Examining n-grams and Multinomial Naïve Bayes Classifier for Identifying the Author of the Text “Epistle to the Hebrews”

Panagiotis Satos[1][a] and Chrysostomos Stylios[1,2][b]
[1]*Department of Informatics and Telecommunications, University of Ioannina, Arta, Greece*
[2]*Industrial Systems Institute, Athena Research Center, Patras, Greece*

Keywords: n-grams, Multinomial Naïve Bayes Classifier, Text Mining, Machine Learning, Natural Language Processing, Theological Texts, Identifying Authors.

Abstract: This work proposes a methodology consisting of splitting and pre-processing of Koine Greek dialect texts, examining word n-grams, character n-grams, multiple-length grams, and then suggests the best value for n of n-grams. The Multinomial Naïve Bayes Classifier is used along with the n-grams to identify the author of the text “Epistle to the Hebrews” between Paul and Luke, who are considered the most likely authors of this Epistle. In order to create a balanced dataset, the texts of Apostle Paul’s Epistles and the book “Acts of the Apostles” by Luke the Evangelist are used. This work aims to identify the author of the text “Epistle to the Hebrews” and reply to the theological question about its paternity.

1 INTRODUCTION

Text is often considered unstructured data, but extracting information from it is imperative, even if success is at a low level. We assume that information remains “hidden” in “buried” text data. Machine Learning (ML) techniques are used for Text Mining. Once words and phrases are converted to numeric values, the information is extracted using machine learning techniques.

Text data is considered either strings or, more commonly, a bag-of-words (BoW). The BoW approach ignores word order and text structure and does not consider the meaning of words so that essential information can be lost. Each term (word) is a vector or a unique point in the vector space. Sometimes, it is desirable to approach the texts semantically to be analyzed on a higher level, and more critical information will be extracted. Based on the latter, Named-Entity Recognition (NER) and the relationships of entities may discover more interesting patterns than the BoW approach. The BoW approach is widely used, and it is simpler in algorithmic terms than the alphanumeric (string-based) approach of texts. However, the effectiveness of a method used to identify (verify) an author depends on the language of the text. One method may have high accuracy in one language (e.g., English) and at the same time low in another (e.g., Greek).

In this paper, we study the Authorship Attribution Problem from prospective authors in theological texts. Usually, many potential authors create a complex problem, while the fact that all texts are theological makes it simpler. When the texts of well-known authors and those of unknown authors (or authors) do not belong to the same type of text, it is challenging to create a training corpus.

Many text classification applications are based on Naïve Bayes models. Here, we selected the Naïve Bayes Classifier because the number of the Epistles of the New Testament is small, and the NB classifier does not require a large amount of training data to achieve maximum accuracy. The most common NB models are two: the Bernoulli (or Binary) Naïve Bayes Model and the Multinomial Naïve Bayes Model. In the first (Bernoulli), data follow a multivariate Bernoulli distribution and are used when each feature is binary (presence or absence of terms in a document), ignoring their frequency. In the second (Multinomial), data follow a multinomial distribution, and the count of terms is significant (every count is a feature). Multinomial Naïve Bayes

[a] https://orcid.org/0000-0001-6099-8894
[b] https://orcid.org/0000-0002-2888-6515

Satos, P. and Stylios, C.
Examining n-grams and Multinomial Naïve Bayes Classifier for Identifying the Author of the Text “Epistle to the Hebrews”.
DOI: 10.5220/0010970300003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 1*, pages 447-458
ISBN: 978-989-758-547-0; ISSN: 2184-433X

Model is considered the basic technique in text classification and is the most widely used.

We investigate text mining methods in theological texts to identify the author. It is a fundamental process since a theological text may (or may not) be accepted, depending on who wrote it. For example, the Orthodox Church officially accepts a theological text written by a divinely inspired author (Apostle, Apostolic Father, Apologist). On the contrary, the exact text by an unknown author probably would not be accepted because it could be considered as written by an atheist or a heretical author.

The management of the Ancient Greek language is crucial. One reason is that it is common for a sentence in the Ancient Greek language to present multiple valid syntactic interpretations. The theological texts of the New Testament, which are used in the present research, are written in the Ancient Greek language and specifically in the Koine Greek (Biblical Greek or Alexandrian) dialect.

The texts of the New Testament are of particular interest, especially the text "Epistle to the Hebrews". Although the author of this Epistle is considered Apostle Paul, most modern theologians question that. A Multinomial Naïve Bayes Classifier is tested in this work using n-grams. The classifier has attempted to answer whether the text "Epistle to the Hebrews" is more similar to Apostle Paul's Epistles or to the book "Acts of the Apostles" of Luke the Evangelist. Theologists have been debating for years about the author of the text "Epistle to the Hebrews" either Apostle Paul or Luke the Evangelist.

Section 2 presents related works and a review of theological texts. Section 3 focuses on the authorship attribution problem and the stylometric characteristics. In Section 4, we describe the proposed methodology. In Section 5, we present our test results, while in Section 6, the conclusions are summarized, and future work directions are indicated.

2 THEOLOGICAL TEXTS & RELATED WORK

We study and use the Apostle Paul's Epistles and the "Acts of the Apostles" of Luke the Evangelist's book. In the text "Epistle to the Hebrews" there is the spirit of Apostle Paul's teaching, but it differs from his other epistles, mainly in style. It is written in a higher common language, more like a treatise than an epistle. For this reason, from the 2nd century, some theologists consider that the meanings are Paul's, but the writing probably belongs to somebody else. Probably one of Paul's followers (Luke the Evangelist, Saint Apollos, Saint Clement of Rome), who on the one hand memorized Paul's teachings, while on the other wrote them in free writing. The inspiration and apostolicity of the Epistle have never been questioned. The hypothesis that someone else may have written the letter on behalf of Paul is also based on the verse of "Epistle to the Romans": *"I Tertius, who wrote this epistle, salute you in the Lord."* (Rom. 16, 22), where it is stated that Tertius wrote Paul's epistle.

Some theologists concluded that Apostle Paul wrote the Epistle, but with the help of Luke the Evangelist. Luke the Evangelist probably undertook its formulation and final drafting. He wrote it on behalf of Paul, but using his summaries and Paul gave the final approval of the Epistle.

In addition to text "Epistle to the Hebrews", many Protestant and Roman Catholic researchers believe that there are six (6) other epistles probably written by Paul's disciples and not by him. So, in some cases, only half (seven) epistles have not been disputed, while for others, only the first three epistles.

M. Ebrahimpour et al. used SVM and Multiple Discriminant Analysis (MDA) to develop two automated author performance schemes that tested to the text "Epistle to the Hebrews". The texts they used were written not in Greek but in Latin characters (they changed each Greek character to the corresponding Latin, i.e. Greeklish), while removing all the characters of the texts (and the punctuation marks, since the original text was written in capital letters, without accents and punctuation) except for lowercase letters (a-z) and spaces. Texts by eight (8) authors were used for comparison. Their results showed that the text "Epistle to the Hebrews" is closer to Paul, but it is further away from the rest of his letters, showing that it is not included in his writing style. Luke appears as the second most probable author. In the case of SVMs with an optimized polynomial kernel, the letter is attributed to Luke. There may be a basis for an earlier statement that the text "Epistle to the Hebrews" was initially written by Paul in Hebrew and translated into Greek by Luke, or someone else is the Epistle's author.

A. Kenny [used 99 criteria (use of conjunctions, particles and prepositions; the cases of nouns, pronouns and adjectives; the moods, tenses and voices of verbs)] attributed to Paul all the Epistles except the text "Epistle to the Hebrews" and the "Epistle to Titus".

D.L. Mealand [used Multivariate Approach (Samples of 1000 words, 25 stylistic variables were analysed & 19 of these were used) & cluster analysis

& discriminant analysis] concluded that the “Epistle to the Colossians” and the “Epistle to the Ephesians” were probably not Paul’s texts.

G. Ledger [Multivariate Statistical Analysis (1000 word sections & orthographic variables are measured)] argued that the “Epistle to the Galatians” and “First Epistle to the Thessalonians” are doubtful.

T. Putnins et al. [used word recurrence interval-based method & trigram Markov method & the third method extracts stylometric measures (such as the frequency of words)] consider the text “Epistle to the Hebrews” and another 11 books as texts with questionable authors, while in addition, they removed the “Epistle to Philemon” and the “Epistle of Jude”, due to their small size, leaving 13 books as a training set. Of the 13 texts consisting of more than 6,000 words, they were divided into texts of about 3,000 words, thus achieving a more extensive training set for 37 texts. Their research showed that the author of the text “Epistle to the Hebrews” could not have been Paul the Apostle, Luke the Evangelist, Mark the Evangelist, Matthew the Apostle and Evangelist or John the Apostle and Evangelist (at a rate of > 99.1%). Saint Barnabas is presented as the most probable author. The text “Epistle of Barnabas” was added to the Dataset, although the latter is considered an occult text and is therefore excluded from the Biblical Canon. In other words, the real author of the “Epistle of Barnabas” was probably not the Saint Barnabas the Apostle but someone posterior who, using the name of Barnabas, sought to gain prestige in his text, making it acceptable to the Church.

M. Koppel et al. presented a learning-based method, measuring the actual “depth of difference” between two collections, which brought high accuracy, being independent of language, period, or type of texts in their examples. They examined how to use negative examples (information) correctly when used in verification problems. Using a little negative data increases the accuracy. They also state that in cases where an author uses a small number of features, he will use it consistently on his works. The subject matter can influence diversity, the type or the purpose of the works, by a chronological stylistic shift or even intentionally for covering the author’s identity.

R. Avros et al. performed two experiments; they applied clustering to various texts by well-known authors to test their algorithm’s performance. Then, using the algorithm, they tried to confirm or not a specific author’s writing of a text. The texts were divided into equal parts (10KB), and using the Bag-of-Words method, a set of vectors represented each book [spectral clustering & Ng-Jordan-Weiss (NJW) algorithm & BoW]. Having proved that the three (3) clusters are the optimal number, he applied the algorithm (the “Epistle to Philemon” is absent, probably due to its small length). The text “Epistle to the Hebrews” was in the same cluster with the “First Epistle to Timothy”, the “Second Epistle to Timothy” and the “Epistle to Titus”.

M. Koppel and S. Seidman [used unsupervised technique & identify textual outliers, novel similarity measures (second-order document similarity measures taken from the authorship verification literature & identify outlier vectors, Hodge and Austin (2004) and Chandola et al. (2009))] considered 7 of the 13 epistles as Paul’s texts, the 4 as epistles that do not belong to Paul and the other 2 epistles as disputed texts. They conducted text comparison experiments using the above Epistles, the texts of the Gospel (of the four Evangelists), and the other epistles of the New Testament.

D. Shalymov et al. [comparison of the randomness of two given texts (incorporation of the Friedman-Rafsky two-sample test into a multistage procedure, n-grams)] compared six of the seven undisputed epistles, while the text “Epistle to the Hebrews” proved to be dissimilar to the above six epistles.

J. Savoy states that the epistles attributed to Paul range from 4 to 13, with the majority agreeing at 7. About the authorship attribution problem for Paul’s epistles, Savoy considers two methods [used hierarchical clustering (Burrows’ Delta & Labbé’s intertextual distance) & verification method (based on the impostors’ strategy)].

3 AUTHORSHIP ATTRIBUTION PROBLEM & STYLOMETRIC CHARACTERISTICS

Author attribution is the process of identifying the creator (or creators) for texts by an unknown or a controversial author. It is a tricky problem because the (known and unknown) documents may come from different fields, and there is a possibility that the unknown author may not be on the list of potential authors.

The Authorship Attribution Problem has been approached with the method of stylometry. “Stylometry” is the use of Natural Language Processing (NLP) methods for detecting the writing style or statistical analysis of the writing style of the author (e.g., syntactic and semantic features, length of sentences and paragraphs, frequency of specific words). The (stylometric) features of the text, such as style, dialect, and writing period, impact the solution of the above problem. However, there are restrictions, especially on electronic texts (e.g., character limit on social media texts). In conclusion, authorship is justified because an author always uses a characteristic vocabulary (a small vocabulary is sufficient for high performance).

A vector is created from each stylometry variable, each dimension corresponding to a different feature. The size of the texts in selecting variables is a significant factor. Results of experiments have shown that separating authors based on their texts' characteristics is more effective when analyzing extensive texts. On the other hand, separating writers by analyzing small texts is more complicated.

Author recognition is a closed-set classification or open-set classification. In the first one, the training set includes samples from all possible authors so that the unknown text is assigned to one of them. However, it is not possible to know all of the candidate authors in advance, so it is possible that the unknown text may not belong to one of the candidate authors who are included in the training set.

Generally, we can divide the author's problem into the following three variants:

- There are many potential authors, and we need to attribute the unknown text to one of them (needle-in-a-haystack problem)
- There is a potential author, and we need to determine if he has (or has not) written the unknown text (verification problem)
- There are no potential authors, and the aim is to extract as much information as possible about the unknown author (profiling problem)

In their research on Author Identification, M. Kocher and J. Savoy suggested the use of the 200 most frequently used terms (including punctuation marks) as features [unsupervised authorship verification model called SPATIUM-L1, using the 200 most frequent terms of the disputed text (isolated words and punctuation symbols)] or removing terms that appear once or twice and limited vocabulary to the 500 most common words [difference between two texts: the L1 norm (e.g. Manhattan, Tanimoto), the L2 norm (e.g. Matusita), the inner product (e.g. Cosine) or the entropy paradigm (e.g. Jeffrey divergence) & high precision (HPrec), (characters, punctuation symbols, or letter n-grams as well as words, lemmas, Part-Of-Speech (POS) tags, and sequences of them)]. However, a writer's writing style does not remain unchanged, as it is influenced by age and other factors such as medical ones.

In most cases of author attribution problems, there is a set of candidate authors, a set of training corpora, and a set of sample texts (or just a text) of an unknown author (test corpus), which must be attributed to one of the candidate authors. Usually, we approach it:

- Through each author's profile, which is extracted from his texts, ignoring the differences between his texts. Common N-Grams (CNGs) are commonly used, which are independent of the language (representing each document as a bag of character n-grams, where the optimal value for "n" depends on the body and language of the texts)
- By approaching the known texts as instance-based approaches, where ML approaches are commonly used (if there is only one training text for an author, it should be broken into smaller sections)
- By combining both approaches (hybrid approaches)

Both profile-based and instance-based approaches have advantages and disadvantages. Profile-based is more effective when there is an uneven distribution of training texts (the problem of class imbalance, that is, when there are many texts for one author and few for another) and when the texts are short. On the other hand, instance-based approaches are more accurate when there are enough training texts for all the candidate authors and when the texts are extensive in size. This paper has chosen the instance-based approach for solving the authorship attribution problem.

3.1 Types of Stylometric Characteristics

Texts are characterized by two main factors: their content and style. In practice, texts are a sequence of words, referred to as strings. The text's set of features (or dimensions) is a dictionary. Feature selection affects the performance of a Text Mining system. In order to detect the author's writing style or stylometry, (stylometric) features of different categories can be selected, such as:

- Lexical features (unique words number)
- Character features
- Syntactic features
- Semantic features
- Application-specific features (for short texts)

In this work, we use features from the first two categories. The text is regarded as a simple sequence of characters for the character Features. The usual procedure is to measure the frequencies of n-grams at the character level, where along with word frequencies, there are essential features for determining stylometric patterns. In many cases of the author’s identification, n-grams are more effective than lexical features. An important parameter is the definition of “n”, which is language-dependent. For the Modern Greek language, the most effective number for “n” is relatively large (n > 4), while for English, a smaller number is suggested. Alternatively, variable-length values are defined. However, it should be noted that for large values of n (e.g. n > 7) the number of n-grams produced becomes vast and most have a very low incidence. This paper proves that a significant value of “n” is also more effective in the Koine Greek dialect of Ancient Greek language.

Word and character n-grams (such as word frequencies, punctuation marks, average phrase length, average word length, etc.) are low-level stylometric features. The syntactic, such as synonyms and semantic dependencies, are considered high-level stylometric features. It should be noted that necessary steps have been taken for the semantic analysis of texts written in the ancient Greek language. Although n-grams are low-level features, experimentally, they have proved to be particularly effective.

By combining different types of features in authorship attribution, the dimension of the features is increased. The combination of multiple features often yields less accurate results than careful feature selection in classification problems and regression. Features selection is significant in text classification and authorship attribution due to the large dimensions and noise features. However, a combination of variables is not efficient in all cases. For example, it can be effective when classifying texts by a particular author, while the same combination may not be effective when performing the same procedure on another author’s texts. Indeed, the ideal combination of stylometric features depends, in any case, on the author.

4 PROPOSED METHODOLOGY

4.1 Dataset

The Dataset consists of the 13 epistles written by Saint Paul and the book “Acts of the Apostles” written by Luke the Evangelist. The 13 Epistles of Paul include 87 chapters with 32,851 words. The book “Acts of the Apostles” consists of 28 chapters with 18,772 words.

Firstly, the Epistles of Paul were divided into 28 different texts (Table 1), and each chapter of the book “Acts of the Apostles” was divided into separate texts (Table 2). These 56 texts were our Dataset (first case).

Secondly, the Epistles of Paul were divided into 14 separate texts (Table 3), and the chapters of the book “Acts of the Apostles” were divided into 14 separate texts (Table 4), so that the Dataset consists of 28 texts, with texts of equal size, as much as possible.

Table 1: Separation of Paul’s Epistles (first case, 28 texts).

“Apostle Paul’s Epistles” Chapters	Words
Epistle of Paul to the Ephesians1-3	1101
Epistle of Paul to the Ephesians4-6	1363
Epistle of Paul to Philemon1	340
Epistle of Paul to the Philippians1-2	936
Epistle of Paul to the Philippians3-4	707
Epistle of Paul to the Galatians1-3	1213
Epistle of Paul to the Galatians4-6	1040
Epistle of Paul to the Colossians1-2	958
Epistle of Paul to the Colossians3-4	663
First Epistle of Paul to the Corinthians1-4	1489
First Epistle of Paul to the Corinthians5-8	1495
First Epistle of Paul to the Corinthians9-12	1950
First Epistle of Paul to the Corinthians13-16	1993
Second Epistle of Paul to the Corinthians1-4	1395
Second Epistle of Paul to the Corinthians5-9	1646
Second Epistle of Paul to the Corinthians1-13	1471
Epistle of Paul to the Romans1-4	1840
Epistle of Paul to the Romans5-8	1936
Epistle of Paul to the Romans9-12	1782
Epistle of Paul to the Romans13-16	1655
First Epistle of Paul to the Thessalonians1-2	616
First Epistle of Paul to the Thessalonians3-5	883
Second Epistle of Paul to the Thessalonians1-3	835
First Epistle of Paul to Timothy1-3	713
First Epistle of Paul to Timothy4-6	915
Second Epistle of Paul to Timothy1-2	678
Second Epistle of Paul to Timothy3-4	573
Epistle of Paul to Titus1-3	665
SUM	**32851**

Table 2: Separation of Paul's Epistles (second case, 14 texts).

"Apostle Paul's Epistles" Chapters	Words
Epistle of Paul to the Ephesians1-6	2464
Epistle of Paul to Philemon1-Epistle of Paul to the Philippians1-4	1983
Epistle of Paul to the Galatians1-6	2253
Epistle of Paul to the Colossians1-4	1621
First Epistle of Paul to the Corinthians1-8	2984
First Epistle of Paul to the Corinthians9-16	3943
Second Epistle of Paul to the Corinthians1-7	2340
Second Epistle of Paul to the Corinthians8-13	2172
Epistle of Paul to the Romans1-8	3776
Epistle of Paul to the Romans9-16	3437
First Epistle of Paul to the Thessalonians1-Second Epistle of Paul to the Thessalonians3	2334
First Epistle of Paul to Timothy1-6	1628
Second Epistle of Paul to Timothy1-4	1251
Epistle of Paul to Titus1-3	665
SUM	**32851**

Table 3: Separation of the chapters of Luke's "Acts of the Apostles" (first case, 28 texts).

"Acts of the Apostles" Chapters	Words
Acts of the Apostles-Chapter1	511
Acts of the Apostles-Chapter2	848
Acts of the Apostles-Chapter3	505
Acts of the Apostles-Chapter4	682
Acts of the Apostles-Chapter5	786
Acts of the Apostles-Chapter6	280
Acts of the Apostles-Chapter7	1143
Acts of the Apostles-Chapter8	714
Acts of the Apostles-Chapter9	793
Acts of the Apostles-Chapter10	859
Acts of the Apostles-Chapter11	534
Acts of the Apostles-Chapter12	495
Acts of the Apostles-Chapter13	953
Acts of the Apostles-Chapter14	479
Acts of the Apostles-Chapter15	723
Acts of the Apostles-Chapter16	724
Acts of the Apostles-Chapter17	675
Acts of the Apostles-Chapter18	528
Acts of the Apostles-Chapter19	766
Acts of the Apostles-Chapter20	694
Acts of the Apostles-Chapter21	809
Acts of the Apostles-Chapter22	586
Acts of the Apostles-Chapter23	678
Acts of the Apostles-Chapter24	497
Acts of the Apostles-Chapter25	538
Acts of the Apostles-Chapter26	598
Acts of the Apostles-Chapter27	753
Acts of the Apostles-Chapter28	621
SUM	**18772**

Table 4: Separation of the chapters of Luke's "Acts of the Apostles" (second case, 14 texts).

"Acts of the Apostles" Chapters	Words
Acts of the Apostles-Chapters1-2	1359
Acts of the Apostles-Chapters3-4	1187
Acts of the Apostles-Chapters5-6	1066
Acts of the Apostles-Chapters7-8	1857
Acts of the Apostles-Chapters9-10	1652
Acts of the Apostles-Chapters11-12	1029
Acts of the Apostles-Chapters13-14	1432
Acts of the Apostles-Chapters15-16	1447
Acts of the Apostles-Chapters17-18	1203
Acts of the Apostles-Chapters19-20	1460
Acts of the Apostles-Chapters21-22	1395
Acts of the Apostles-Chapters23-24	1175
Acts of the Apostles-Chapters25-26	1136
Acts of the Apostles-Chapters27-28	1374
SUM	**18772**

We divided the Dataset into training set and test set, via "model_selection.train_test_split[1]" method of "scikit-learn" (sklearn) machine-learning python library (train_size = 0.75 and shuffle=True).

In the first case, out of the fifty-six (56) texts, 75% represented the training set (42 texts), and the remaining 25% accounted for the test set (14 texts). In the second case, out of the fifty-six (28) texts, 75% represented the training set (21 texts), and the remaining 25% accounted for the test set (7 texts).

Based on the proposed approach and since the epistles that made up the test set was not specific but were random (shuffle = True), each case of the following experiments was repeated five (5) times, with the lowest accuracy recorded.

4.2 Proposed Methodology

During text pre-processing, words with a high frequency of occurrence (prepositions and conjunctions, articles) are usually removed. However, in author identification problems, it is often important not to remove them. In this research, we examined and compared both approaches.

At first, the pre-processing of the texts included the following actions:

- Characters' conversion to lowercase
- Stopwords removal
- Lemmatization
- Punctuation marks removal

Then, a combination of the last two was applied.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

For lemmatization we chose the CLTK [2] lemmatizer. “GLEM” is considered a more accurate lemmatizer than “CLTK” for the ancient Greek language. However, both have been developed and evaluated using texts in an earlier dialect (Ionic-Attic) related to the Koine Greek dialect used in the New Testament. Therefore, their accuracy should be compared in texts, which have been written in the Koine Greek dialect. The latter, however, goes beyond the subject matter of this research paper.

As mentioned above, n-grams have been proven to be particularly effective. The terms (words or characters) can take binary values, zero (0) value when absent from the text, and the value one (1) when they appear. The term tf-idf (Term Frequency - Inverse Document Frequency) was used, where (a) the more often a term appears in a text, the more critical it is for its content and (b) the more texts a term appears in, the less information it is likely to have. The frequency of the term in a text appears as TF = (Multiple occurrences of term T) / (Number of terms), since it may be necessary when a term occurs several times in a text. The reverse document frequency occurs as IDF=log(N/n), where N is the total number of texts and n is the number of texts that contain the term since when a term appears in many texts, it cannot help to distinguish them. The product TF x IDF gives the TFIDF value. The higher IDF value, the more unique the term is. Features with low tf-idf are either commonly used in all documents or rarely used and appear only in large documents. Features with high tf-idf are often used in specific documents, but they are rarely used in all documents. In other words, tf-idf is a statistical tool from which we can discover how important a term is in a document from a collection of documents.

In the case of unprocessed text, the most common terms were stopwords. The stopwords are initially removed during text pre-processing, but some terms were converted to stopwords due to lemmatization. From the last observation, we concluded that if lemmatization took place before stopwords removal, the same texts would return fewer features, meaning that feature dimensions would be reduced. However, since stopwords have been removed before lemmatization, the terms that appear after lemmatization should not be considered as stopwords but as terms with important information. Also, it should be mentioned that we used the CLTK tool, which has 126 stopwords for Ancient Greek language, whereas in Python’s NLTK there are 256.

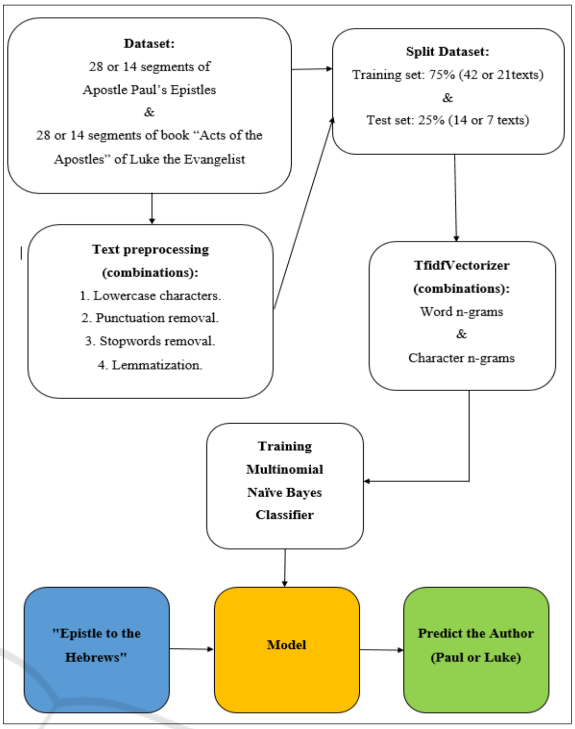

Figure 1: Proposed Methodology.

5 RESULTS

In order to test the proposed approach, we applied tf-idf (Term Frequency–Inverse Document Frequency) with various selections of n, in order to extract the features:

- word n-grams level (1 <= n <= 10)
- character n-grams level (1 <= n <= 10)

Thus, we tested the use of 10 different word n-grams Tfidf Vectors and 10 different character Tfidf Vectors in the texts; one time with pre-processing and one time without pre-processing. All the results are presented in Table 5.

The lowest accuracy is 21.4%, and it is for the case of word 7-grams with pre-processed texts.. The combination of pre-processed text with character 8-grams proved the most effective. Table 6 presents the comparison between text with and without pre-processing based on the average accuracy.

Concerning the character 8-gram terms of the text “Epistle to the Hebrews”, in the case of not pre-processed text the 2 most common terms were ‘τοῦ θεο’ and ‘τοῦ θεοῦ’. In the case of pre-processed text, the two most common character 8-gram terms were

[2] https://docs.cltk.org/en/latest/index.html

‘αὐτός’ and ‘ς αὐτός’. In the last case, we observed that the first term ‘αὐτός’ was the same as the first most common character 8-gram term of the training set.

Then, we repeated the above experiment with the second separation of the texts (28 texts). From the comparison of results (Table 5 & Table 6), we concluded that in the second experiment, there is no n-gram with absolute accuracy (100%); while the same texts were maintained, but with a different separation (from 28-28 to 14-14), we observe that some n-grams, while having high accuracy in the firstcase, showed low in the second (and vice versa). The same percentage of accuracy was maintained (42.9% in character 1-gram and in character 2-grams in unprocessed texts). In the case of 3-grams, there is an accuracy of only 35.7%, while it had one of the four highest with a percentage of 85.7%; so pre-processed texts seem to be more accurate. Character n-grams give better results than word n-grams; the character 8-grams seems to be the most effective in texts of the Koine Greek (dialect). The second experiment was the only ones with 85.7% accuracy and f-measure > 0.85; small texts return better results than large texts.

Table 5: Results of Multinomial Naïve Bayes Classifier (1st and 2nd case).

Tfidf Vectorizer	Pre-processing	Accuracy (1st case)	Accuracy (2nd case)
word 1-gram	×	85.7%	28.6%
word 2-grams	×	92.9%	71.4%
word 3-grams	×	50.0%	42.9%
word 4-grams	×	50.0%	71.4%
word 5-grams	×	64.3%	57.1%
word 6-grams	×	28.6%	28.6%
word 7-grams	×	50.0%	28.6%
word 8-grams	×	35.7%	57.1%
word 9-grams	×	28.6%	0.0%
word 10-grams	×	50.0%	14.3%
word 1-gram	✓	92.9%	85.7%
word 2-grams	✓	71.4%	14.3%
word 3-grams	✓	35.7%	85.7%
word 4-grams	✓	57.1%	28.6%
word 5-grams	✓	50.0%	28.6%
word 6-grams	✓	35.7%	42.9%
word 7-grams	✓	21.4%	28.6%
word 8-grams	✓	50.0%	28.6%
word 9-grams	✓	50.0%	71.4%
word 10-grams	✓	35.7%	42.8%
char 1-gram	×	42.9%	42.9%
char 2-grams	×	42.9%	42.9%
char 3-grams	×	85.7%	28.6%
char 4-grams	×	92.9%	57.1%
char 5-grams	×	71.4%	57.1%
char 6-grams	×	92.9%	28.6%
char 7-grams	×	78.6%	42.9%
char 8-grams	×	92.9%	14.3%
char 9-grams	×	92.9%	14.3%
char 10-grams	×	78.6%	85.7%
char 1-gram	✓	42.9%	42.9%
char 2-grams	✓	78.6%	42.9%
char 3-grams	✓	92.9%	28.6%
char 4-grams	✓	85.7%	71.4%
char 5-grams	✓	92.9%	57.1%
char 6-grams	✓	92.9%	71.4%
char 7-grams	✓	92.9%	71.4%
char 8-grams	✓	100.0%	85.7%
char 9-grams	✓	92.9%	71.4%
char 10-grams	✓	85.7%	57.1%

Table 6: Average accuracy between word and character n-grams (1st and 2nd case).

Tfidf Vectorizer	Texts with pre-processing	Accuracy avg. (1st case)	Accuracy avg. (2nd case)
word n-grams	×	53.6%	40.0%
word n-grams	✓	50.0%	45.7%
character n-grams	×	77.2%	41.4%
character n-grams	✓	85.7%	60.0%

In the third experiment, we used multiple-length n-grams. We observed that: while in the case of 56 no preprocessed texts, the non-multiple-length character n-grams had achieved an accuracy of 77.2%, the multiple-length character n-grams achieved 79.8%. Similarly, while in the 28 unprocessed texts, the non-multiple-length n-grams had an accuracy of 41.4%, the multiple-length character n-grams achieved an accuracy of 57.1%. In contrast, in the 56 pre-processed texts the non-multiple-length n-grams had achieved an accuracy of 85.7%, while the multiple-length character n-grams had only 76.2%. In all 28 pre-processed texts, the non-multiple-length n-grams had achieved an accuracy of 60.0%, while the multiple-length character n-grams had only 47.6%. In other words, in unprocessed texts the non-multiple-length character n-grams showed less accuracy than the multiple-length character n-grams. On the contrary, in pre-processed texts, the non-multiple-length character n-grams were more accurate than the multiple-length character n-grams.

Table 7: Accuracy comparison of multiple-length character n-grams, by dividing the Dataset into 56 and 28 texts.

Tfidf Vect. Character	Texts	Pre-proc.	Acc.	Acc. avg.
5-6-grams	28+ 28	×	57.1%	79.8%
5-6-7-grams	28+ 28	×	85.7%	
5-6-7-8-grams	28+ 28	×	92.9%	
6-7-grams	28+ 28	×	57.1%	
6-7-8-grams	28+ 28	×	92.9%	
7-8-grams	28+ 28	×	92.9%	
5-6-grams	28+ 28	✓	92.9%	76.2%
5-6-7-grams	28+ 28	✓	85.7%	
5-6-7-8-grams	28+ 28	✓	85.7%	
6-7-grams	28+ 28	✓	92.9%	
6-7-8-grams	28+ 28	✓	85.7%	
7-8-grams	28+ 28	✓	14.3%	
5-6-grams	14+ 14	×	71.4%	57.1%
5-6-7-grams	14+ 14	×	14.3%	
5-6-7-8-grams	14+ 14	×	71.4%	
6-7-grams	14+ 14	×	57.1%	
6-7-8-grams	14+ 14	×	85.7%	
7-8-grams	14+ 14	×	42.9%	
5-6-grams	14+ 14	✓	57.1%	47.6%
5-6-7-grams	14+ 14	✓	71.4%	
5-6-7-8-grams	14+ 14	✓	85.7%	
6-7-grams	14+ 14	✓	14.3%	
6-7-8-grams	14+ 14	✓	28.6%	
7-8-grams	14+ 14	✓	28.6%	

The above tests were performed for all features. For example, in the case of the 56 texts of Dataset, with characters’ conversion to lowercase, Stopwords removal but without Lemmatization and Punctuation marks removal, in one case the training set included 174,270 features. Looking for the ideal number of features in character 8-grams, we tested twenty different numbers of features (5, 10, 25, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 1000, 2000, 5000, 10000, 20000, 50000 and 100000 features) in character 8-grams. The percentage of accuracy did not decrease from 150 to 10,000 features.

Given the subtraction of the stopwords and the conversion of the characters to lowercase, we observed that the application of lemmatization and/or the removal of punctuation marks did not affect the accuracy of the classifier. This occurred when we divided the training set into 56 and 28 texts.

When the number of features changes from 5 to 100, the accuracy decreases. Also, the lemmatization application seems to return better results than removing punctuation marks when few features (5 or 10) are used. The opposite is observed when more features are used (25 or 100).

Then, we processed the texts of the Dataset” (a. characters’ conversion to lowercase, b. Stopwords removal, but without Lemmatization and Punctuation marks removal) and we classified the “Epistle to the Hebrews” using the following:

- character 8-grams
- Dataset with 56 texts
- 150, 200, 250, 300, 350, 400, 450, 500, 1000, 2000, 5000 and 10000 features

In all cases (for 12 different sets of features), the classifier classified the text “Epistle to the Hebrews” as a text that more closely resembles the writing style of Luke than that of Paul. For visualization, we reduced the dimensions (features) to 2, using the feature_selection class [3] of the sklearn library. However, the text “Epistle to the Hebrews” was not clear on Luke’s side. We observed that the best separation was achieved for a number of features equal to 300, while the more the number of features increased, the separation became less distinct

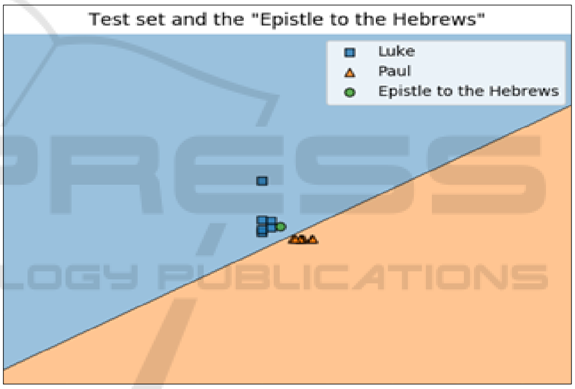

Figure 2: Paternity of the text “Epistle to the Hebrews”, for 8-grams character, Dataset with 56 texts and 300 features.

Having come up with the options that increase the accuracy of the classifier (character 8-grams, 300 more common features) from our experiments, we applied them in two (2) new experiments. Since from the previous research there is no doubt about the first three epistles (“Epistle to the Romans”, “First Epistle to the Corinthians” and “Second Epistle to the Corinthians”) that they belong to Paul, we used them together with the book “Acts of the Apostles” by Luke as the only texts of our Dataset.

The punctuation marks were removed in pre-processing, and the characters were capitalized to take their original form. The stopwords were not removed in the first experiment, with Paul’s texts

[3] https://scikit-learn.org/stable/modules/feature_selection.html

consisting of 18,640 words, while Luke's 18,772. In the second, the stopwords were removed, with Paul's texts consisting of 11,353 words, while Luke's 11,876. We reduced the number of words per text (to 1,200 words) to produce a larger number of texts.

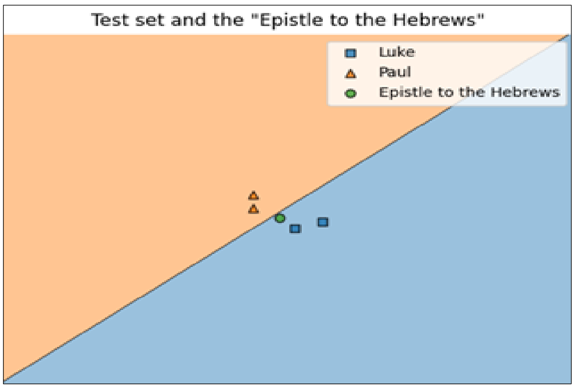

Figure 3: Paternity of the text "Epistle to the Hebrews", for 8-grams character and 300 features without stopwords removal, using only the first 3 epistles of Paul.

In first experiment, we divided the terms of Paul and Luke into 32 texts (16 and 16 respectively, with 1,200 terms per text for the first 15 texts of each, with the last of Paul consisting of 640 terms and the last of Luke from 772). In the second, we divided 20 texts (10 and 10 respectively, similarly with 1,200 terms per text for the first 9 texts of each, with the last of Paul consisting of 553 terms and the last of Luke from 1,076). After the corresponding pre-processing in the text "Epistle to the Hebrews" for both experiments (4,996 and 3,216 terms, respectively), we asked the classifier to classify it. Since Dataset is smaller than the equivalent of the above experiments (only 3 of Paul's 13 epistles were used), we increased the percentage of the training set from 75% to 90%. So the training set in both experiments consisted of 28 and 18 texts, respectively, and the test set of 4 and 2 texts, respectively.

We noted that although the classifier classified the training set without errors (100%), the classification of the text "Epistle to the Hebrews" did not improve concerning the above experiments. The classification turned out to be worse in the case of stopwords subtraction. The non-removal of the stopwords achieved a better separation in the test set, with the text "Epistle to the Hebrews" continuing to be classified on the side of Luke, while remaining very close to the decision boundary, thus preventing safe conclusions. As for the most common character 8-grams terms, in the first case (the stopwords were not removed) were: "ου θεου" and "του θεου", while in the second case (the stopwords were removed) were: "αυτους" and "πνευματ".

6 CONCLUSIONS

Various combinations of testing the available texts were investigated, with and without pre-processing. After selecting the best approach (character 8-grams), the text "Epistle to the Hebrews" was checked between Apostle Paul and Luke the Evangelist, proving that it is more like to belong to Luke. It is seen that although the text "Epistle to the Hebrews" appears in Luke's area, it is very close to the decision boundary, which justifies the fact that theologists debate about its paternity.

In our experiments, we applied n-grams and Multinomial Naïve Bayes Classifier in texts of Koine Greek dialect, and we concluded that:

- Character 8-grams are the most effective
- Suggested the use of the 300 most frequently used terms (number of features), including punctuation marks
- The training set is more effective when divided into smaller texts, compared to larger texts
- The pre-processing of the texts of the training set brings greater accuracy
- Basic pre-processing involves the removal of stopwords and converting characters to lowercase

Future work will include various investigations. The text "Epistle to the Hebrews" could be checked among the other possible writers. We will also apply other classifiers, such as Support Vector Machines and Artificial Neural Networks.

ACKNOWLEDGEMENTS

This work is funded by the project: "Enhancing research activities of Laboratory of Knowledge and Intelligent Computing", supported by the Research Committee of the University of Ioannina No 61240.

REFERENCES

Aggarwal, C. (2015). *Data Mining: The Textbook*. Springer. Switzerland, pages 31, 342, 429-430.

Aggarwal, C. (2018). *Machine Learning for Text*. Springer. Switzerland, pages 118, 122.

Aggarwal, C., ChengXiang, Z. (2012). *Mining text data*. Springer. Switzerland, pages 3-6, 52, 167, 213.

Apostolidis, A. (2016). *The Man of Faith in His "Word of Faith" to the Hebrew Letter (Heb. 11, 1-40)*. National and Kapodistrian University of Athens. Athens, pages 76, 77, 78, 87, 88, 93, 100-101, 534.

Aragon, M.E., Lopez-Monroy, A.PAGES (2018). A Straightforward Multimodal Approach for Author Profiling. In *Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier, editors, CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers. ISSN 1613-0073*, pages 7.

Athanasopoulou, E.E. (2006). *Applications of Machine Learning to Text Categorization*. University of Patras. Patras, page 21.

Avros, R., Soffer, A., Volkovich, Z., Yahalom, O. (2012). An Approach to Model Selection in Spectral Clustering with Application to the Writing Style Determination Problem. In *IC3K 2012*. pages 19-36.

Bamman, D., Mambrini, F., Crane, G. (2009). An ownership model of annotation: The Ancient Greek dependency treebank. In *Proceedings of the eighth international workshop on treebanks and linguistic theories (TLT 8)*. pages 5-15.

Bary, C., Berck, P., Hendrickx, I. (2017). A Memory-Based Lemmatizer for Ancient Greek. In *Proceedings of DATeCH2017*. Gottingen. Germany.

Berry, D. (2011). The computational turn: Thinking about the digital humanities. In *Culture Machine. Vol 12*, pages 1, 13.

Bizzoni, Y., Boschetti, F., Diakoff, H., Gratta, R.D., Monachini, M., Crane, G.R. (2014). The Making of Ancient Greek WordNet. In *LREC*. pages 1-7.

Boulis, C., Ostendorf, M. (2005). Text Classification by Augmenting the Bag-of-Words Representation with Redundancy-Compensated Bigrams, page 1.

Celano, G., Crane, G. (2015). Semantic Role Annotation in the Ancient Greek Dependency Treebank. In *Dickinson, Markus et al. Proc. of the 14th International Workshop on Treebanks and linguistic Theories (TLT14)*. pages 26-34.

Dabagh, R. (2007). Authorship attribution and statistical text analysis. *Vol 4. No 2*, pages 150.

Ebrahimpour, M., Putnins, TJ., Berryman, MJ., Allison, A., Ng, BWH., Abbott, D. (2013). Automated Authorship Attribution Using Advanced Signal Classification Techniques. In *PLOS ONE 8(2): e54998*.

Eder, M., Piasecki, M., Walkowiak, T. (2017). An open stylometric system based on multilevel text analysis. In *Cognitive Studies / Etudes cognitives*. pages 1-2.

Eusebius. (1890). *The Church History (Translated by Arthur Cushman McGiffert)*. Christian Literature Publishing Co. Buffalo. NY.

Eyheramendy, S., Madigan, D. (2005). *A Novel Feature Selection Score for Text Categorization*, page 5.

HaCohen-Kerner, Y., Miller, D., Yigal, Y., Shayovitz, E. (2018). Cross-domain Authorship Attribution: Author Identification using Char Sequences, Word Uni-grams, and POS-tags Features — Notebook for PAN at CLEF 2018. In *Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier. editors. ISSN 1613-0073*, page 1.

Jankowska, M., Milios, E., Keselj, V. (2014). Author verification using common n-gram profiles of text documents. In *Proc.of 25th International Conference on Computational Linguistics: Technical Papers*, page 395.

Jockers, M., Witten, D. (2010). A comparative study of machine learning methods for authorship attribution. In *LLC. 25. 215-223, 10.1093/llc/fqq001*, page 221.

Juola, PAGES (2008). Authorship Attribution. In *Foundations and Trends in Information Retrieval 1*, page 233.

Karavidopoulos, I. (1998). *Introduction to the New Testament*. Pournaras. Thessaloniki, pages 394-395, 404, 406.

Katakis, I. (2015). *Machine Learning Methods for Automatic Text Classification*. Aristotle University of Thessaloniki. Thessaloniki. pages 29-30.

Kenny, A. (1986). *A Stylometric Study of the New Testament*. Oxford University Press. UK.

Kestemont, M., Stamatatos, E., Manjavacas, E., Daelemans, W., Potthast, M., Stein, B. (2019). Overview of the Cross-domain Authorship Attribution Task at PAN 2019. In *Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller eds. CLEF 2019 Labs and Workshops, Notebook Papers*, pages 1, 13.

Kocher, M., Savoy, J. (2015). UniNE at CLEF 2015: Author Identification. In *Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, CLEF 2015 Evaluation Labs and Workshopages ISSN 1613-0073*, pages 1, 7.

Kocher, M., Savoy, J. (2017). A simple and efficient algorithm for authorship verification. In *Journal of the Association for Information Science and Technology 68.1*, page 259.

Kocher, M., Savoy, J. (2019). Evaluation of text representation schemes and distance measures for authorship linking. In *Digital Scholarship in the Humanities. 34(1)*. pages 189–207.

Koppel, M., Schler, J., Argamon, S. (2009). Computational methods in authorship attribution. In *Journal of the American Society for information Science and Technology 60.1*, page 9.

Koppel, M., Schler, J., Argamon, S., Winter, Y. (2012). The "Fundamental Problem" of Authorship Attribution. In *English Studies, 93*, page 284.

Koppel, M., Schler, J., Bonchek Dokow, E. (2007). Measuring differentiability: Unmasking pseudonymous authors. In *Journal of Machine Learning Research. 8(6)*. pages 1261-1276.

Koppel, M., Seidman, S. (2018). Detecting pseudoepigraphic texts using novel similarity measures. In *Digital Scholarship in the Humanities. 33(1)*. pages 79–80.

Kotsiantis, S. (2007). Supervised machine learning: A review of classification techniques. In *Emerging artificial intelligence applications in computer engineering 160*, pages 255, 262.

Kourtis, I., Stamatatos, E. (2011). Author Identification Using Semi-supervised Learning. In *Vivien Petras, Pamela Forner, and Paul D. Clough, editors, Notebook Papers of CLEF 2011. ISBN 978-88-904810-1-7. ISSN 2038-4963*. pages 1-3.

Ledger, G. (1995). An Exploration of Dierences in the Pauline Epistles using Multivariate Statistical Analysis. In *Literary and Linguistic Computing. 10. 85-97*.

Lopez-Escobedo, F., Mendes-Cruz, C.F., Sierra, G., Solorzano-Soto, J. (2013). Analysis of stylometric variables in long and short texts. In *5th Int. Conf. Corpus Linguistics. University of Alicante. Spain*, pages 1-3, 5.

Maharjan, S., Solorio, T. (2015). Using Wide Range of Features for Author Profiling. In *Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, CLEF 2015, ISSN 1613-0073*, page 4.

McCallum, A., Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. In *Learning for Text Categorization: Papers from the 1998 AAAI Workshopages*, page 6.

Mealand, D.L. (1995). The Extent of the Pauline Corpus: A Multivariate Approach. In *Journal for the Study of the New Testament 18 (59)*. ppages 61-92.

Niforas, N. (2016). *Use of Text Mining in Classification of Legislation*. University of Patras. Patras, page 17.

Panagopoulou, I. (1994). *Introduction to the New Testament*. Akritas. Athens, page 330.

Porter, S. (1995). Pauline Authorship and the Pastoral Epistles: Implications for Canon. In *Bulletin for Biblical Research 5*. pages 105-123.

Potha, N., Stamatatos, E. (2019). *Authorship Verification*. University of Aegean. Greece. pages 112-113.

Putnins, T., Signoriello, D.J., Jain, S., Berryman, M.J., Abbott, D. (2006). Advanced text authorship detection methods and their application to biblical texts. In *Complex Systems. Vol 6039. doi:10.1117/12.639281*.

Rehurek, R. (2011). *Scalability of Semantic Analysis in Natural Language Processing*. Masaryk University. Brno, page 7.

Sagayam, R., Srinivasan, S., Roshni, S. (2012). A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques. In *International Journal of Computational Engineering Research, Vol 2, Issue 5*, pages 1443.

Savoy J. (2019). Authorship of Pauline epistles revisited. In *Journal of the Association for Information Science and Technology*.

Savoy J. (2020). *Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling*, Springer. Switzerland.

Shalymov, D., Granichin, O., Klebanov, L., Volkovich, Z. (2016). Literary writing style recognition via a minimal spanning tree-based approach. In *Expert Systems with Applications. 61*, pages 145–153.

Stamatatos, E. (2007). Author Identification Using Imbalanced and Limited Training Texts. In *18th International Workshop on Database and Expert Systems Applications*, page 5.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. In *Journal of the American Society for Information Science and Technology, Vol 60, No 3*, pages 1, 3-12.

Stamatatos. E. (2017). Authorship Attribution Using Text Distortion. In *Proceedings of EACL, 15th International Conference on Computational Linguistics*, pages 1.

Stamatatos, E., Fakotakis, N., Kokkinakis, G. (2000). Automatic Text Categorization in Terms of Genre and Author. In *Computational Linguistics 26.4. ISSN: 0891-2017*, page 472.

Stanko, S., Lu, D., Hsu, I. (2013). *Whose Book is it Anyway? Using Machine Learning to Identify the Author of Unknown Texts*. Stanford University. Stanford, page 2.

Su, J., Sayyad-Shirabad, J., Matwin, S. (2011). Large Scale Text Classification using Semi-supervised Multinomial Naive Bayes. In *ICML*, pages 1.

Van Dam, M. (2013). A Basic Character N-gram Approach to Authorship Verification. In *Pamela Forner, Roberto Navigli, and Dan Tufis, editors, CLEF 2013, ISBN 978-88-904810-3-1, ISSN 2038-4963*, page 3.

Witten, I., Frank, E. (2005). *Data mining: Practical Machine Learning Tools and Techniques*. Elsevier. UK, 2nd edition, page 351.

Xu, S., Li, Y., Wang, Z. (2017). Bayesian Multinomial Naïve Bayes Classifier to Text Classification. In *10.1007/978-981-10-5041-1_57*, page 1.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. In *Information retrieval 1.1-2*, page 3.

Yang, Y., Olafsson, S. (2005). *Near-Optimal Feature Selection*, page 1.

Zangerle, E., Tschuggnall, M., Specht, G., Potthast, M., Stein, B. (2019). Overview of the Style Change Detection Task at PAN 2019. In *Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers*, page 1.

(2013). *The New Testament*. Brotherhood of Theologians «THE SAVIOR». Athens, page 929.