Authorship Authentication System for Arabic Texts Using Machine Learning Algorithms

A Thesis Submitted to Faculty of Engineering, Fayoum University
In Partial Fulfillment of the Requirements for Degree of
Master of Engineering Science (M.Sc.)

Electrical Engineering Department
(Computer Engineering and Systems)

Submitted by

Eng. Mai Shaaban Ahmed Mohamed Barani

Supervised by

Prof. Dr. Rania Ahmed Abdel Azim
Professor of digital signals, Electrical Engineering Department
Faculty of Engineering, Fayoum University
(Main Supervisor)

Dr. Ahmed Salama Ismail
Lecturer at Information Systems Department
Faculty of Computer Science and Information Systems - Fayoum University
(Co-Supervisor)

Faculty of Engineering, Fayoum University
Fayoum, Egypt
April, 2024

ABSTRACT

Owing to the world's fast-growing internet usage, authorship authentication today plays a critical role in identifying the true identity of an author or the origin of a written work. It is the method through which a linguist tries to figure out who wrote an anonymous work based on the language used and the writer's linguistic style. There has been considerable research effort into trying to solve this problem. Initially these efforts were based on statistical patterns, and more recently they have centered on a range of techniques from artificial intelligence. An important early breakthrough was achieved by Mosteller and Wallace in 1964 [1], who pioneered the use of 'function words' – typically pronouns, conjunctions and prepositions – as the features on which to base the discovery of patterns of usage relevant to specific authors.

The main focus was studying English, Spanish, and German. Because of the difficulty and length of Arabic sentences, academics have paid less attention to the language. Over the past two decades, Arabic texts on the World Wide Web have significantly expanded, making it the fourth most utilized language. Hence, there is a high demand for efficient Arabic text classification.

Text classification is a big domain area in applied artificial intelligence. Using machine learning, recent advanced technology, classification task can be automated that leads to superfast and efficient processing. In text classification tasks, preprocessing of the dataset and features selection process are very important to achieve the best result. A poor classifier fed with meaningful features may perform better than a robust classifier fed with low-quality features.

In this thesis, authorship authentication problem is discussed concerning many Arabic concerns using a dataset consisted of ten different writers (Alfarabi, Alghazali, Aljahedh, Almasoody, Almeqrezi, Altabary, Altowhedy, Ibnaljawzy, Ibnrshd, and Ibnsena). Each writer has 60 text files, so the dataset contains 600 files. Our approach structure starts by preprocessing of the dataset which includes stopwords removal, normalization, tokenization and stemming to study the effect of preprocessing on the performance of different classifiers. We tested different features extraction techniques such as Bag of Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF) besides features of the structure, lexicon, syntax, semantics, and writing style to study their effect on Arabic authorship authentication. The process that follows features extraction is features selection,

also known as attributes selection. It seeks to extract a subset of the most essential features from the sparse feature space without impairing the classifier's performance. Then, the selected feature vectors are evaluated using several classification methods. we investigated the performance of ten classification algorithms, including Support Vector Classifier (SVC), Random Forest (RF), Logistic Regression (LR), Multinomial Naive Bayesian (MNB), k-Nearest Neighbors (KNN), Stochastic Gradient Descent (SGD), Decision Trees (DT), Bagging Classifier (BC), Gradient Boosting Classifier (GBC), and Multilayer Perceptron (MLP). We also describe and evaluate a hybrid approach that combines classifiers using stacking classifier to improve overall predictive performance effectively. The results show that the stacking model with LR as the meta-classifier with two base classifiers has achieved an accuracy of 96%, while the stacking model with LR as the meta-classifier with three base classifiers has achieved an accuracy of 96.67%. In addition, the stacking model with RF as the meta-classifier with two base classifiers has achieved an accuracy of 97%, while the stacking model with RF as the meta-classifier with three base classifiers has achieved an accuracy of 98.33%.

Chi-squared feature selection was also employed to enhance the effectiveness of the selected features and improve the stacking model's performance.