). *Anaphora Resolution*. London; New York: Routledge.
- Momina, B. F., Kulkarnia, P. J., & Chaudharia, A. A. (2007). Web Document Clustering Using Document Index Graph. *International Journal of Information Processing*, 1(2), 49 - 57.
- Mosteller, F., & Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Reading, Mass.: Addison-Wesley Pub. Co.
- Nakamura, J., & Sinclair, J. (1995). The World of Woman in the Bank of English: Internal Criteria for the Classification of Corpora. *Literary and Linguistic Computing*, 10(2), 99-110. doi:10.1093/LLC/10.2.99
- Nawar, M. (2014). Improving Arabic Tokenization and POS Tagging Using Morphological Analyzer. In H. A.E., T. M.F., & T. Azar (Eds.), *Advanced Machine Learning Technologies and Applications* (Vol. 488, pp.46-53). Cham: Springer.
- Omar, A. (2020). Feature Selection in Text Clustering Applications of Literary Texts: A Hybrid of Term Weighting Methods. *International Journal of Advanced Computer Science and Applications*, 11(2), 99-107. DOI: [10.14569/IJACSA.2020.0110214](https://doi.org/10.14569/IJACSA.2020.0110214)
- Omar, A. & Deraan, B. (2019) Towards a Linguistic Stylometric Model for the Authorship Detection in Cybercrime Investigations. *International Journal of English Linguistics*. 9 (5), 182-192
- Omar, A., Elghayesh, B. I., & Kassem, M. A. M. (2019). Authorship Attribution Revisited: The Problem of Flash Fiction: A morphological-based Linguistic Stylometry Approach. *Arab World English Journal*, 10(3)318-329. DOI: <https://dx.doi.org/10.24093/awej/vol10no3.22>
- Omar, A. & Hamouda, W. (2020) The Effectiveness of Stemming in the Stylometric Authorship Attribution in Arabic. *International Journal of Advanced Computer Science and Applications*. 11 (1), 116-121
- Ozgur, Y. (2006). *Empirical selection of NLP-driven document representations for text categorization*. Syracuse, New York: Syracuse University.
- Paton, J. M., & Can, F. (2004). A Stylometric Analysis of Yas, ar Kemal's I_nce Memed Tetralogy. *Computers and the Humanities* 38, 457-467.
- Pedersen, T. (2008). Computational Approaches to Measuring the Similarity of Short Contexts: A Review of Applications and Methods. <https://arxiv.org/abs/0806.3787>
- Pedrycz, W. (2005). *Knowledge-Based Clustering: From Data to Information Granules*: Wiley.
- Purandare, A., & Pedersen, T. (2004). SenseClusters - Finding Clusters that Represent Word Senses. *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*. San Jose, USA, July 25-29, 2004.
- Raghavan, S. (2010). Authorship Attribution Using Probabilistic Context-Free Grammars. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. (ACL 2010), Uppsala, Sweden.

- Ramsay, S. (2005). In Praise of Pattern. *TEXT Technology: the Journal of Computer Text Processing*, 14(2), 177-190. DOI:10.1093/LLC/18.2.167
- Reader, B. (2012). Free press vs. free speech? The rhetoric of “civility” in regard to anonymous online comments. *Journalism & Mass Communication Quarterly*, 89(3), 495-513. DOI: 10.1177/1077699012447923
- Rexha, A., Kröll, M., Ziak, H., & Kern, R. (2018). Authorship identification of documents with high content similarity. *Scientometrics*, 115(1), 223-237. doi:10.1007/s11192-018-2661-6
- Roth, R., Rambow, O., Habash, N., Diab, M., & Rudin, C. (2008). Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. *Proceedings of ACL-08: HLT*. Short Papers, Columbus, Ohio.
- Rudman, J. (1997). The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities*, 31(4), 351-365.
- Rudman, J. (2012). The State of Non-Traditional Authorship Attribution Studies—2012: Some Problems and Solutions. *English Studies*, 93(3), 259-274. doi:10.1080/0013838X.2012.668785
- Salton, G. (1971). *The Smart retrieval system: experiments in automatic document processing*. Englewood Cliffs: Prentice Hall Inc.
- Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11), 613–620. <https://doi.org/10.1145/361219.361220>
- Savoy, J. (2020). *Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling*. New York: Springer International Publishing
- Schallbruch, M., & Skierka, I. (2018). *Cybersecurity in Germany*. Cham: Switzerland: Springer.
- Sebastiani, F. (2006). Classification of Text, Automatic. . In K. Brown (Ed.), *Encyclopedia of Language & Linguistics* (2nd ed., Vol. 2, pp. 457-462). Oxford: Elsevier.
- Soudi, A., Farghaly, A., Neumann, G., & Zibib, R. (2012). *Challenges for Arabic Machine Translation*. John Benjamins Publishing.
- Srivastava, A. N., & Sahami, M. (Eds.). (2009). *Text Mining Classification, Clustering, and Applications*. London: Chapman and Hall.
- Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of American Society for Information, Science, and Technology*, 60(3), 538-556. doi:10.1002/asi.v60:3
- Tambouratzis, G., & Vassiliou, M. (2007). Employing Thematic Variables for Enhancing Classification Accuracy Within Author Discrimination Experiments. *Literary and Linguistic Computing*, 22(2), 207-224. doi:10.1093/LLC/fqm003
- Theodoridis, S., & Koutroubas, K. (2003). *Pattern Recognition* (2nd ed.). San Diego, CA: Academic Press.
- Unsworth, J. (2000). Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This? *The Symposium on Humanities Computing: Formal Methods, Experimental Practice*. King’s College, London.
- Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge: Cambridge University Press.
- Yu, B. (2008). An Evaluation of Text Classification Methods for Literary Study. *Literary and Linguistic Computing*, 23(3), 327-343. doi:10.1093/LLC/fqn015
- Yu, B., & Unsworth, J. (2006). Toward Discovering Potential Data Mining Applications in Literary Criticism. *The Digital Humanities*, Paris-Sorbo.
- Zaki, Y., Hajjar, H., Hajjar, M., & Bernard, G. (2017). Towards the development of a statistical parser of the Arabic language. *The 2017 Computing Conference*, 85-87. London, United Kingdom

Appendix One An example of a user profile

User	Profile	Transliteration	English translation
ID01	نيكي ميناج بتسوي حفله ف السعوديه؟ شقول ، شخلي	nyky mynaj btswy hafla f als'ewdyh? shaqul , shakhly	Nicki Minaj will headline a music

	يقولج نص الكلام بيكون مشفر ليش الانفجار الي ورا	yaqulj ns alklam bykwn mshafr lysh alanfjar aly wra	festival in Saudi Arabia. What to do? They say that half the speech would be coded. Why is the explosion in the back?
ID02	نشكر #السعودية على إلغاء حفل #نيكي ميناج للمرة الثانية أشكر #السعودية على إلغاء حفل #نيكي ميناج	nashkr als'ewdyh 'ela elgha' hafll nyky_mynaj llmara althanyh ashkr als'ewdyh 'ela elgha' hafll nyky_mynaj	We thank Saudi Arabia for cancelling Nicki Minaj's music festival Again, I thank Saudi Arabia for cancelling Nicki Minaj's music festival
ID03	بيقولوا نيكي ميناج هتعمل حفله ف السعودية دى ست مش راجل انا سمعت الخبر على الراديو اسمها لا تستطيع تحديد راجل او امرأة اللي اختارتة لنفسها في انجلترا	byqulu nyky mynaj ht'eml hflh f als'ewdyh da sit msh rajl ana sm'et alkhabr 'ela alradyw asmha la tastty'e thadyd rajl aw amrah alla akhtarh lnfsa fa anjltra	They claim that Nicki Minaj will headline a music festival in Saudi Arabia. She is a woman, not a man. I heard this news item on the radio The name she chose for herself in England goes for both men and women
ID04	شصار على #نيكي ميناج وين وروني خل العن والدينها بالانجليزي	shusar 'ela nyky_mynaj wyn warwny khl al'en waldynha balanqlyzy	What is new about Nicki Minaj? Where is she? I want to insult her parents in English.
ID05	خبر إلغاء الحفلة خبر سار فله الحمد والمنة	khabr elgha' alhafllh khabr sar flh alhmd walmnh	The cancellation of the music festival is really a news item. All praise and thanks are due to Allah.
ID06	حتى اسمها يجيب الغثيان	hta asmha yjyb alghthyan	Even her name causes nausea
ID07	والله سررنا بهذا الخير. وان كنت متأكد من البداية ان القصة مختلقة. من أعداء السعودية. لتشويه صورة المملكة خاصة وأننا في موسم الحج.	wallh srna bhda alkhbr. wan knt mtakd mn albdyayh an alqsa mkhtlqa. mn a'eda' als'ewdyh. ltshwyh surt almmlka khashan wanna fy mwsam alhj.	We are so pleased with this news item. I was so sure from the very beginning that this story is fabricated by the enemies of Saudi Arabia to distort its image, especially we are in the season of pilgrimage.

ID08	اوامر عليا من قيادة الدولة تلغي حفل هذي الشاذه وليس هي من اعتذرت ...	awamr 'elya mn qyadat aldwla tlghy hafl hady alshadh wlaya hy mn a'etdrt ...	Supreme orders have been issued by state leaders to cancel the music festival scheduled to be performed by this gay woman. She did not excuse herself.
ID09	صدرت توجيهات عليا بإلغاء حفل #نيكي_ميناج! ...	sadrt twjyhat 'elya belgha' hafl nyky_mynaj ... !	Supreme orders have been given to cancel Nicki Minaj's music festival!
ID10	لحقّ او ما تلحق اغنية مسربه من حفل #نيكي_ميناج في #جده	lhq aw ma tlhq aghnyh msrbh mn hafl nyky_mynaj fy jdth	Hurry up! Don't miss a leaked song from Nicki Minaj's concert in Jeddah
ID11	كل شي بصوتها حلو فديتها انهرتتت رسميا كنسلت	kul shy bswtha hlw fdytha anhrttt rsmyan kanslt	Everything she sings is fantastic. I adore her. Officially, the music festival has been canceled.
ID12			
ID13	هما الجماعة متغاضين من #نيكي_ميناج ليه,, ولا امكن انها موزة طحن	hma aljma'eh mtghazyn mn nyky_mynaj lyh , wla a_kmn anha mwzh thn	Why are those people displeased with Nicki Minaj? Is this because she is really hot chick?
ID14	طب و بعدين يا جدعان يسطا انت عارف يعني ايه نيكي ميناج تعمل حفلة في السعودية؟! تسمع عن نيكي ميناج طيب	tb w b'edyn ya jd'ean ysta ant 'earf y'eny ayh nyky mynaj t'eml hafla fy als'ewdyh?! tsm'e 'en nyky mynaj tyb	What shall we do guys? You man, do you know what does it mean that Nicki Minaj will headline a music festival in Saudi Arabia?! Have you ever heard of Nicki Minaj
	بيقولوا نيكي ميناج هتعمل حفله ف السعودية الله غالب حبيبنا كلنا وبس بقي عشان فيها تأبيده دي	byqwlwa nyky mynaj ht'eml hflh f als'ewdyh allh ghalb hbybna klna wbs bqy 'eshan fyha tabydh dy	It is said that Nicki Minaj will headline a music festival in Saudi Arabia Nothing to do You are our darling, indeed Calm down; otherwise we would be sentenced for life imprisonment.