ISSN 2664-9969

UDC 004.8
DOI: 10.15587/2706-5448.2024.301706

Yuliia Ulyanovska,
Oleksandr Firsov,
Victoria Kostenko,
Oleksiy Pryadka

STUDY OF THE PROCESS OF IDENTIFYING THE AUTHORSHIP OF TEXTS WRITTEN IN NATURAL LANGUAGE

The object of the research is the process of identifying the authorship of a text using computer technologies with the application of machine learning. The full process of solving the problem from text preparation to evaluation of the results was considered. Identification of the authorship of a text is a very complex and time-consuming task that requires maximum attention. This is because the identification process always requires taking into account a very large number of different factors and information related to each specific author. As a result, various problems and errors related to the human factor may arise in the identification process, which may ultimately lead to a deterioration in the results obtained.

The subject of the work is the methods and means of analyzing the process of identifying the authorship of a text using existing computer technologies. As part of the work, the authors have developed a web application for identifying the authorship of a text. The software application was written using machine learning technologies, has a user-friendly interface and an advanced error tracking system, and can recognize both text written by one author and that written in collaboration.

The effectiveness of different types of machine learning models and data fitting tools is analyzed. Computer technologies for identifying the authorship of a text are defined. The main advantages of using computer technology to identify text authorship are:

– Speed: computer algorithms can analyze large amounts of text in an extremely short period of time.

– Objectivity: computer algorithms use only proven algorithms to analyze text features and are not subject to emotional influence or preconceived opinions during the analysis process.

The result of the work is a web application for identifying the authorship of a text developed on the basis of research on the process of identifying the authorship of a text using computer technology.

Keywords: *normalization, toning, lemmatization, stop word, machine learning, classical model, deep model, LSTM, GRU, web-application.*

Received date: 18.02.2024
Accepted date: 08.04.2024
Published date: 15.04.2024

How to cite
Ulyanovska, Y., Firsov, O., Kostenko, V., Pryadka, O. (2024). Study of the process of identifying the authorship of texts written in natural language. Technology Audit and Production Reserves, 2 (2 (76)), 32–37. doi: https://doi.org/10.15587/2706-5448.2024.301706

1. Introduction

Identification and analysis of necessary information in arrays of unstructured text documents written in natural language is one of the most important tasks implemented in information systems of any complexity and scale of processed data.

We live in a world that is increasingly filled with digital assistants that allow to interact with vast information resources. Part of the appeal of these smart devices is that they don't just transmit information, but also understand it to some degree, facilitating high-level communication. At the same time, the data is combined, filtered and summarized into a form that is easier to digest. Applications such as machine translators, Q&A systems, voice information transcription and text summarization tools, as well as chatbots are becoming an integral part of our lives in the computer world [1].

Traditional methods of processing based on fast and simple classical algorithms no longer correspond to the growth rate of volumes of information resources in the form of unstructured documents, mail messages, etc. That is, traditional algorithms and methods of working with text documents do not meet modern conditions.

Identifying the authorship of texts is a very useful, but at the same time rather difficult task. This prompts both to look for new and to improve old ways to solve this problem in order to facilitate its process, which leads to the appearance of many studies within the framework of this topic. For example, in work [2] various methods of classical machine learning were studied, with the help of which an experiment was conducted to identify the author of the text. This experiment used support vector machines (SVM), naive Bayes method, multilayer perceptron (MLP), logistic regression (LR), stochastic gradient descent (SGD) and ensemble

learning methods, including extra trees. Each of these methods was applied to three data sets of different sizes, which were formed from newspaper articles and included 10, 15, and 20 authors. The data vectors were formed using a statistical indicator (TF-IDF) using 1-gram and 2-gram word tokens. According to the results of this experiment, it can be said with confidence that the use of machine learning methods to identify the authorship of a text is quite an effective solution, because during the experiment it was possible to achieve the accuracy of author recognition in the region of 93–97 %, which is a rather good result.

It is also possible to highlight the work [3], which describes a system for identifying the authorship of the text, the main elements of which are blocks of stemming and embedding, classification and visualization of results. When studying the work of the system on the data of the works of Ukrainian writers, a classification result was obtained at the level of 98 % when comparing two authors in pairs.

Mathematical methods of identifying the author of the text are investigated in [4]. With the help of the method of analysis of hierarchies, the optimal method for solving the problem is chosen according to certain criteria, the method of statistical analysis for the identification of the text of Ukrainian literature.

In [5], it is about the use of a deep learning approach to solve the problem of authorship identification. The work with recurrent LSTM and GRU networks, as well as models trained on the C50 and BBC data corpus using various algorithms for data embedding, were considered in the study. As a result, it turned out that although GRU is definitely less accurate than LSTM, if it is about relatively small data sets, it is more appropriate to use GRU, because on such sets it usually outperforms LSTM and at the same time works much faster.

All this indicates that it is extremely important to look for new approaches and means of processing information resources and their integration into information systems that already exist or are being developed.

The aim of research is to investigate the possibilities and expediency of using machine learning models to solve the problem of text authorship identification, the use of Python tools, online cloud services and libraries for practical implementation. In accordance with the goal, the following research tasks were set and solved in the work:

– Get acquainted with the subject area.
– Get acquainted with research in the subject area.
– Investigate available methods for identifying text authorship.
– Investigate the process of identifying text authorship using machine learning methods.
– Investigate the effectiveness of using classic machine learning models to solve the problem of identifying text authorship.
– Develop deep machine learning models to solve the problem of text authorship identification.
– Develop an application for identifying the authorship of the text based on the conducted research.
– Test the application for identifying the authorship of the text.

2. Materials and Methods

2.1. Methodology for identification of text authorship. Identification of text authorship is the process of trying to identify the probable authorship of a document based on documents whose authorship is known [6]. It can be useful in various areas of our lives, for example, in the field of journalism in law enforcement agencies or literary studies. In the field of journalism, knowledge of authorship can be useful for verifying the reliability of published or received information. In turn, in the field of law enforcement agencies, establishing the authorship of the required text can become an important evidentiary basis for opening a criminal case. And in the field of literary studies, knowledge of the authorship of a text can help to follow the creative evolution of a writer or even discover new styles and genres.

The authorship identification process is based on the search and further analysis of stylistic features of the author's writing and speech in order to reveal features that will be unique to it. It can be anything, for example, the use of rarely used vocabulary or peculiarities of the placement of punctuation marks in the text. The main thing is that it should be a unique set of factors that would allow identifying the right author.

2.2. Problem solving. The main approaches to solving the task of identifying the authorship of a text can be considered: the use of statistical analysis methods, the involvement of experts, and the use of machine learning methods. Let's consider in more detail what the use of each of these approaches involves, highlighting their strengths and weaknesses in parallel.

The use of statistical methods is an approach that uses different statistical methods to analyze and compare different characteristics of texts, for example, the frequency of occurrence of words, phrases and sentences can be compared. Statistical methods can be divided into univariate and multivariate Univariate methods are methods of data analysis that are used in cases where there is a single measurer to evaluate each element of the sample, or there are several measurers, but each variable is analyzed separately from all others [7]. Multivariate methods are used to evaluate the data of each sample element, two or more measures are used, and variables are analyzed simultaneously [7].

Advantages of using statistical methods: it is possible to work with a small amount of texts; it is not necessary to have deep knowledge of linguistics or other related sciences.

Disadvantages of using statistical methods: it is vulnerable to changes in the author's style; does not take into account the context and semantics of the text; authors with a similar writing style cannot be distinguished.

Expert text analysis – this approach involves the involvement of linguists or other experts from related fields who are engaged in text analysis and have sufficient experience to determine the stylistic and other characteristics of the text.

The advantages of expert analysis are: high accuracy of determining authorship; the possibility of identifying unique characteristics of the author's style that cannot be detected by other methods.

Disadvantages of expert text analysis: expert assessments are subjective; it is necessary to spend time looking for experts; high price for expert services.

The use of machine learning methods – this approach consists in the use of various computer algorithms and machine learning methods to recognize the style and characteristics of the text that may indicate its author.

The approach has the following advantages:

– high accuracy of determining authorship, especially when using a large amount of text data for training;

ISSN 2664-9969

– the possibility of using for automated analysis of large volumes of text;
– the possibility of identifying unique characteristics of the author's style, which cannot be detected by statistical methods.

Disadvantages of using machine learning methods: the need for a large amount of text data for training; limitation in recognition of similar authors or co-authorship.

2.3. Authorship identification using machine learning methods. Before talking about the authorship identification process itself, it is necessary to define what machine learning is a branch of artificial intelligence and informatics that focuses on the use of data and algorithms to imitate the way of human learning [8]. This branch is focused on solving a sufficiently large number of different problems, such as classification, generation, or data analysis. It is divided into different areas of research, each of which has its own set of techniques and algorithms for solving problems of a certain type. Looking at the task of identifying authorship in the context of the field of machine learning, it can be attributed to the tasks of text classification, which are solved in the field of natural language processing (NLP), which is a field of machine learning focused on human languages. This includes both written and spoken languages [9]. Solving any natural language processing problem (in our case, it is a classification problem) can be divided into the following main stages: the stage of text data preparation; data embedding stage; stage of preparation and training of models; stage of evaluation of results.

These steps must be performed one after the other in the same order as they are presented in the list above.

The text data preparation stage is a complex process of preparing text data in order to be able to use it further in machine learning methods. This stage includes processes such as collection, normalization of text, extraction of stop words and lemmatization.

The final result of this stage is the creation of a corpus. In the field of NLP, a corpus is a set of texts, which can be a collection of movie reviews, comments or author's works [9].

The first thing to do when creating a corpus is to normalize the text – this is a process that is needed to clean the text from unnecessary information. Usually, these are punctuation marks and different case of characters. The final result of the normalization process is usually lowercase text that has no punctuation marks.

The second thing to do when creating a corpus is to remove all stop words from the normalized text – these are words that are found in a large volume of texts and do not have a special semantic load [4].

Extraction of stop words helps to facilitate further analysis of the corpus in machine learning models, increases the accuracy of the obtained results.

The third thing to do when creating a corpus is to perform lemmatization of the text. Lemmatization is a method of normalization, i. e. grouping of changed terms to their basic form determined by the corresponding part of speech in the given text [9].

The main goal of lemmatization is to reduce the dimensionality of the vocabulary space and improve the accuracy of text processing by machine learning models. After performing all the previous actions, the body will be received, which will be prepared for further work.

The data embedding stage is a complex process that transforms words or tokens of text into numerical vectors that are used for further analysis and understanding of the text by mainframe computers. The goal of this process is to represent words that are close in context and/or semantics with close numerical vectors. This stage includes processes such as toning and applying data embedding techniques. The first thing to do at this stage is to tone the data.

After the tokenization process is complete, it is necessary to choose one of the data embedding methods to convert the text into numeric vectors. The main data embedding methods are bag of words and TF-IDF.

To complete the embedding stage, it is necessary to choose and use one of the above methods, after which it is possible to receive a corpus of data that is completely ready for use in the field of machine learning.

The stage of preparation and training of models is the most important for solving any problem in any field of machine learning. This stage includes such processes as creating or selecting an existing model that may be suitable for solving the given task and training this model on the prepared data.

A machine learning model is an algorithm that has been trained to recognize certain types of patterns [10]. Models play a key role in solving any problems related to machine learning and are built to predict the probabilities of regularities or determine sequences. Machine learning models can be divided into two categories, namely classical and deep learning.

Classical learning models are models based on classical mathematical algorithms [11].

Advantages of classical learning models: well suited for small data sets; easy to use; do not require a lot of computing power.

Disadvantages of classical learning models: low accuracy on large data sets; difficult to detect complex functions.

Deep learning models are models based on the use of neural networks [11].

Advantages of deep learning models: suitable for tasks of high complexity; good accuracy on large datasets.

Disadvantages of deep learning models: they require significant computing power.

In the framework of this work, both classical and deep learning models were used, which are suitable for solving the problem of authorship identification. The work algorithms of some classic models are considered: Logistic Regression, Decision Tree Multinominal NB, Multi-layer Perceptron (MLP) [12]. In addition, the algorithm of some deep models: Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) [13] was studied.

The stage of evaluating the results helped to determine how well the model fulfills the task set before it.

The models were evaluated according to the following metrics:
– confusion matrix is a matrix that shows the number of correct and incorrect classifications for each class;
– accuracy is the ratio of the number of correctly classified examples to the total number of examples;
– loss is a measure of model error that measures how well the model predicts the output values;
– *F*-measure is the average value between precision and completeness, needed to evaluate the balance between precision and completeness.

2.4. LAV software application. The application was implemented in the Python programming language. To work with machine learning models, the Keras library was used, which is a deep learning API written in Python that runs on top of the TensorFlow machine learning platform. It was developed with an emphasis on providing fast experiments [8].

To work with data, the Pandas library, which is used to work with data sets, is used [9].

To work with Ukrainian texts, the simplemma library was used – it is a library that provides a simple and multilingual approach to finding basic word forms. And also, langdetect is a library for recognizing the language in which the text was written.

Streamlit, an open source library that makes it easy to build and share beautiful, custom web applications for machine learning and data science [10], was used to develop the user interface. This library made it possible to create a beautiful and functional web user interface without the need to go outside the Python ecosystem.

The LAV software application partially implements a general algorithm for solving problems in the field of natural language processing. But with one difference – the application does not train models, but has ready-made ones.

The application has an object-oriented structure, its functionality is divided into two main classes:

1) TextClassifier is responsible for all work with machine learning models;

2) program is responsible for working with class functions and displays the user interface.

These classes have their own set of methods.

2.5. Computational experiment. The computational experiment was performed in the Google Claboratory cloud service – this web service is aimed at simplifying the process of research in the field of machine learning. The service provides access to machines with a connected video card and a large amount of RAM, which significantly speeds up the work process. Also, the service provides an already prepared ecosystem for the research process due to the fact that the main libraries that may be needed for work in the field of machine learning are already installed by default.

The experiment was conducted using the Python language and the following libraries:

– Keras was used to work with deep learning models;

– Pandas was used to work with data;

– Simplemma was used to work with Ukrainian texts.

Libraries for working with data are also additionally applied: NLTK is a package of libraries and programs for symbolic and statistical text processing. It is used to develop programs that work with language, in particular for computational or empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning. And NumPy is a library that adds support for large multidimensional arrays and matrices, along with a large collection of functions for mathematical processing of data from these arrays.

For the presentation of graphs and charts, the following was used:

– Matplotlib is a library that allows to create graphs, charts, histograms and other data visualizations;

– Scikit-learn is a library of classic machine learning methods that includes a set of tools for classification, regression, etc.

Data is the basis for any problem in the field of machine learning, because it is based on them that a model is built to solve the given problem. As part of this work, a separate data corpus was formed, which contains the texts of six well-known Ukrainian authors. It contains the works of: Ivan Bagrianyi, Volodymyr Vynnychenko, Panas Myrnyi, Olha Kobylianska, Ivan Nechui-Levytskyi and Taras Shevchenko. The corpus is presented in CSV file format.

3. Results and Discussions

As a result of the computational experiment, four popular models of classical learning were tested, as well as work with deep learning models was carried out by creating and testing 2 models of the LSTM and GRU types.

The effectiveness of the following models was investigated: Logistic Regression; Decision Tree; Multinominal NB; Multi-layer Perceptron.

The following configuration was used to train these models: input data to the models is a corpus of authors of Ukrainian literature. The embedding methods are bag of words and TF-IDF. 1-grams and 2-grams are used as data toning methods. All model algorithms are implemented using scikit-learn library and they all work in default mode. The analysis was carried out and the models that give the best indicators were selected.

All studies were conducted with different types of data embedding, with the aim of finding the best one. Bag of words and TF-IDF methods were used during the research with two types of n-grams: 1-grams and a combination of 1-grams and 2-grams. The models were trained on the created data corpus of Ukrainian authors, which includes: a collection of works by 6 Ukrainian authors. The results are presented in Fig. 1, 2.

All studied models can be used to solve the problem of authorship identification. If to look at the classic models, then the models whose data passed through TF-IDF performed best. On average, such models have 2 % higher accuracy than similar models, but coded with a bag of words.

And using combinations of n-grams together with TF-IDF helped increase the accuracy by 1–4 %. But in the case of bag of words, on this data set the models showed even lower accuracy, it can drop from 2 % to more than 10 %.

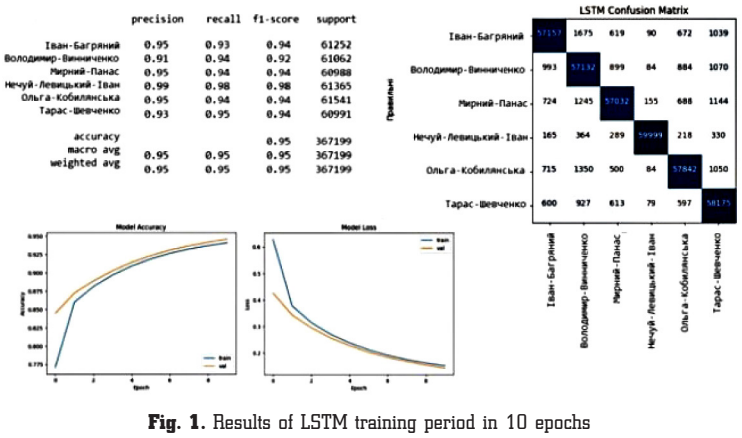

Fig. 1. Results of LSTM training period in 10 epochs

ISSN 2664-9969

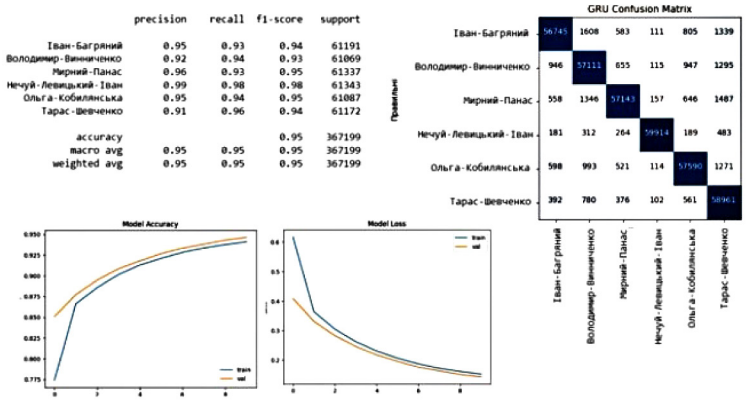

Fig. 2. The results of the GRU training period in 10 epochs

The following models showed the best results: Decision Tree – 95.80 % accuracy (Bag of words+1-gram) Multilayer Perceptron 88.58 % accuracy (TF-IDF+*n*-gram combination). As for deep learning models, they showed quite similar results. For the same period of their training, their accuracy fluctuates around 95 %. GRU 94.63 % LSTM 94.59 %. One GRU training epoch took 800 seconds on average, while one LSTM training epoch took 900 seconds on average. The GRU proved to be more effective.

Currently, texts remain the main way to receive, store or distribute information, so the need to identify their authorship has always existed and in the most diverse areas of everyday life. Despite the rapid rate of increase in the amount of information, its quality leaves much to be desired. Most of the textual information that surrounds us is unstructured, fake or unreliable, so the need to identify the authorship of textual information has become more relevant than ever before.

Many different tools have been developed to solve the problem of identifying the authorship of texts. But to solve this problem in the conditions of a large amount of information with the help of existing identification methods, it is necessary to make considerable efforts.

Increasingly, there is a need for a certain level of automation of the identification process. That is why the results of the use of machine learning technologies obtained in the course of the study are of practical importance and allow to achieve a certain level of automation of the process of identifying the authorship of texts.

Certain limitations are created by the fact that text data can be stored in different formats, and when processing such data, separate mechanisms and algorithms for preprocessing, indexing, and information extraction must be taken into account. A promising direction of further research in this subject area is the development of an algorithm for the classification of text files, which may include additional processing of texts, namely, the creation of such structures as thesauruses, evaluation of the possibility of text classification; selection of fragments using regular expressions.

Knowing about the authorship of a text provides many advantages and can influence many things. For example, it can change its general perception, calling into question the reliability of the information in this text, or help determine whether a given text has any artistic historical or scientific value. In addition, the possibility of applying the results of the linguistic examination of the authorship of textual information to detect plagiarism in student works in the conditions of distance education is interesting and promising from the point of view of further research.

4. Conclusions

It is shown that the identification of the authorship of texts is a very useful, but at the same time rather difficult task.

This prompts both to look for new and to improve old ways to solve this problem, in order to facilitate its process, which leads to the appearance of many studies within the framework of this topic [3–6].

Within the framework of this study, an analysis of the effectiveness of four classic machine learning models and two deep machine learning models based on the LSTM and GRU architecture, which are suitable for solving the task of identifying text authorship, was carried out.

In NLP terms, the task can be classified as a multicluster classification task – assigning text to two or more categories. In this case, the author of the text can be considered a category.

Based on the research results, an application was implemented in the Python programming language. Using Python provided the following benefits:

– cross-platform – allows to write the code only once and then simply use it on different platforms;
– versatility makes it possible to completely change the type of application, if necessary, Python supports writing application programs and web applications;
– easy syntax allows to be less distracted by the nuances of the language itself and to concentrate more on solving the given task;
– large database of ready-made libraries allows to use many possibilities without thinking about how they work in the middle, concentrating attention on the solved tasks.

Conflict of interest

The authors declare that they have no conflict of interest concerning this research, whether financial, personal, authorship or otherwise, that could affect the study and its results presented in this paper.

Financing

The study was performed without financial support.

Data availability

The paper has no associated data.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating this work.

References

1. Bengfort, B., Bilbro, R., Ojeda, T. (2018). *Applied Text Analysis with Python*. O'Reilly Media, Inc., 330.
2. Yülüce, İ., Dalkılıç, F. (2022). Author Identification with Machine Learning Algorithms. *International Journal of Multidisciplinary Studies and Innovative Technologies, 6 (1),* 45–50. doi: https://doi.org/10.36287/ijmsit.6.1.45
3. Lupey, M. (2020). Determining the author's affiliation of a Ukrainian-language text using a neuro-system for determining the affiliation of a text. *Science and Education a New Dimension,*

VIII (233) (28), 34–37. doi: https://doi.org/10.31174/send-nt2020-233viii28-07
4. Podshyvalenko, B. O. (2021). Zastosuvannia metodiv statystychnoho analizu dlia rozviazannia zadachi identyfikatsii tekstiv. *Radioelektronika ta molod u XXI stolitti, 7 (10)*, 65–66.
5. Gupta, S. T., Sahoo, J. K., Roul, R. K. (2019). Authorship Identification using Recurrent Neural Networks. *Proceedings of the 2019 3rd International Conference on Information System and Data Mining*, 133–137. doi: https://doi.org/10.1145/3325917.3325935
6. Zhao, Y., Zobel, J. (2007). Searching with Style. *Authorship Attribution in Classic Literature, 148*, 89–111.
7. *Statystychnyi analiz*. Available at: https://stud.com.ua/49878/marketing/statistichniy_analiz
8. *What is machine learning (ML)?* Available at: https://www.ibm.com/topics/machine-learning
9. *Slovnyk NLP*. Available at: https://medium.com/
10. *Windows Machine Learning (WinML)*. Available at: https://learn.microsoft.com/en-us/windows/ai/windows-ml/
11. Lamiae, H. (2020). *Classical ML vs. Deep Learning*. Available at: https://lamiae-hana.medium.com/classical-ml-vs-deep-learning-f8e28a52132d
12. *Scikit-learn User Guide*. Available at: https://scikitlearn.org/stable/user_guide.html
13. Lendave, V. (2021). *LSTM Vs GRU in Recurrent Neural Network: A Comparative Study*. Available at: https://analyticsindiamag-com.translate.goog/lstm-vs-gru-in-recurrent-neural-network-a-comparative-study/

Yuliia Ulianovska, *PhD, Associate Professor, Department of Computer Science and Software Engineering, University of Customs and Finance, Dnipro, Ukraine, ORCID: https://orcid.org/0000-0001-5945-5251*

Oleksandr Firsov, *PhD, Associate Professor, Department of Computer Science and Software Engineering, University of Customs and Finance, Dnipro, Ukraine, ORCID: https://orcid.org/0000-0002-6528-6447*

✉ ***Victoria Kostenko,*** *Senior Lecturer, Department of Computer Science and Software Engineering, University of Customs and Finance, Dnipro, Ukraine, e-mail: viktko@ukr.net, ORCID: https://orcid.org/0000-0003-3847-2110*

Oleksiy Pryadka, *Department of Computer Science and Software Engineering, University of Customs and Finance, Dnipro, Ukraine, ORCID: https://orcid.org/0009-0009-2835-8569*

✉ *Corresponding author*