ANALYSIS OF EFFECTIVE TECHNIQUES AND ALGORITHMS IN TERMS OF "TEXT MINING" TO PREDICT THE AUTHORSHIP IN ALBANIAN LANGUAGE

Dr. Miranda Harizaj*, Msc. Arli Minga**, Dr. Alfons Harizaj***

* Polytechnical Univistry of Tirana, Tirana, Albania miranda.harizaj@fie.edu.al
** National Agency for Information Society,Tirana, Albania arli.minga@cit.edu.al
*** Canadian Institute of Technology, Rr: „Xhanfize KEKO", Tirana, Albania alfons.harizaj@cit.edu.al

ABSTRACT

Natural Language Processing has gained a special importance and development in recent years, where the analysis of written texts through various techniques of "text mining" and the extraction of all their features is a prerequisite to be used and be further implemented for various purposes. In this paper it will be compared some of the most effective techniques and algorithms in terms of "text mining", to predict the authorship of a written text in the Albanian language, using for training the model, a fund of articles written by some of the most well- known bloggers of Albanian journalism. When talking about finding the authorship of a text it must kept in mind many important elements such as: number of sentences, sentence structure, number of words in a sentence, repetition of the same word, length of words used, frequency of the use of punctuation, literary figures used; elements which best display the unique narrative style for each author. This paper can serve as a good starting point to go further to its specific objective, predicting the authorship of an anonymous text, but also for other applications related to "text mining", referring to the Albanian language.

1 INTRODUCTION

Natural Language Processing is a field of study, which has gained a special importance and development momentum in recent years, where the analysis of written texts through various techniques of "text mining" and extracting all their features is a prerequisite to be used and be further implemented for various purposes [11 & 13]. A special and interesting case is the implementation in the Albanian language.

This paper will compare some of the most effective techniques and algorithms in terms of "text mining", to predict the authorship of a text written in the Albanian language[4 & 18], having as a model training fund articles of some of the well-known bloggers of Albanian journalism [18].

When talking about finding the authorship of a text it must be kept in mind many important elements such as: number of sentences, sentence structure, number of words in a sentence, repetition of the same word, length of words used, frequency of the use of punctuation, literary figures used; elements which best display the unique narrative style for each author [20 &22].

One method of determining the authorship of a handwritten document, especially a text document as content, is ancient. The importance of the author attribution problem stems from its use in forensic analysis, textbooks, e-commerce and the development of innovative techniques to enable such a process [16].

Data mining techniques have become more popular in recent years for a variety of reasons [14 & 31]. Data in any format, including text, pictures, mass data and multimedia, is now available and endlessly spread online. "Data mining" in itself has evolved as a concept and with an even greater speed have evolved the techniques that make it possible, according to each case or problem that arises.

One of them is Stylometry, which is the study of different language styles and writing habits in order to determine the authorship of a written part of the text. A writing style refers to a writer's language choices that remain the same throughout his or her work. What is achieved as a result after the stylometric research is that all the lyricists have a unique and special writing style, which can

be evaluated and learned, thus serving as input for the training of Machine Learning models [7, 8 & 17]. A writer's narrative style refers to a set of characteristics he or she typically uses, such as word length, sentence length, use of certain words, and syntactic structure of phrases.

It's possible to know the author, the translator of some works only by the way of writing. In this context, two main problems may arise:

- Decide who wrote a certain text among some well-known authors; this problem is called Author Recognition [23].
- Identify whether or not an author has written a particular document, using a small collection of documents, all written by that author is called Author Identification[1, 2 &3].

In the paper are presented two approaches that can be used, which are: profile-based classification, probabilistic model. Also best methods in the field are researched and presented as:

1. SVM (Support Vector Machine) that can be used to successfully solve textbook classification problems [6].
2. The Naive Bayes Classifier is a simple classifier that is classified based on the probabilities of events and it is usually applied in text classification.
3. A "neural network" is a set of algorithms that attempts to recognize the interrelated relationships between a set of data [6]. An approach to textbook classification based on Deep Learning is the implementation of a "one layer neural network".

In the paper it is presented the implementation of "logistic regression", which in itself can be seen as an algorithm found between Machine Learning and Deep Learning, as it is interpretable as a "one layer NN".

The proposed built model presented in this paper is based on the use of the "Logistic Regression" algorithm, which is an algorithm that can be seen as a method that lies somewhere between "machine learning" and "deep learning", as it is easily conceived as a "one- layer neural network". This algorithm is most effective in cases where it should be a decision between two possible choices: in this case whether or not an author has written a specific text. The Python programming language has been chosen for the software component, taking into account the wide range of libraries that can be used in the field of "data science" and "text mining".

In conclusion the results achieved during the simulation are satisfactory while the proposed model, trained on a fund, which should be said to be limited in terms of the number of writings involved. Over 60% of cases finding the authorship of articles written in the Albanian media by the selected analysts, is successful.

2 RELATED WORKS

In the last couple years, the usage of data mining techniques is increasing and being applied in many areas, Data available might be in any format like text, images, binary, and multimedia. And several techniques of mining increased, modified, improved over the time. [31, 32 & 33]. The focus of this paper is on author identification techniques. Today the availability of text document in electronic form increases the importance of using automatic methods to analyze the content of text documents [2, 11, 17 &27].Initially identifying document was very time consuming, expensive and has its limit. That emerges text categorization in predefined categories called as classification. Categorization is based on certain properties called as features. There18 are various method for extraction of features [18 &20]. Writeprint is a method that is similar to finger printing. Another property of ngrams is that they provide information on rising word sequences based on their length [23 & 34]. Following that, another option is stylometric features, which offers a set of style indicators that have been tailored for automatic text analysis [5 & 18].

3 AUTHORSHIP IDENTIFICATION

3.1 METHODOLOGY

There are two approaches that can be used, which are: profile-based classification, probabilistic model. Using profile-based classification, first it should be managed a set of texts known to a particular author, which should be collected and stored in a database. This "great bookstore" is used to derive the characteristics of the narrative style of a particular author. The authorship of a text by an unknown author, ie

anonymous, is judged on the basis of this previous collection of texts where the style is analyzed and then it is sufficient to make a comparison. Judging the authorship of a text based on stylistic profile analysis is simple.

On the other hand, probabilistic models are one of the oldest approaches to identifying authorship that has been widely used in recent studies [6,7,8,17 & 18]. These approaches aim to maximize the probability P (x | a) that a text X belongs to a candidate author a. This approach is much more intriguing because it can be used both in character sequences and for words.

In fact, the best results for identifying authorship have been achieved using models that judge the level of word usage, without going into sentences, and compromising accuracy.

3.2 BEST METHODS

3.2.1 SUPPORT VECTOR MACHINE

SVM (Machine Vector Support) is a good Machine Learning method [7, 8, 17 and 27] that can be used to successfully solve textbook classification problems. However, it is mainly used to resolve categorization issues. Each data item is represented as an n-dimensional point in space, where the value of each property is the value of the SVM algorithm for a given coordinate. Scikit-learn is a well-known Python library for implementing Machine Learning algorithms [16]. In the scikit-learn library, SVM is also applicable where it should be judged the authorship of an article written in the Albanian media.

Support vector machines is an algorithm that determines the best decision boundary between vectors that belong to a given group (or category) and vectors that do not belong to it [10]. It can be applied to any kind of vectors which encode any kind of data. This means that in order to leverage the power of svm text classification, texts have to be transformed into vectors. On the other hand, Vectors are (sometimes huge) lists of numbers which represent a set of coordinates in some space.

So, when SVM determines the decision boundary mentioned above, SVM decides where to draw the best "line" (or the best hyperplane) that divides the space into two subspaces: one for the vectors which belong to the given category and one for the vectors which do not belong to it. It can find vector representations which encode as much information from the texts as possible to be able to apply the SVM algorithm to text classification problems and obtain very good results. Examples are widely spread over internet with accuracy score mainly over 80% [10].

3.2.2 NAIVE BAYES CLASSIFIER

The Naive Bayes Classifier is a simple classifier that is classified based on the probabilities of events. It is usually applied in text classification. Although it is a simple algorithm, it works well on many text classification problems [18]. The advantage is that it requires less training time and less training data. This means less CPU and RAM consumption.

As with any Machine Learning model [7, 8, 17 and 27], it is the need to have an existing set of examples (training group) for each data category. Let's consider sentence classification to classify a sentence into "question mark" or "demonstrative". In this case, there are two classes ("question mark" and "demonstrative"). With the training group, it can train a model based on the Naive Bayes Classifier, which can be used to automatically categorize a new sentence.

However the raw data, a sequence of symbols (i.e. strings) cannot be fed directly to the algorithms themselves as most of them expect numerical feature vectors with a fixed size rather than the raw text documents with variable length [16]. In order to address this, scikit-learn provides utilities for the most common ways to extract numerical features from text content, namely:

- Tokenizing strings and giving an integer id for each possible token, for instance by using white-spaces and punctuation as token separators.
- Counting the occurrences of tokens in each document.

In this scheme, features and samples are defined as follows:

- Each individual token occurrence frequency is treated as a feature.
- The vector of all the token frequencies for a given document is considered a multivariate sample.

3.2.3 NEURAL NETWORK

A "neural network" is a set of algorithms that attempts to recognize the interrelated relationships between a set of data using a method that mimics how the human brain works. NNs refer to neuronal systems which may be of an organic or artificial nature. Because NNs can be adapted to inputs that change from time to time, they can produce the best possible result without requiring the production criteria to be "recreated". The notion based on NN artificial intelligence has established an unrivaled reputation in creating trading systems.

Text classification is one of the popular tasks in NLP that allows a program to classify free-text documents based on pre-defined classes [11 & 13]. The classes can be based on topic, genre, or sentiment. Today's emergence of large digital documents makes the text classification task more crucial, especially for companies to maximize their workflow or even profits.

Recently, the progress of NLP research on text classification has arrived at the state-of-the-art (SOTA). It has achieved awesome results, showing Deep Learning methods as the cutting-edge technology to perform such tasks [11 and 15].

Also, keep in mind that beyond accuracy, Deep Learning techniques are also complex [1]. Complexity in implementation brings the need for a good, abundant and well-structured data fund, which in this case, due to the limitations of the Albanian language, but also for the lack of electronic materials, this is not possible.

NN can also be used in textbook classification. Deep Learning methods are proving successful day by day in textbook classification, achieving top-level results in a range of academic problems [19, 21 & 22]. An approach to textbook classification based on Deep Learning is the implementation of a "one layer neural network". Further in the paper it is presented the implementation of "logistic regression", which in itself can be seen as an algorithm found between Machine Learning and Deep Learning, as it is interpretable as a "one layer NN".

4 TEXT MINING FOR ALBANIAN LANGUAGE

In the case of this paper, but also in other cases when the Natural Language Processing is applied to the Albanian language, the challenges are mainly linguistic and not technical [4, 9 & 21]. The Albanian language has a complex morphological system. There are many forms for nouns, adjectives and numbers, which have five cases, two numbers (singular and plural) and determinability (indefinite and definite). Linguistic issues need to be mentioned and considered for textbook classification, including inflections, negation, homonyms, dialects, irony, and sarcasm[19,20 & 21].

- Inflections and word order - The first point noticed is that Albanian is a very inflectional language, compared to English. Albanian is a difficult language to learn due to the large number of different word forms.
- Negation - Negation is another topic where English and Albanian languages differ. English uses a negation form, but Albanian can use double, triple, or even quadruple negations.
- Homonyms - the presence of the same word in different meanings.
- Dialects - With two main dialects, the Albanian language presents an important challenge for Text Mining applications
- Irony and sarcasm - Irony and sarcasm are two types of communication in which the speaker writes the opposite of what he means.
- Sentence with vague meaning - There are many vague phrases that can have particular emotional polarity depending on the situation and context

5 PROPOSED MODEL

5.1 LOGISTIC REGRESSION

For the implementation it is selected Python programming language and we have developed the prediction model by using Logistic Regression [14 & 15]. The reasons for this choice will be presented below, but the main one is that dataset is not complex just authors assigned to the sentence that they have written in their previous articles.

Binary classification is the "classic" application of the logistic regression model. However, it can utilize many "flavors" of logistic to solve multiclass classification issues, such as the One-vs-All

or One-vs-One techniques, as well as the related softmax regression/multinomial logistic regression. Despite the existence of kernelized logistic regression versions, the conventional "model" is a linear classifier. If it is the case working with a dataset where the classes are more or less "linearly separable" logistic regression comes in handy.

Logistic regression is used to determine the probability of occurrence of a binary event, i.e with only two possibilities. As you can see, logistic regression is used to predict the probability of a series of events. Logistic regression assists data analysts in making well-informed decisions by predicting as accurately as possible. If the assessment and judge is still simple, logistic regression comes into play to increase efficiency and reduce costs. In this case, the judge should be whether or not an author is the real writer of a certain article? So logistic regression it is seen as a method that chooses only in two ways: Yes and No.

5.2 STOPWORDS

Stopwords are a group of words used in the everyday use of spoken and written language. Examples of stopwords in English are "a", "the", "is", "are" etc. Stopwords are commonly used in Text Mining and Natural Language Processing (NLP) to eliminate words that are used so often that they contain very little useful information [15]. Sothey are words that do not significantly affect the narrative style of an author [20 & 26] as in the case where a judge should be on the authorship of an article.

To make possible and implement in the most effective way possible the model of predicting authorship, it is created a fund of "stopwords" for the Albanian language such as: "I", "is", "you", "do", i.e. words that should be used for the sentence to make sense, but that do not distinguish one author from another, as they are words that are used with "compulsion". After creating this vocabulary, should proceed with the "clean" of the available items from these "stopwords" and then do the model training and then the forecasting.

It can be argued that the elimination of stopwords is not necessary because removing them could distort the meaning of the sentence or destroy the structure, but the technical context must be taken into account above all. In this case, if are removed from the word personal pronouns like: "unë", "ti" or "ne", it is not a problem, because in this case should be analyzed the words and not the phrase. Even removing from a sentence of conjunctions ("të", "e", etc.) is not a problem, as in this case it interprets them as necessary morphological, but not technical parts. As per the method they have been chosen to analyze the text, it is needed mostly the unique words, the ones that actually constitute and an important stylistic element.

5.3 DATASET THAT WILL BE USED

In the purpose of this paper it is created a dataset for the training of the proposed model. This dataset was populated from online data of news portals, a bunch of articles written by the most well-known opinionists in Albania.

First the articles were collected and then divided into sentences. The sentence is selected as the main cell for this kind of prediction and because the sentence is the essence of writing style.

The dataset structure is showed in:

Table No. 4.1 Dataset with Albanian language articles

	id	type	article	sentence	text	author
0	1001001	Opinion	1	1	Apeli i Gjykatës së Posatçme i ka dhën	Bushati
1	1001002	Opinion	1	2	Një trupë prej tre gjyqtarësh e kryesua	Bushati
2	1001003	Opinion	1	3	Natyrisht, që nga ajo kohë ndryshuan s	Bushati
3	1001004	Opinion	1	4	Së pari gjykata e të drejtave të njeriut	Bushati
4	1001005	Opinion	1	5	Pas kësaj, një sërë mediash ndërkomb	Bushati
5	1001006	Opinion	1	6	Ajo po merrej me median dhe gazetarë	Bushati
6	1001007	Opinion	1	7	Por, elementi kryesor, që e ndryshoi lo	Bushati
7	1001008	Opinion	1	8	Përpara 25 prillit SPAK mori përsipër ro	Bushati
8	1001009	Opinion	1	9	Ai bleu kohë për ti lënë hapsirë partisë	Bushati
9	1001010	Opinion	1	10	Dhe tani që ky objektiv u arrit, vazhdim	Bushati
10	1001011	Opinion	1	11	Ndonëse përfaqësuesja e institucionit	Bushati
11	1001012	Opinion	1	12	Ata e bënë të tyren duke e zvarritur pa	Bushati
12	1001013	Opinion	1	13	Por, tani kur gjithcka u mbyll me "happ	Bushati

5.4 IMPLEMENTATION OF PROPOSED MODEL

For implementation it is selected the Python programming language and developed the forecasting model using Logistic Regression and also used the "Sci-kit Learn" library, which is the best Python library for Text Mining and has many integrated techniques and algorithms, which provide a high effectiveness for model training as in this case.

Pandas is a data manipulation and analysis software package for the Python programming language. It mostly consists of data structures and methods for manipulating numerical tables and time series. It's BSD-licensed open-source soft-

ware with three clauses.

NumPy is a Python library that provides support for massive, multi-dimensional arrays and matrices, as well as a large number of high-level mathematical functions that may be used to manipulate these arrays.

The scikit-learn toolkit in Python has a fantastic utility called ***CountVectorizer***. It's used to turn a text into a vector based on the frequency (count) of each word in the text. This is useful when dealing with a large number of such texts and converting each word into a vector (for using in further text analysis).

The ***train_test_split*** function in Sklearn model selection divides data arrays into two subsets: training data and testing data. You don't have to divide the dataset manually with this function. Sklearn train test split creates random divisions for the two subsets by default.

Accurracy_score - In multi label classification, this function computes subset accuracy: the set of labels predicted for a sample must precisely match the corresponding set of labels in y true.

It is used Pandas function read, to read the used dataset saved as a *Comma-separated values* file. If it is carefully seen in that line of code, the first thing that comes in mind is that why are used the encoding equal to "latin-1". The answer is simple. Python and any other programming language has a default encoder set to "utf-8". In this case, that it is required to read from a file that has in it text written in Albanian language it should be read it with encoder set to "latin-1". The reason behind this decision is to make the computer "able" to understand "ë" and "ç" which are the most characteristic letters in albanian language. By its default encoder, it cannot read these letters because with "utf-8" encoder those letters are classified as special characters.

Then to check that everything has gone well it is used df.head, which is a line that makes us able to see the first 5 lines of a data frame. The result from that piece of code is shown in the Figure below:

1001001	Opinion	1	1	Apeli i Gjykatës së Posatçme i ka dhënë k
1001002	Opinion	1	2	Një trupë prej tre gjyqtarësh e kryesuar D
1001003	Opinion	1	3	Natyrisht, që nga ajo kohë ndryshuan shumë
1001004	Opinion	1	4	Së pari gjykata e të drejtave të njeriut n
1001005	Opinion	1	5	Pas kësaj, një sërë mediash ndërkombëtare

Figure 4. 1 Dataset to data frame

As a result the data frame has been created successfully and the first read sentences in the dataset are from Andi Bushati articles.

An important problem that needs to be addressed before starting with text mining methods is the elimination of stop words. Stop words are terms that are taken out of natural language data before or after processing. ***Stop words*** are the most prevalent terms avoided by most search engines in order to save space and time while processing huge amounts of data. Examples of stop words in English are "a", "the", "is", "are" and etc. In the Albanian language, it should be created from the beginning a list of most used stop words. Examples of stop words for the Albanian language can be mentioned: "unë", "një", "është", "janë", etc., words which are found in many sentences, where for Albanian it is inevitable and the use of conjunctions such as: "të".

```
with open('stopwords_alb.txt', encoding='latin-1') as f:
    content = f.readlines()
content = [x.strip() for x in content]
```

In the snippet of above code it is created the option for computer to read line by line in a text document all the stop words for the Albanian language, which will be used for cleaning the data.

```
df['text'] = df.text.str.replace("[^\w\s]", "").str.lower()
df['text'] = df['text'].apply(lambda x: [word for word in x.split() if word not in content])
```

In the code snippet above it is doable to search dataset line by line in "text" column where are stored the sentences written by Albanian analysts. The code takes the sentence and "cuts" it in pieces by creating an array. After that there is the possibility to analyze every sentence and where the program finds words that are in the list of stop words, it decides to delete it. This is an important step to work with strings and text classification.

The result of the operation described above is shown in the figure below:

Opinion	1	1	[apeli, gjykatës, së, posatçme, dhënë, enjte, ...	Bushati	apeli gjykatës së posatçme dhënë enjte drejtë ...
Opinion	1	2	[trupë, tre, gjyqtarësh, kryesuar, dhimitër, l...	Bushati	trupë tre gjyqtarësh kryesuar dhimitër lara ho...
Opinion	1	3	[natyrisht, kohë, ndryshuan, gjëra]	Bushati	natyrisht kohë ndryshuan gjëra ...
Opinion	1	4	[së, pari, gjykata, drejtave, njeriut, strasbu...	Bushati	së pari gjykata drejtave njeriut strasburg pre...
Opinion	1	5	[sërë, mediash, ndërkombëtare, shoqatash, liri...	Bushati	sërë mediash ndërkombëtare shoqatash lirisë së...

Figure 4. 2 Data frame with cleaned text using the selected list of stop words

Xfeatures = df['test_without_sw']

ylabels = df['author']

cv = CountVectorizer()

X = cv.fit_transform(Xfeatures)

Then in order to build the model using logistic regression firstly all sentences of the dataframe are vectorizes, which has been firstly cleaned from stopwords.

Fit_transform – is a function that transforms text in a sparse matrix. The majority of the elements in a sparse matrix are zero, thus they are not kept to conserve memory. The index of the value in the matrix (row, column) is indicated by the numbers in brackets, and 1 is the value (The number of times a term appeared in the document represented by the row of the matrix).

The result of fit_transform is like:

(0, 1) 1

(0, 2) 1

(0, 6) 1

(0, 3) 1

(0, 8) 1

(1, 5) 2

It can describe this result as *"(sentence_index, feature_index) count".* As there are 2 sentence: it starts from 0 and ends at 1. So for the example (0, 1) 1

- *0 : row [the sentence index]*
- *1 : get feature index (word) from vectorizer. vocabulary_[1]*
- *1 : count/tfidf it is the count of how many times a specific word has appeared in the text*

After that it has to split the selected data in *train* portion and *test* one.

x_train,x_test,y_train,y_test = train_test_split(X,y-labels,test_size=0.33,random_state=42)

test_size - *float or int, default=None* - If float, it should be between 0.0 and 1.0. If int, it refers to the total number of test samples. If the value is None, the complement of the train size is used.

train_size - *float or int, default=None* - If float, it should be between 0.0 and 1.0 and reflect the percentage of the dataset that should be included in the train split. If int, the absolute number of train samples is represented. If None, the value is automatically set to the complement of the test size.

random_state - *int, RandomState instance or None, default=None* - Controls how the data is shuffled before the split is done. Pass an int for reproducible output across multiple function calls.

logit = LogisticRegression()

logit.fit(x_train,y_train)

print("Accuracy of Logit Model :",logit.score(x_test,y_test))

In the code snippet above it is created the prediction model and printed out the result of the accuracy that it produces, which in this case is above 60 %.

anonymous

ext = "Është e vështirë sot të parashikosh nëse do të vazhdonin të qëndronin në pushtet në Ballkan autokratë si Vuçiçi, Rama apo Gjukanoviçi, nëse vende të BE nuk do të drejtoheshin nga Orbani, vëllezërtit Kazinski apo Janesh Janza në Sllloveni."

querywords = anonymous_text.split()

resultwords = [word for word in querywords if word.lower() not in content]

anonymous_text_without_sw = [' '.join(resultwords)]

In the other step, written in the code snippet above, it is one sentence from another article of Andi Bushati, which is not part of the selected dataframe. First it should be split word by word to initially remove all the stop words from it, as previously proceeded with all the sentences inside dataframe, then it should be returned it as it was, a string.

vect4 = cv.transform(anonymous_text_without_sw).toarray()

logit.predict(vect4)

Result: array(['*Bushati*'], dtype=object)

In the final step of the program the proposed model is set into prediction stuff. In this case the model has successfully and correctly predicted that the sentence (anonymous text written above) is originally written by Andi Bushati.

6 CONCLUSIONS

Although in different parts of the world, where there is great technological progress, but also a more efficient penetration of new technologies related to Big Data, where Authentication Using NLP is a well-known practice, in fact, for Regarding the Albanian language, and a new field of study remains.

This paper contributed in the creation of a fund of stopwords" that makes the classification of text more effective, while these types of words are not part of the analysis, because they are deleted from the text as they do not affect the narrative style.

It is concluded that for this type of data, logistic regression is a good approach to get the desired results in terms of textbook classification, where the proposed model achieved a satisfactory accuracy.

The proposed model can predict the author of a particular piece of text using logistic regression and achieved an accuracy of above 60%. Although we had an accurate result when analyzing the public text written by one of the authors in the used database, this kind of accuracy requires from us additional work, mainly the development of a wider database, a richer vocabulary and the next step is to develop a model for sentence classification.

REFERENCES

1. Chen Qian, Tianchang He, R.Zhang, "Deep Learning based Authorship Identification", Stanford, 2018
2. J.Frery, M.J.Mathieu, "Author Identification by Automatic Learning", August 2015
3. M.Sh. Tamboli, R.S. Prasad, "Authorship Analysis and Identification Techniques: A Review", International Journal of Computer Applications, September 2013.
4. T.K. Mustafa, "Text Mining authorship detection methods development", August 2018
5. K. Hoxha, "Albanian language identification in text documents", Tirana, 2017
6. A.Romanov, A. Kurtukova, A. Shelupanov, A. Fedotova, V. Goncharov, "Authorship identification of Russian language text, using SVM and DNN", Basel, December 2020
7. W. Anwar, I.S. Bajwa, Sh. Ramzan, "Design and Implementation of a Machine Learning-Based Authorship Identification Model", Pakistan, January 2019
8. R.R. Iyer, C.P. Rose, "A Machine Learning Framework for Authorship Identification From Texts", USA, December 2019
9. N. Zanini, V. Dhawan, "Text Mining: An introduction to theory and some applications", UCLES, 2015
10. E. Stamatatos, "A Survey of Modern Authorship Attribution Methods", Greece, 2017
11. N.K. Alhuqail, "Author Identification based on NLP", European Journal of Computer Science and Information Technology, 2021
12. E. Uyar, "Authorship attribution", December 2007
13. R. Kibble, "Introduction to NLP", Goldsmiths, 2013
14. R. H. R. Tan and F. S. Tsai, "Authorship Identification for Online Text," 2010 International Conference on Cyberworlds, 2010, pp. 155-162, doi: 10.1109/CW.2010.50
15. R. Kukunuri, "Logistic Regression and it's applications in Natural Language Processing", Medium.com, December 2018
16. K. Perifanos, "Authorship Attribution and Forensic Linguistics with Python/Scikit-Learn/Pandas", March 2014
17. N. Chakrabarty, "A Machine Learning Approach to Author Identification of Horror Novels from Text Snippets", Towardscience.com, January 2019
18. Kadriu, L. Abazi, "A Comparison of Algorithms for Text Classification of Albanian News Articles", Zagreb, September 2017
19. Vasili, E. Xhina, I. Ninka, Dh. Terpo, "Senti-

ment Analysis on Social Media for Albanian Language", Tirana, 2021

20. A.Kadriu, L.Abazi, H. Abazi, "Albanian Text Classification: Bag of Words Model and Word Analogies", Tetovo, 2019
21. B.Kabashi, "A Lexicon of Albanian for Natural Language Processing", 2018
22. M.Axhiu, "Language challenges in aspect-based sentiment analysis: A review of Albanian language", June 2019
23. T. R. R. Raju Dara, "Authorship Attribution using Content based Features and N-gram features," International Journal of Engineering and Advanced Technology (IJEAT), vol. 9, no. 1, pp. 1152-1156, 2019.
24. Using Latent Semantic Analysis," Notebook for PAN at CLEF, pp. 1143-1147, 2014.
25. R. U. K. M. Barathi Ganesh H B, "Author identification based on word distribution in word space," in 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, India, 28 September 2015.
26. I. S. B. S. R. Waheed Anwar, "Design and Implementation of a Machine Learning-Based Authorship Identification Model," Hindawi Scientific Programming, vol. 2019, pp. 1-15, 2019.
27. N. M. E.-M. G. Ahmed M. Mohsen, "Author Identification Using Deep Learning," in 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 2016.
28. X. W, Z. N, W. D. Chunxia Zhang, "Authorship identification from unstructured texts," Knowledge-Based Systems, pp. 99-111, 2014.
29. Beck, J., & Mostow, J. (2008). How Who Should Practice: Using Learning Decomposition to Evaluate the Efficacy of Different Types of Practice for Different Types of Students. In B. Woolf, E. Aïmeur, R. Nkambou & S. Lajoie (Eds.), Intelligent Tutoring Systems, (5091), 353–362. Springer Berlin Heidelberg.
30. Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief. U.S. Department of Educational, Office of Educational Technology. Retrieved from: Blikstein, P. (2011). Using learning analytics to assess students' behavior in openended programming tasks. Paper presented at the Proceedings of the 1st international conference on learning analytics and knowledge.
31. Centre for Digital Education (CDE) (2013). Big Data, Big Expectations. The Promise and Practicability of Big Data for Education. The Centre for Digital Education.
32. Dhawan, V., & Zanini, N. (2014). Big data and social media analytics. *Research Matters: A Cambridge Assessment Publication*, *18*, 36–41.
33. Johannes Furnkranz, "A Study using n-gram Feature for Text Categorization", Technical report OEFAI-TR-98- 30, 1998
34. Maria Fernanda Caropreso, "Statistical Phrases in Automated Text Categorization," IEI-B4-07-2000. Pisa, IT, (2000).
35. E. Stamatatos, N. Fakotakis and G. Kokkinakis, "Computer-Based Authorship Attribution without Lexical Measures", Kluwer Academic Publishers, Computers and the Humanities 35, 2001, pp 193-214.
36. Chade-Meng Tan, Yuan-Fang Wang, Chan-Do Lee, "The use of Bigrams to enhance Categorization," Inf. Process. Manage. 38(4): 529-546 (2002).
37. Moshe Koppel, Jonathan Schler, Shlomo Argamon, "Computational Methods in Authorship Attribution".
38. B. Rama Krishna, J. Ramesh, "An Efficient Self Constructing Algorithm for Text Categorization" International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 7, 2012, ISSN: 2278- 0181.

APPENDIX

Attached files:

- Authors.csv – all articles assembled used for training of the model (CSV file)
- Author_Identification_Model.ipynb – Implementation, all code in Python
- Stopwords_alb.txt – all stopwords