جامعة بيرزيت
BIRZEIT UNIVERSITY

Authorship Attribution of Arabic Articles

By:
Maha Hajja
Ahmad Yahya
Adnan Yahya

Presented at ICALP'19 Conference, Nancy, France
October 16-17, 2019

Outline

- Motivation
- Problem Statement & Applications
- Dataset
- Feature Extraction
- Feature Selection
- Experiments & Results
- Conclusions & Future Work

1 Motivation

Motivation

- The web content is growing very fast
  - Many articles are posted anonymously with the different social media platforms and blog websites
- This resulted in articles, blogs, essays and emails being published under assumed identities or have no known author
- Copyright and other legal issues like plagiarism may occur

2

Problem Statement & Applications

What is Authorship Attribution

- A sub-task of the text classification (TC) paradigm
- Authorship Attribution deals with identifying the author of an anonymous text.
- By attributing each test text of unknown authorship to one of a set of known authors, whose training texts are given.

Authorship Attribution Application

- Plagiarism detection
  - (for example: College Essays)
- Identifying writers for inappropriate documents and texts that were sent anonymously
  - (for example: dangerous or slanderous e-mails)
- Solving copyright issues
  - Determining the source of anonymous posts in blogs
  - Resolving problems of unclear authorship for important historical documents.

AAA System Stages

3 Dataset

Dataset

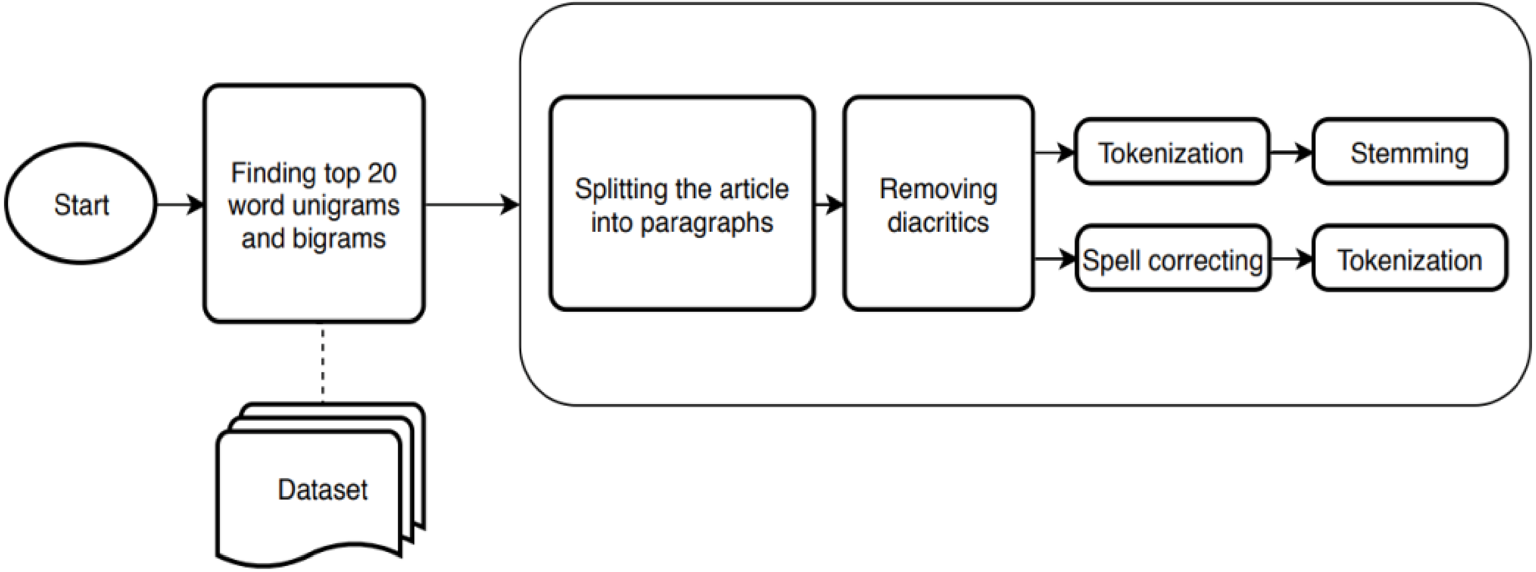
- Proper dataset for Arabic articles authorship attribution was not found
- A Dataset was manually collected
  - 7 authors
  - 10 articles each
- All the articles were collected from the website blogs.aljazeera.net except for one author
- Homogenous articles, hence writing style features will be addressed and emphasized
- For the purpose of having larger data, another dataset with the same properties was combined with ours through the experiments
- All texts are MSA
- For each article we created a metadata file to contain items such as author’s name, class index for author, title of article, article link, size, date of publication and language
- The dataset and its expansions and metadata are being made available for other researchers

4 Feature Extraction

Feature Extraction

- The Feature Extraction included the following sub-stages:
  - Preprocessing
  - Feature Types Identification
  - Feature Extraction

Preprocessing



Feature Types

- Style Features
  - Lexical Features
  - Syntactic Features
- PoS Features
- Content Features

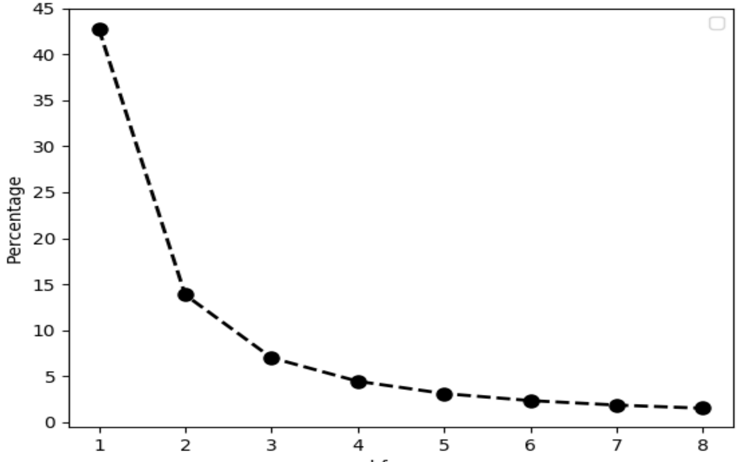
Style Features-Lexical Features

- Features that represent statistics about the text
- Proved to have good results in the literature
- Readability score measures the complexity of the text
- $Readability\ Score = \frac{\sum_{i}^{N} rank(token(i))}{N}$
- The rank of a word depends on it's usage(frequency)

Lexical Features	Average word length
	Average sentence length
	Percentage of short words
	Percentage of hapax-legomena
	Percentage of numbers
	Percentage of typos
	Percentage of diacritics
	Type to token ratio
	Nuraihan readability score

Style Features-Lexical Features – Cont.

- The frequencies of the words were collected from Al Jazeera documents
- The distribution of the frequencies proved to have a Zipfian distribution



Word Rank	Word	Frequency
1	في	3671564
2	من	2330186
3	أن	1534168
4	على	1520368
5	إلى	1103242
6	عن	641711
7	التي	597724
8	إن	463476
9	مع	460636
10	ما	442644

Style Features- Syntactic Features

- The Syntactic features were split into two categories:
  - Percentage of the punctuations in the text
  - The function words are the words used to connect two parts of a sentence

Function Type	Function Words Used
Conditional function words (أدوات الشرط)	ان، من، ما، متى، اين، أينما، لو، لولا، ما
Accusative function words (أدوات النصب)	لن، حتى، ان، كي، اللام، لام الجحود، الفاء
Questioning function words (أدوات الاستفهام)	من، ما، متى، اين، كيف، كم، لماذا، هل
Simile function words (أدوات التشبيه)	الكاف، كأن
Preposition and postposition function words (أدوات الجر)	من، الى، على، في، عن، حتى، رب، الباء، الكاف، اللام، الواو، التاء، مذ، منذ

Part of Speech (PoS) Features

- PoS Features can strongly determine the writing style of an author
- Different PoS features were extracted from FARASA PoS tagger

PoS Code	PoS Description
NSUFF	Noun Suffix
PRON	Pronoun
ADJ	Adjective
NUM	Number
PREP	Preposition
CASE	alef of tanween fatha
DET	determiner
ADV	Adverb
PART	Particles
V	Verb
CONJ	Conjunction
NOUN	Noun
PUNC	Punctuation

Content Features

- Content features are the features that deal with the content of the text itself:
  - The frequency of the top unigram/bigram words
  - The percentage of positive, negative and neutral words used

Content Features	Frequency of top 20 unigrams
	Frequency of top 20 bigrams
	Percentage of positive words
	Percentage of negative words
	Percentage of neutral words

Extraction Methodology

- All the features presented need robust tools and prior knowledge of frequencies in large dataset. Therefore the following were used:
  - pre-collected set of unigrams frequencies on Al Jazeera documents from FARASA
  - FARASA PoS tagger
  - The sentiment analyzer from ArabicTools - Ali Salhi

5 Features Selection

Features Selection

- Some features may be less informative and decrease the model’s accuracy. Therefore, the a subset from the features was taken using:
  - **Statistical Approach (Information Gain)**
  - **Search Approach (Greedy Search)**

Features Selection – Statistical Approach

- The Information Gain (IG) was calculated for all the features
- The PoS percentages features have relatively large IG compared to the other features
- Token to term ratio, punctuation percentage and word length also had a high IG value

Feature	Information Gain
Determiner (PoS, ال التعريف)	1.000582
Type to token ratio (TTR)	0.990532
Percentage of punctuation	0.937544
Average word length	0.886882
Adverb (PoS)	0.847305
Percentage of short words	0.75187
Adjective (PoS)	0.728352
Pronoun (PoS)	0.707754
Average sentence length	0.654821
NOUN (PoS)	0.631447
Unigrams (average IG value)	0.6155228
VERB (PoS)	0.583294
Nuraihan readability score	0.583221
Particles (PoS)	0.569106
Noun suffix (PoS)	0.480477
Neutral words percentage	0.467596
Percentage of Hepax-Legomena	0.463519
Conjunction (PoS)	0.42083
Bigrams (average IG value)	0.319848

Features Selection – Search Approach

- Finding the best subset of features is an exhaustive and an NP-hard problem
- The greedy search was chosen for its low time complexity
- Many important features like the function words frequencies and percentage of diacritics were discarded because of the large number of authors to choose from.

Feature Type	Feature
Style Features	Average word length
	Average sentence length
	Percentage of punctuation
	Percentage of short words
	Percentage of hapax-legomena
	Percentage of typos
	Type to token ratio
	Nuraihan readability score
PoS Features	PoS percentages for top PoS
Content Features	Frequency of top 20 unigrams
	Frequency of top 20 bigrams
	Percentage of neutral words

6 Experiments and Results

Evaluating classifiers

- 10-fold cross validation
- Pairs of 2 authors
- Macro measurements were used*

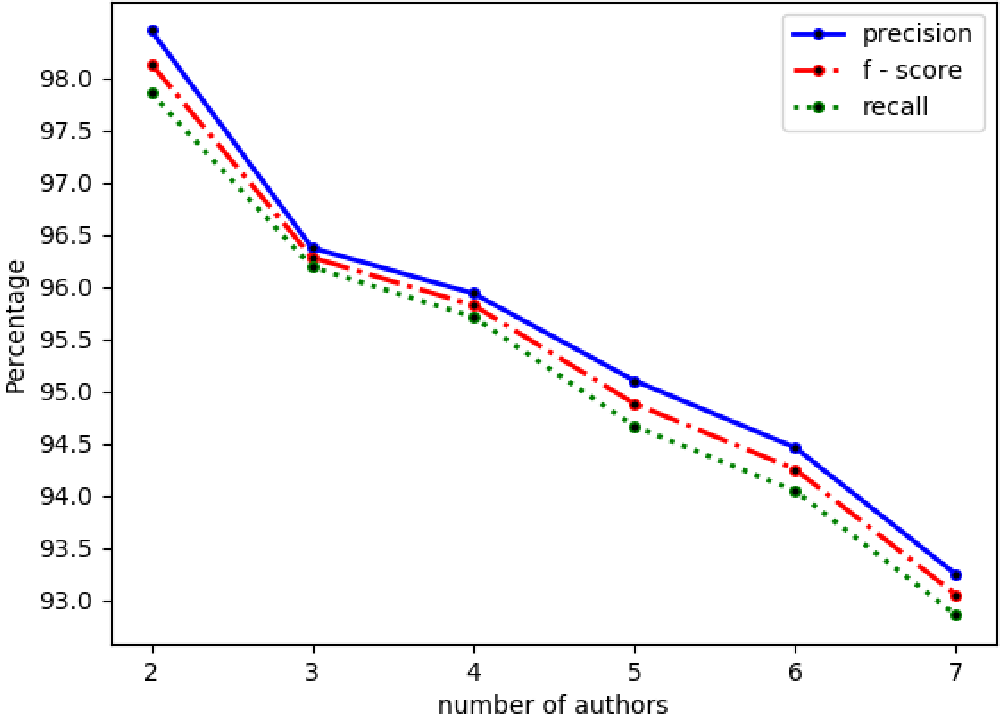
Classifier	Macro Precision	Macro Recall	Macro F-Score
SVM	98.24%	98.10%	98.17%
Decision Tree	84.97%	84.52%	84.75%
Naive Bayes	97.97%	97.61%	97.79%

*Macro measurement: taking the average over different sets.

e.g. $Macro\ Presicion = \frac{P_1 + P_2 + \ldots + P_N}{N}$

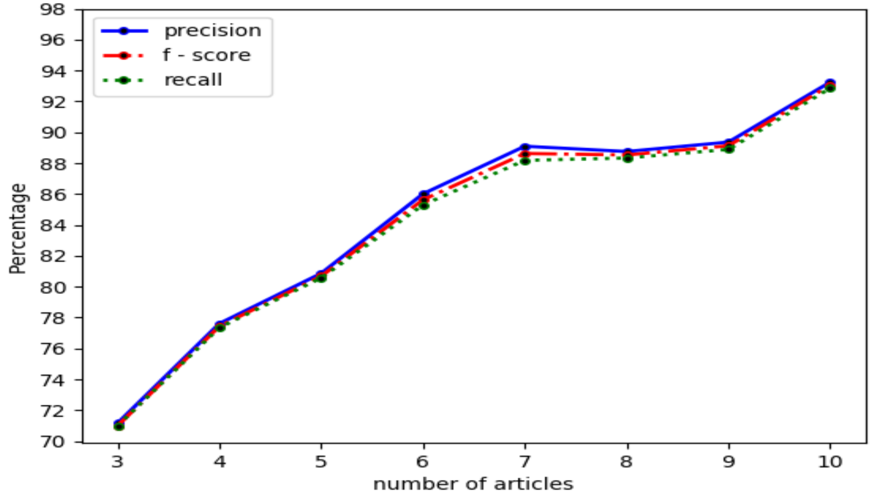
Performance vs number of authors

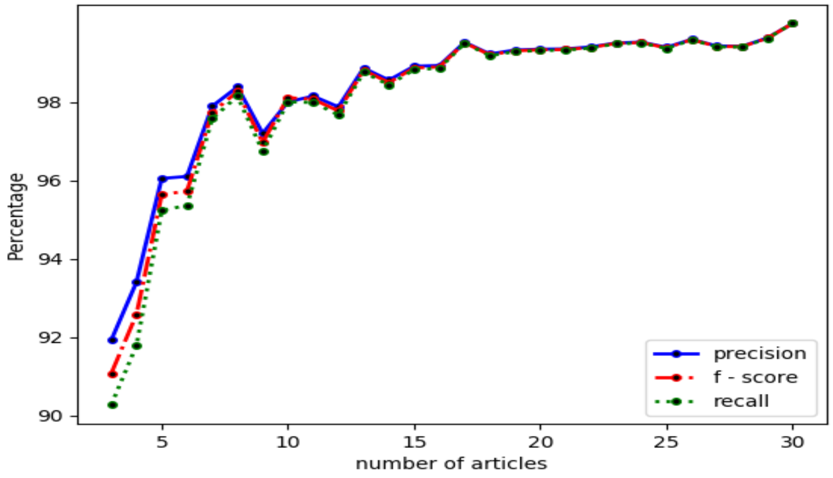
- The SVM proved to have the best results, hence was chosen for this experiment and the remaining ones
- Taking the subset of k authors
- Starting from k=2 to k=7
- All possible combinations were evaluated using 10-fold cross validation
- The experiment was combined with another dataset to have a total of 16 authors.
- The metrics still went down as the number of authors increases, but remained above 93%.



Performance vs number of training articles

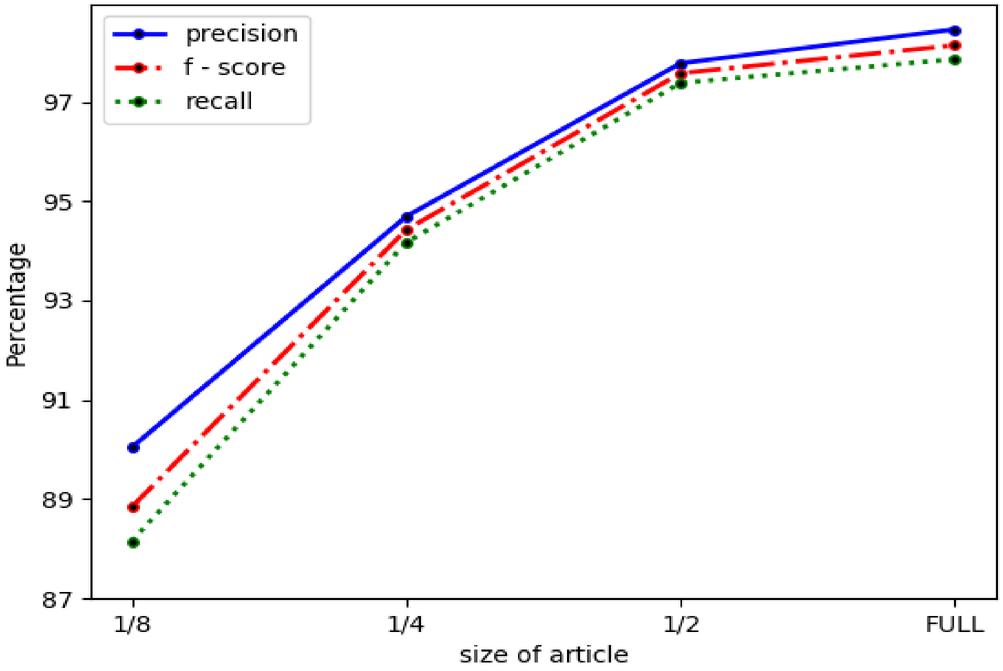
- Each of the values n = 3, 4, …, 10 articles were tested
- All the article combinations from a given n were evaluated and averaged
- Subset of 6 authors with 30 articles were tested and it showed a convergence when the number of articles reached 12–14





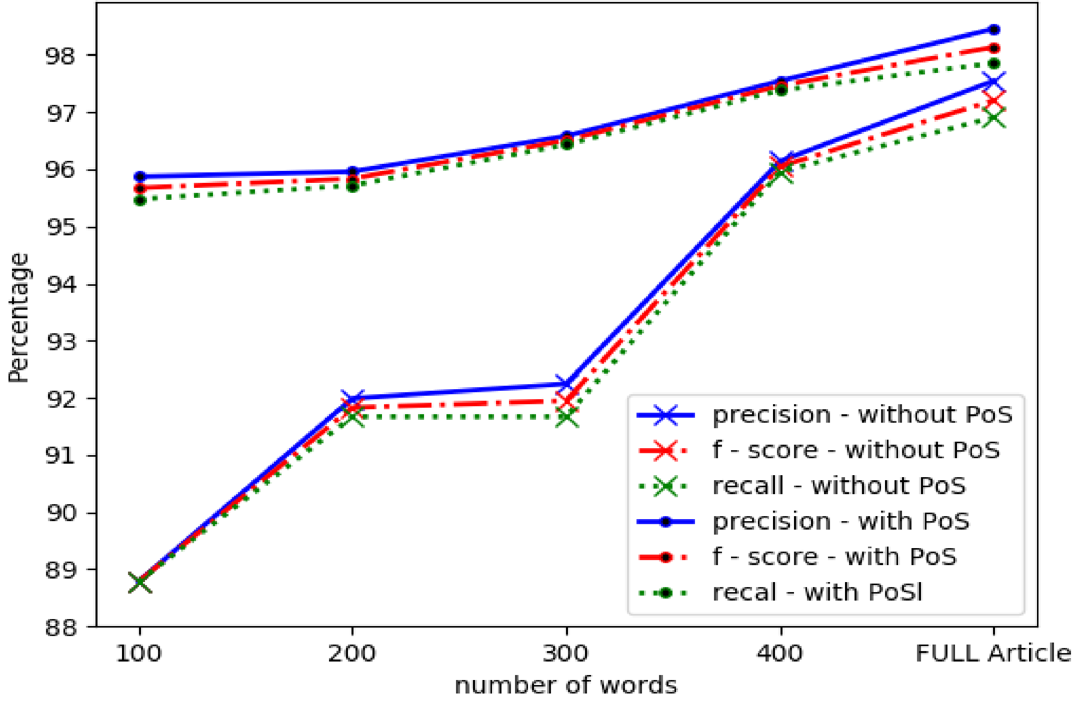
Variating the size of taken context – continuous chunks

- the first and second halves of each article were taken separately and used as the training set then evaluated and averaged
- Same done with the four quarters
- Then the eight eighths of each article
- Using all the possible subsets from pairs of authors (7 choose 2), then averaging the results for each pair



Variating the size of taken context – Random bag of words

- bigrams frequencies feature was not used
- PoS features were pre-calculated
- The experiment was done with and without PoS features for the pairs of 2 authors (k = 2), for a different number of randomly selected words
- continuous chunks were not as negative as choosing random words because features (like word bigrams and POS tags) stayed alive and meaningful
- significant improvement when POS tags were included



7

Conclusions and Future Work

Conclusions

- Reducing the number of articles affects the results negatively
- Increasing the number of articles increases the accuracy to a certain point "convergence threshold"
- reducing the number of authors for a classifier to choose from resulted in better results
- the continuity of the text preserves a lot of useful features that is lost in randomized word selection which resulted in worse performance.
- The more words for the random tokens the better, with improvements when they had their POS tags as additional features

Future Work

- Testing on a topic-specific dataset
- Testing the results with short chuck text (small context)
  - Tweets
  - Facebook Posts
- Trying to identify other attributes than the author name itself
  - Gender
  - Age
  - Interests
- Trying to include metadata in the training process, i.e. The available profiling information in case of Twitter or Facebook
  - Location, The timestamp of posting, Liked pages, etc..

Thanks!

Any ***questions*** *?*