IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 16, Issue 2, Ser. II (Mar - Apr 2021), PP 01-05
www.iosrjournals.org

Author Prediction in Text Mining of the Opinion Articles in Arabic Newspapers

Naji Shukri Alzaza
Department of Software Engineering
University of Palestine, Gaza, Palestine

Abstract
Text mining provides a valuable technique for author prediction of the opinion articles in Arabic newspapers. Opinion articles are one of the most components in any newspaper, especially daily newspapers. Analyzing such articles, especially in Arabic, did not receive adequate attention from scientific research. Meanwhile, from a professional perspective, the media and journalism organizations need to ensure the authenticity of the article concerning the original author and verify its objective content. Dataset was collected officially via the newspaper from the last six years of the opinion Arabic articles that already published online between 2016-2021. 8109 articles of 428 authors were preprocessed. The model was evaluated and tested for two months of new random articles for the same authors. The Neural Network algorithm were the highest scores (81.1%), with a none significant difference for Logistic Regression (80.8%). Indeed, the Neural Network algorithm was utilized in the real evaluation. Analyzing opinion articles could make a theme for each author which identifies his/her uniqueness. The model provides a new paradigm for originality test of the opinion's articles in Arabic language. Results show that an author with 10 articles is the minimum to get quite a prediction of his/her new articles.

Keywords: *Arabic Text Mining, Author Prediction, Opinion Articles Analysis, Newspaper Mining*

Date of Submission: 02-04-2021 Date of Acceptance: 16-04-2021

I. Introduction

Text is mostly the objects that is used to express among other multimedia such as images, sounds, and videos. People in all media use text to express their feelings, opinions, or ideas [1, 2]. Furthermore, social media, newspapers, websites utilize Search Engine Optimization (SEO) on text to speed up people's need for the search. The multiplicity and ease of electronic publishing tools have given humanity a great opportunity for expression and communication, which has increased the volume of data steadily.

Articles have reserved their space in all types of media as well as new media. Articles could be published daily, weekly or more. Moreover, text mining has been utilized in all publication media such as newspapers, social media, company press releases, user reviews about experiences and products as well as scientific articles and discourse [1]. However, there are several synonyms of text mining such as opinion mining, text analysis, unstructured text, and knowledge extraction, natural language processing (NLP) [1]. According to IDC report by 2024, 50% of organizations will rely on automated data analysis to make their decisions, of which unstructured data will account for 80% [3].

Articles play an important role in shaping public opinion and reflecting attitudes and trends towards current events and developments. Meanwhile, from a professional perspective, the media and journalism organizations need to ensure the authenticity of the article concerning the original author and verify its objective content. However, text mining provides effective classification tools and algorithms to predict the author based on the analysis of articles. Articles long whose word count ranges between 350-1000 words, make the decision-maker (editor) unable to audit and decide in a manual way. Indeed, most qualitative studies other than Information Technology on the newspaper, do not utilize text mining.

Several studies have covered analyzing publications and press news, knowledge extraction, and topic modeling via social media especially Facebook and Twitter [1, 4, 5]. Trends and patterns investigated in two million articles of English newspapers from 1989 to 2012 [6].

Text mining in the Arabic Language has its special issues [7, 8] such as writing direction from right-to-left; different shapes for some letters based on their position in the same word; there is no capitalization for letters; has two types of sentence: noun sentence and actual sentence which include many parts accordingly and many of the Arabic morphological issues, which is not in this research scope.

Despite opinion articles are one of the most components in any newspaper, especially daily newspapers such as the Felesteen newspaper, analyzing opinion articles, especially in Arabic, did not receive adequate attention from scientific research.

II. Text Mining Prediction Models

Choose the proper model for the specific data set is one of the most critical issues in data mining. In text mining that is usually depends on the technique such as classification, clustering, knowledge extraction (topic modeling),sentiment analysis (opining mining), etc. [9]. However, comparing the utilized models could figure out which is better based on its accuracy.

Several studies that are reviewed by a comprehensive study, which are included in the classification models such as Support Vector Machines (SVM), Naive Bayes, Decision Tree, K-nearest neighbor (KNN), and Artificial Neural Networks (ANN) [9, 6].

Classification and clustering models are the main catogries in text mining [1, 6], while sentiment analysis or opinion miningis mostly utilized in social media analysis [10, 11, 4, 12, 13], arabic knowledg extraction is also has researchers attention [4, 14, 15].

Antons, Grünwald, Cichy, and Salge (2020) argued that utilizing text mining need not much literature demonstration about its benefits and case studies since it has reached of maturity stage in innovation. However, investigation of 124 articles found that 58.9% of them used less than 10,000 corpora (texts) [1].

III. Methodology

As shown in Figure 1, the text mining processing has underlying steps [1] that comprises data gathering, text preprocessing, algorithms testing/evaluation, choose the best algorithm. The prediction model needs one more step that is the combining of new corpora (articles) with the algorithm was chosen. However, iteration is the normal process to improve results in each step.

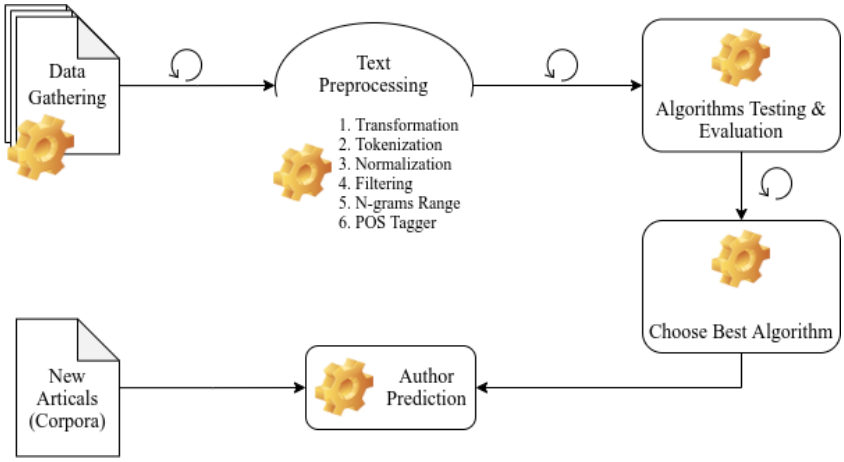

Figure 1: Text mining process

However, the new corpora should be in the same structure of the texts which collecting for testing (i.e. training and testing) and evaluation of the algorithm.

A. Data Gathering

A real dataset was collected officially from the newspaper, from the last six years of the opinion Arabic articles that already published online between 2016-2021. 8109 articles of 428 authors were preprocessed (see Table 1). Each corpus represents one article which includes a full Arabic text between 300-1000 words. The model was evaluated for two months of new random articles for the same authors.

Table 1: Dataset for a valid processing

Articles	Authors	Tokens (Words)
8109	**428**	**1,410,761**

B. Text Preprocessing

Preprocessing is the major step in text mining preparing the text for analysis [1] and data cleaning is the most sensitive stage of the data mining process. However, text mining needs many efforts for text preprocessing especially in Arabic corpora. Six steps of the text preprocessing were proceeded that are Transformation,

Tokenization, Normalization, Filtering, and POS Tagger. Word cloud has been utilized to check the results of the text preprocessing. The most words of 1,410,761 tokens are in political issues which reflects the nature of the daily newspaper.

1. Transformation: Text transformation is the process to clean up the not-needed characters of each corpus. Corpora have been parsed to omit HTML tags and Uniform Resource Locator (URL), i.e. the websites links. Furthermore, all non-Arabic characters have been omitted to decrease the noise that may affect significantly the text analysis process.

2. Tokenization: Text tokenization is the process of breaking up a sequence of strings into pieces such as words, keywords, phrases, statements, symbols, and other elements called tokens. Tokenization may proceed based on word punctuation, whitespace, sentence or regular expression (i.e. \w+). This research found word punctuation is the most proper token for the Arabic articles' dataset. Indeed, each article has tokenized into words. As shown in Table 1, the dataset provides 1,410,761 tokens (i.e. words) from 8109 articles for 428 authors.

3. **Normalization:** Text normalization aims to put all text on the same level in processing. Stemming and Lemmatization are mostly used in text mining. Stemming is defined as the process that returns the segment of the word left after removing some prefixes and suffixes from the word [16]. Lemmatization is the process of converting a word to its base form. While lemmatization considers the text context and converts the word to its meaningful base form, stemming just removes some extra characters. Furthermore, stemming may lead to incorrect meanings and spelling errors. However, this research utilized UDPipelemmatizer with special Arabic package supporting [17]. The UDPipelemmatizer provide four services that are: Tokenization, Parts of Speech tagging, Lemmatization, and Dependency Parsing.

4. Filtering: Text filtering is very important to cleanup some noisy terms. This research utilized five filters: a predefined list of Stopwords removal, Lexicon, word frequency between 10%-90% in the article, English words removals, Regular expression removal for special signs such as "?, !, $,&, .. ".

5. N-grams Range: Text meaning, especially in Arabic depends on two or more words to form a complete sentence. N-grams provide several ranges for joint meaning which may be lost in a single word, i.e. N-grams range =1 means that process based on unigrams or one word, etc. This research utilized N-grams range between 1-2, that means process should be less than or equal two words.

6. **POS Tagger:** Part-Of-Speech Tagger (POS Tagger) is an algorithm reads text in a specific language and assigns parts of speech to each word (token), such as noun, verb, adjective, etc. [18]. However, this research utilized Treebank POS Tagger (MaxEnt) with Arabic supporting package. The MaxEnd "Maximum Entropy Model" provides high level of accuracy (96.6%) [19].

IV. Testing and Evaluation

Six classification algorithms have been evaluated and tested on the dataset. As shown in Table 2, the Classification Accuracy (CA) and Area Under the Curve (AUC) of the Neural Network algorithm were the highest scores (81.1%), with a none significant difference for Logistic Regression (80.8%). Indeed, the Neural Network algorithm was utilized in the real evaluation and testing. The final model is shown in figure 2, which utilizes the Neural Network and Logistic Regression as the most two predictors algorithms among the six algorithms.

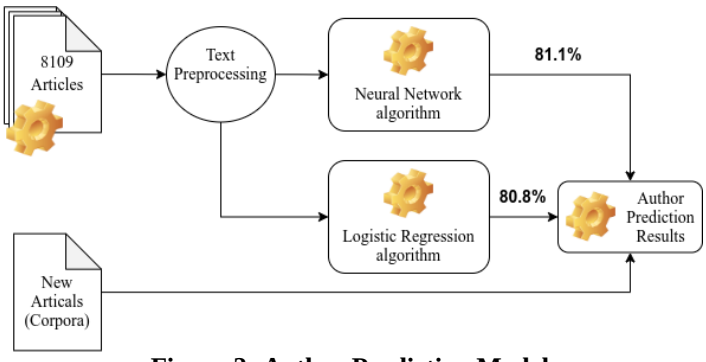

Figure 2: Author Prediction Model

Table 2: Test and scores of classification

Model	AUC	CA	F1	Precision
Neural Network	**0.960**	**0.811**	**0.783**	**0.766**
Logistic Regression	**0.958**	**0.808**	**0.775**	**0.756**
SVM	0.951	0.783	0.750	0.739
Random Forest	0.857	0.567	0.521	0.536
kNN	0.812	0.401	0.392	0.447
Naive Bayes	0.721	0.002	0.001	0.001

Furthermore, Figure 3 shows the Results of Classifiers' (ROC) tests on the dataset which presents the mean of comparison between the six classification models. However, the more accurate classifier is the closer curve that follows the left-hand border and then the top border of the ROC space.

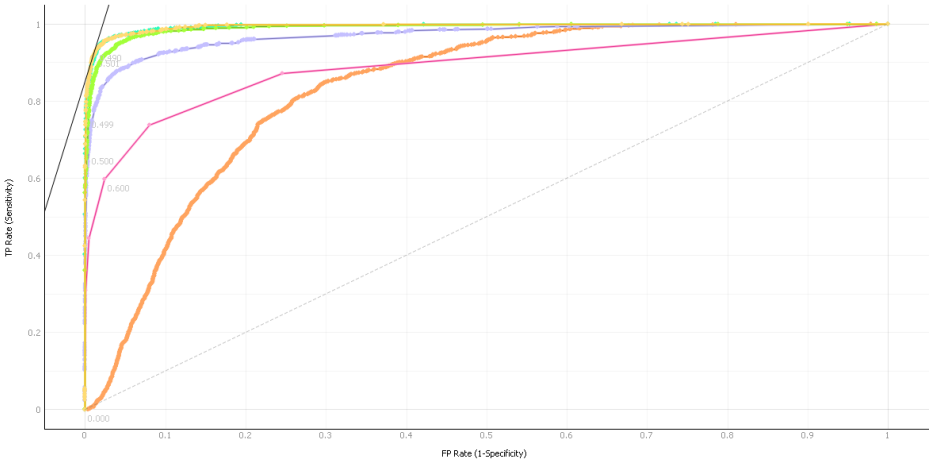

Figure 3: ROC Analysis

As shown in figure 4, the model provided a true prediction for a particular author between 68%-100%. However, there were three articles predicted for the mistaken author. When checking back manually of such mistaken prediction articles, found that the author changed his writing interesting for a different writing theme. Indeed, the model accuracy depends on the author's writing theme and his/her topic interesting which normally, enhances his/her tokens.

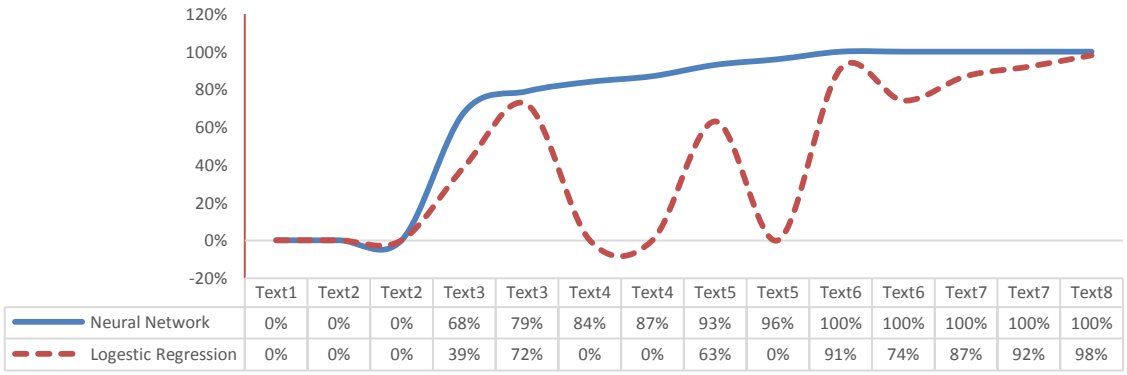

Figure 4: Sample result of an author prediction

V. Conclusion

Text mining provides a valuable technique for author prediction of the opinion articles in Arabic newspapers. Data clearing is a major effort that needs several preprocessing, especially in Arabic corpora. Furthermore, preprocessing, training, and testing of algorithms in articles text mining need high-performance hardware. Analyzing opinion articles could make a theme for each author which identifies his/her uniqueness. The model provides a new paradigm for originality test of the opinion's articles in Arabic language.

Furthermore, results show that an author with 10 articles is the minimum to get quite a prediction of his/her new articles. This research may encourage researchers in media-scope to utilize the text mining instead of the conventional way which depends on the manual process.

References

[1] D. Antons, E. Grünwald, P. Cichy und T. O. Salge, „The application of text mining methods in innovation research: current state, evolution patterns, and development priorities," *R&D Management*, pp. 297-308, 2020.

[2] S. A. Salloum, M. Al-Emran and K. Shaalan, "Mining Social Media Text: Extracting Knowledge from Facebook," *International Journal of Computing and Digital Systems*, vol. 6, no. 2, pp. 73-81, 2017.

[3] IDC Corporate, „IDC FutureScape: Worldwide Data and Analytics 2021 Predictions," IDC Corporate USA, 2020.

[4] S. A. Salloum, C. Mhamdi, M. Al-Emran und K. Shaalan, „Analysis and Classification of Arabic Newspapers' Facebook Pages using Text Mining Techniques," *International Journal of Information Technology and Language Studies (IJITLS)*, Bd. 1, Nr. 2, pp. 8-17, 2017.

[5] A. Al-Rawi, „Assessing public sentiments and news preferences on Al Jazeera and Al Arabiya," *International Communication Gazette*, Bd. 79, Nr. 1, pp. 1-19, 2016.

[6] N. Kim, H. Lee, W. Kim, H. Lee und J. H. Suh, „Dynamic patterns of industry convergence: Evidence from a large amount of unstructured data," *Volume 44, Issue 9*, pp. 1734-1748, 2015.

[7] Q. Al-Radaideh, „Applications of Mining Arabic Text: A Review," in *Recent Trends in Computational Intelligence*, Intechopen.com, 2020.

[8] A. Farghaly und K. Shaalan, „Arabic Natural Language Processing: Challenges and Solutions," *ACM Transactions on Asian Language Information Processing*, 2009.

[9] S. A. Salloum, A. Q. AlHamad, M. Al-Emran und K. Shaalan, „A Survey of Arabic Text Mining," in *Intelligent Natural Language Processing: Trends and Applications*, Springer, Cham, 2018.

[10] H. K. Aldayel und A. M. Azmi, „Arabic tweets sentiment analysis - a hybrid scheme," *Journal of Information Science*, Bd. 42, Nr. 6, 2016.

[11] „A New Modeling Approach for Arabic Opinion Mining Recognition," in *Intelligent Systems and Computer Vision (ISCV), 25-26 March 2015*, 2015.

[12] M. A. Elmasry, T. Soliman und A.-R. Hedar, „Sentiment Analysis of Arabic Slang Comments on Facebook," *International Journal of Computers & Technology*, Bd. 12, Nr. 5, 2014.

[13] F. H. Mahyoub, M. A. Siddiqui und M. Y. Dahab, „Building an Arabic Sentiment Lexicon Using Semi-supervised Learning," *Journal of King Saud University – Computer and Information Sciences*, Bd. 26, Nr. 4, p. 417–424, 2014.

[14] A.-B. Sharaf und E. Atwell, „Knowledge Representation of the Quran through Frame Semantics: a Corpus-based Approach," in *The Fifth Corpus Linguistics Conference*, Liverpool: UK, 2009.

[15] F. Harrag, „Text mining approach for knowledge extraction in Sahîh Al-Bukhari," *Computers in Human Behavior*, Nr. 30, pp. 558-566, 2014.

[16] tartarus.org, „The Porter Stemming Algorithm," Jan 2006. [Online]. Available: https://tartarus.org/martin/PorterStemmer/. [Zugriff am 29 January 2021].

[17] J. Wijffels, „UDPipe Natural Language Processing - Text Annotation," 10 Dec 2020. [Online]. Available: https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html. [Zugriff am 3 March 2021].

[18] K. Toutanova, D. Klein, C. Manning und Y. Singer, „Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network," in *HLT-NAACL*, 2003.

[19] A. Ratnaparkhi, „A Maximum Entropy Model for Part-Of-Speech Tagging," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, USA, 1996.

ACKNOWLEDGEMENT

I gratefully thank Felesteen newspaper for providing the dataset and full cooperation in data analysis and real testing of the author prediction model. I hope this model could help it to ensure the authenticity of the article concerning the original author and verify its objective content.

Dr. Naji Shukri Alzaza *obtained his BSc degree in Computer/Mathematics in 2000 from the Islamic University of Gaza (IUG), Palestine, and MSc and PhD degrees in Information Technology (IT) in 2007 from the University Utara Malaysia (UUM), Malaysia. He is an Oracle Certified Professional since 2004. He got a Golden Medal from Malaysia Technology Expo (MTE2010) and a Silver Medal from MTE2008. His research interests include mobile learning, mobile commerce, virtual reality, data mining, and information security. Author of several books in IT, community and cultural issues. A political and social community analyst. In 2013, he was the Dean of the Faculty of Information Technology, University of Palestine (UP). In 2014, he was the Dean of the Community Service and Continuing Education at UP. Currently, he is an Associate Professor of Mobile Technology.*