Gender Classification of Surnames: Ukrainian aspect

Natalia Borysova[1], Karina Melnyk[1], Nadiia Babkova[1], Zoia Kochuieva[1] and Viktoriia Melnyk[2]

[1] *National Technical University "Kharkiv Polytechnic Institute", Kirpichova street, 2, Kharkiv, 61002, Ukraine*
[2] *Kharkiv general education school of I-III degrees № 145, Amosova street, 24a, Kharkiv,61171, Ukraine*

Abstract

This research focuses on resolving the problem of gender classification of Ukrainian surnames of texts' authors from GRAC corpus. An analytical review of existing solutions and classification methods for gender determination of texts have been carried out. The mathematical model of the given task has been developed. The functional model of the gender classification process has been proposed in the form of BPMN-diagram. The developed approach works with two groups of surnames: classification according to endings and classification without explicit gender features. The set of indicators for the second group according to texts' characteristics has been proposed. The efficiency of proposed classifier has been calculated.

Keywords

Gender classification of surnames, GRAC corpus, gender determination, naive Bayesian classifier, classification efficiency

1. Introduction

The paradigm change in the humanities around the world has led to the actualization of research in the study of gender aspects of language. This area is called gender linguistics. Recently, the problem of determining gender identity has become more and more urgent. There are many different classification criteria in determining gender task. Usually such criteria are various characteristics of the text, which can be automatically identified in a text. They reflect the morphological, lexical, syntactic and stylistic features of the author of the text. The problem of gender determination can be resolved in various areas of human activity, such as: authorship's expertise, banking, insurance, etc. This task is relevant in domain areas where it is impossible to analyze texts. In this case, if some personal information is available, namely a person's surname, it is easy to determine a gender using the last letters of the surname. However, there are some problems, when surnames do not have a division into the so-called "male" and "female" options. If additional information of person such as name and/or patronymic is presented, it can be used for gender determination. However, the practice of using names in different countries is different, so it is necessary to use databases of names of a certain language that already indicate gender. Automation of such processes requires the involvement of natural language processing methods.

This study proposes an approach to solution of the problem of determining gender for some authors of the General Regionally Annotated Corpus of Ukrainian (GRAC). GRAC is the Ukrainian language corpus with a volume of more than 650 million tokens. It is designed for linguistic research in grammar, vocabulary, history of the Ukrainian literary language, as well as for use in compiling dictionaries and grammars. The developers of the corpus are Maria Shvedova and Vasyl Starko. GRAC contains texts of various genres, styles, topics, regions [1, 2]. To date the gender affiliation of 23089 authors of this corpus is unknown. Thus, the purpose of the study is to develop approach for resolving the given issue.

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12-13, 2022, Gliwice, Poland
EMAIL: borysova.n.v@gmail.com (N. Borysova); karina.v.melnyk@gmail.com (K. Melnyk); nadjenna@gmail.com (N. Babkova); aliseiko@gmail.com (Z. Kochuieva); v13121423@gmail.com (V. Melnyk)
ORCID: 0000-0002-8834-2536 (N. Borysova); 0000-0001-9642-5414 (K. Melnyk); 0000-0002-2200-7794 (N. Babkova); 0000-0002-4300-3370 (Z. Kochuieva); 0000-0003-2958-3935 (V. Melnyk)

2. Formal problem statement

Let's consider the wording of the gender determination task of Ukrainian surnames of the authors and their texts. This task is a classification task in general terms. The object of classification is a set of Ukrainian surnames of the authors of the GRAC corpus with unknown gender. Each record of the list of authors is subject to the following rule: authors' surnames and initials should be presented in Latin only. Foreign names have been removed at the pre-processing stage. Therefore, input data is the various characteristics and indicators of the texts and surnames. Let designate $C = \{c_1, \dots c_n\}$ as a set of such characteristics. The result of the gender determination task is two classes: male and female. Let designate $G = \{g_1, g_2\}$ as a set of possible classes. Therefore, the gender classification task is a mapping of one set to another $f: C \to G$.

This issue can be divided to the following tasks:

- form a set of input indicators for the resolving the classification task;
- conduct an analytical review of approaches for resolving the given task;
- carry out the review of mathematical classification methods and choose suitable method;
- develop the model of the gender classification of surnames;
- assess the proposed approach of resolving the classification task.

3. Literature Review

3.1. Review of related works

Software that automatically recognizes and classifies people based on the electronic footprint that the user leaves on the network has become increasingly widespread recently. The rapid growth of the Internet has created many ways to share information across time and space. Social networks (Twitter, Myspace, Facebook), e-commerce (eBay, Craiglist), newsgroups and other sites allow to accumulate large amounts of data about users and the surrounding space. Gender information is no longer required when registering in some companies. It is expected, that the percentage of users without a declared gender will increase. Some tasks, such as testing the accuracy and efficiency of machine learning algorithms for recommending content need the user's gender. Therefore, it is necessary to determine the gender of registered users who do not disclose it, using the information collected during their registration and/or their further activities. The surname is the main source of information about a person, so it is the main factor in this study. There are two types of surnames in Ukrainian:

- surnames with gender characteristics;
- surnames without clearly defined gender characteristics.

The problem of determining the gender of the author of a written text cannot be solved without basic knowledge about the system of linguistic gender markers: linguistic units, models, structures that characterize the language in the gender aspect. Early studies of the characteristics of men and women speech behavior on the materials of various linguistic cultures show the following: discursive markers of male dominance in language practices, lexical indicators of gender specificity, differences in strategies and tactics of communication, etc. The nature of communication between homogeneous and heterogeneous groups of communicants in the gender aspect also differs [3, 4]. The works [4-7] demonstrate communicative models, which allow highlighting signs of typically male and typically female speech.

A review of a wide range of gender studies allows concluding that researchers usually analyze one type of discursive realization of communicatives: they study either spoken or written business, artistic speech, but the question of the stability/variability of signs of gender opposition in different types of speech situations is not raised. However, this problem is very significant, because its solution also requires an adjustment methodology, specific procedures for gender studies.

An analysis of available sources of information has shown that there are many solutions for determining a gender identity used by full name (last and first name and patronymic) or last and first name based on the various methods. The article [8] describes the usage of the statistical method for determining a gender identity in detail. The research [9] indicates how to use the neural network for it. The forum [10] proposes discussion of various methods and algorithms, for example, the using of

databases with names, or the defining the gender based on the end of patronymic, etc. The website [11] describes how to use the Gender API to determine the gender of customers by the names, which they write in the registration forms. Moreover, the authors provide for free use both their programs and training datasets.

Input data of the given task is a list of authors with surnames and initials, so name and patronymic are not specified. Therefore, the usage of available solutions in solving given problem is not possible. Therefore, it is necessary to develop approach for the resolution of the given task.

3.2. An analytical review of classification methods

Aforementioned analysis of domain area allow seeing that the gender determination task of Ukrainian surnames of the authors and their texts is a classification task. It is proposed to use machine-learning methods to solve the problem. Machine learning is algorithms for independently finding solutions to various problems through the integrated use of statistical data, which one are source for creating forecasts and patterns. There are many types of issue, where the machine learning techniques is useful: regression task, classification task, clustering task, dimensionality reduction problem, and anomaly detection problem. There are three types of methods in machine learning: supervised learning method, unsupervised learning method, and reinforcement learning one. Most often, supervised learning is used to analyze text data, since algorithms of this class work faster and better with texts. With the help of machine learning, a machine classifier can be built that can recognize different classes of text. The classifier is built on a pre-labeled text corpus (training sample), in which labels are assigned to data that encode their features. Learning can be defined as identifying common patterns based on training data. The primary task is to identify features in the data that can predict the target variable (label). However, classifiers are not transparent to understanding and interpretation. Machine learning uses various technologies and algorithms. Scientists can use discriminant analysis, Bayesian classifiers, artificial neural networks, and many other mathematical methods. An analytical review of classification methods for the given issue has been conducted using a limited but representative set of objects. The finding of this process is presented in [12].

Bayes' method refers to probabilistic classification methods [12, 13]. This classification approach adopts the principle of class conditional independence from Bayes' theorem. Naive Bayesian classifier is a simple and easy to implement algorithm. This classifier is mainly used in text classification, spam identification, and recommendation systems. It process well numerical and categorical data. This classifier shows good result when the amount of data is limited in comparing with models that are more complex.

Linear regression is used to identify the relationship between a dependent variable and one or more independent variables, and is generally used to predict future outcomes [13].

The support vector machine is a linear classification method. It gives good results when processing documents. However, there may be documents that will be assigned to one class by the algorithm, but in reality, they should belong to another. Such data are called outliers because they introduce method error. Such documents are best ignored when using this classification method.

The k-nearest neighbor method, also known as the KNN algorithm, is a non-parametric algorithm that classifies data based on its proximity and association with other available data. The ease of use of this method and low computation time make it the most popular algorithm for data scientists, however, as the test set increases, the processing time increases, making it less attractive for classification problems. KNN is commonly used for recommendation engines and image recognition [12, 13].

The decision tree method refers to the logical methods of classification. In practice, binary decision trees are usually used, because in them the decision to move along the edges is carried out by simply checking the presence of a feature in the document. When the feature value is less than a certain value, one branch is selected, and when it is greater than or equal to, another branch is selected. Compared to other approaches, the decision tree approach is a symbolic (i.e., non-numeric) algorithm [12].

Every method has both advantages and disadvantages. After analyzing and comparing the aforementioned methods of classification and their results, it has been decided to use Bayesian classifier.

4. Materials and methods

4.1. Gender differences of texts

Differences between male and female languages are manifested at different levels of the language: in vocabulary, in phonetics, in grammar. In addition, there are differences in the tactics of conducting conversations. Linguists claim that most often gender differences appear at the level of vocabulary. E. A. Zemskoy, M. V. Kitaigorodskaya and M. M. Rozanova [5] note in their works that women tend to use diminutive forms, especially when talking with children and animals, namely the use of approximate designations, a tendency to hyperbolic expression and a high concentration of emotionally evaluating words. Men have a tendency to coarsen the language with lexical means, a tendency to exact nomination, the use of terms, the use of stylistically neutral evaluative vocabulary, and the active use of professional knowledge outside the sphere of professional communication. In addition, women widely use adjectives and adverbs that express a general positive assessment. Such typically female evaluative words are adjectives and corresponding adverbs: wonderful, magical, amazing, unsurpassed, beautiful.

According to O. Espersen, women are more prone to euphemisms and less to obscene expressions compared with men. They are also more conservative in their use of neoplasms in the language [14].

Women use more pronouns, verbs, particles than men do. Women tend to use emotional qualities of objects and states, while men prefer concrete nouns. Men use qualitative adjectives mostly in the highest degree, and not in the comparative or high degree. Women prefer to use exclamations: “ouch” is the most common one. Constructions with the pronouns “such”, yes, “which” marked with both positive and negative connotations, are entrenched in women. Women prefer to use diminutives to convey multifaceted connections with the world, while men prefer to use diminutives when describing situations with children or loved ones. Men’s speech have rationalistic assessments. [15, 16].

N. L. Pushkareva [17] notes that women often use inversions, exclamation marks and questions. Their texts are characterized by detailed and expressive sentences. Sentences and texts of men are laconic, specific and less dynamic.

The study of written speech E.I. Goroshko have showed that the following features are entrenched in the male language: men more often use contractual rather than composing communication, less often use incomplete sentences, elliptical constructions, reverse word order [4, 18, 19].

According to A. V. Kirilina, women prefer to focus on their inner world, because their vocabulary contains more words that can describe feelings and emotions. They also often use verbs that can convey the emotional and psychological state of a person. [3].

Thus, various studies of the oral and written speech of men and women show that the difference in the use of language units is not an accident. The originality of the speeches of people of different sex really exists at all levels and in any language. It is noted that the gender factor is more fully realized at the level of vocabulary. Linguists claim that the female language is more emotional and expressive, it is characterized by a rare use of stylistically reduced means and vulgar vocabulary. Women prefer more detailed sentences and texts. Men tend to use professionalisms and terms in the language, coarsen the language with lexical means, and tend to concise sentences and texts. However, it should be noted that although there is a general trend of gender differentiation of language means, the above characteristics might vary depending on the communicative situation, cultural level and social status of the speaker.

4.2. The forming process of input data

Let’s consider the process of forming input data based on the aforementioned analysis on the example of the Ukrainian anthroponyms. Anthroponyms or proper personal names are an important part of every language. Native speakers, as well as people who study the language, use surnames, first names, patronymics, and sometimes nicknames to distinguish people in society. Anthroponyms are necessary factors of verbal communication. Their importance lies in the fact that they contain important theoretical-linguistic, historical, ethnographic and other scientific and everyday information [20]. In Ukraine, the three-term name of a person is most common: last name, first name and patronymic.

It is obvious that all Ukrainian surnames can be divided into two groups: surnames that have explicit gender characteristics (for example, Markov and Markova, Svintsytsky and Svintsytska, Isayev and Isayeva, Sanin and Sanina, Berezivsky and Berezivska, Kutsev and Kutseva etc.), and surnames without such features (Shevchenko, Potebnya, Solomka, Plaksiy, Melnyk, Movchan, Sklyar, Stelmakh, Markovych, Koval etc.). This division determines the solution of the problem of gender classification of surnames separately for these two groups.

The most important stage in the resolving of the classification task is the forming of the input data. Therefore, the classification's features of the first group are the last four letters of the surname.

The input data of the second group of the surnames are the various characteristics of the texts written by the authors with these surnames. The analytical review of the literature sources [20-26] would determine the set of texts' characteristics. The main advantage of the set that the features do not depend on the context and have a linguistic interpretation. Chosen classification features have been divided into five groups.

The first group is the frequency indicators of using the punctuation marks and special characters (comma, period, exclamation point, question point, parentheses, dashes, quotation marks, single quotation marks, colon, and semicolon). Each sign are characterized by the following indicators:

- the number of occurrences of the signs in the text is divided by the total number of sentences;
- the number of sentences with the particular sign is divided by the total number of sentences;
- the number of occurrences of all signs is divided by the total number of sentences;
- the number of sentences with at least one sign is divided by the total number of sentences;
- the average number of different signs in a sentence;
- the breadth of the author's use of punctuation marks: the maximum number of different signs in sentences is divided by the number of different signs.

These features cannot be consciously controlled by a person, in contrast to the syntactic and semantic characteristics of the text, so their use in the author's expertise is most acceptable. At the same time, the question arises of how these parameters correlate with semantically relevant features.

Second group characterizes the frequency indicators of using the different parts of speech and their combinations. The main parts of speech and their forms are the following: noun, verb, personal pronoun, pronoun (all other types), adjective, short form of adjective, adverb, predicate ("unfortunately", "good", "bad"), introductory words, service parts languages (preposition, conjunction, particle), as well as two combinations: "adverb + adjective" and "adverb + adverb". Lists of introductory words and service parts of speech have been taken from dictionaries. The following values are calculated for each of these groups:

- the number of occurrences of a part of speech or combination in the text is divided by the total number of sentences;
- the number of sentences that contain a certain part of speech or combination is divided by the total number of sentences.

Third group indicates the length of sentences and words:

- the average length of sentences in the text that expressed in words;
- the average length of words in the text that expressed in symbols.

Fourth group of indicators shows the frequency data of using the idioms and phraseologies. They contribute to greater diversity of language and recovery. The use of idioms can indicate the age, education, mood of the person who speaks or writes. It is necessary to use the appropriate dictionaries for calculating the following indicators.

- the number of idioms in the text is divided by the total number of sentences;
- the number of sentences with at least one idiom is divided by the total number of sentences;
- the number of phraseologies in the text is divided by the total number of sentences;
- the number of sentences with at least one phraseology from the list is divided by the total number of sentences.

The last group contains indicators of the vocabulary:

- the richness of vocabulary: the number of different words used in the text is divided by the total number of sentences;
- the number of words with the frequency of appearance in the text equal to 1 and 2 is divided by the total number of sentences;

- the number of words not found in the dictionary is divided by the total number of sentences.

4.3. Model of gender classification of surnames

Let's consider the process of solving the problem of gender classification of surnames in more detailed way. Before the process starts, it is necessary to preprocess the input data. Records without surnames (for example, Biofarm, Orhanizatsiia hromadska etc.) as well as not Ukrainian surnames (for example, Dao, Wu, Aktürk, Algül etc.) have been removed from the analyzed file. Surnames in Cyrillic have been transliterated in accordance with modern rules. All extra characters except letters, spaces, hyphens, apostrophes have been removed from records. All records have been aligned with one register.

The gender classification process consists of two stages. The first one realizes the implementation of gender classification of surnames with clear gender characteristics. The model of this business process is presented in the form of BPMN-diagram in Figure 1.

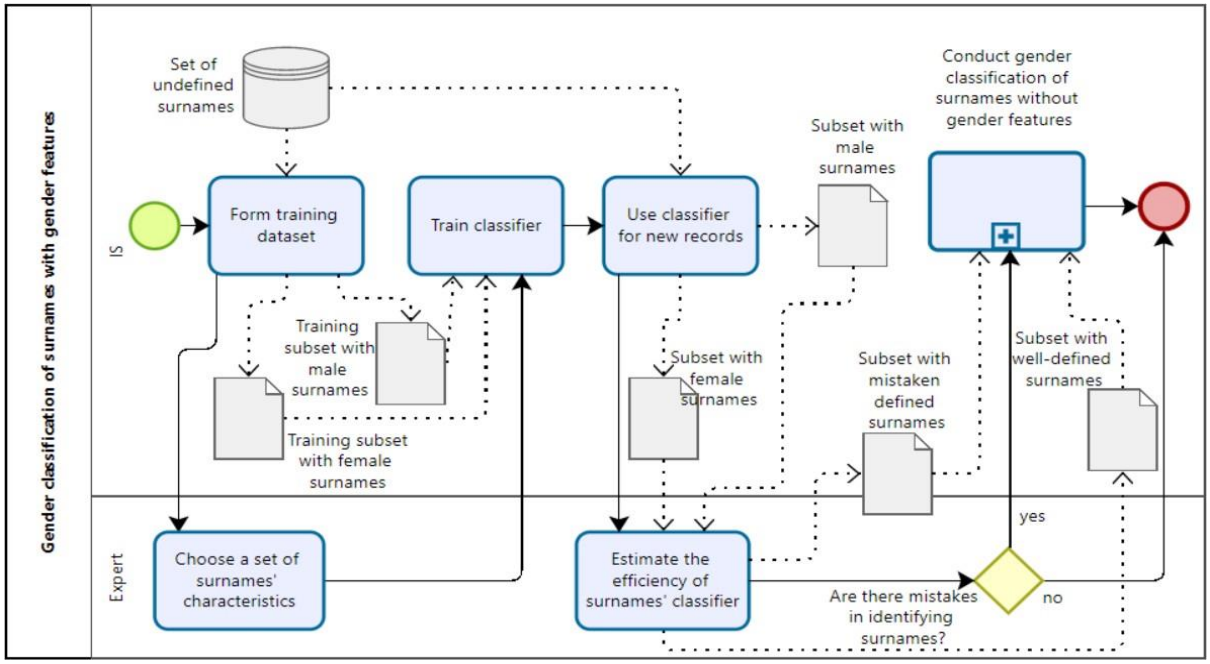

Figure 1: BPMN-diagram of gender classification of surnames with gender features

The process starts from the forming of a training dataset with female and male surnames. There are many suitable endings of female surnames: -ska, -s'ka, -skaya, -skaia, -ova, -ina, -eva, -ieva, -eeva, -yeva, -lna, -la. The endings of male surnames are the following: -iev, -kov, -sky, -nov, -shev, -skyi, -khin, -mov, -rov, -skiy, -chev, -cheev, -l'ev, -bov, -eyev, -gin, -kin, -bin, -skii, -gii, -hiy, -ckiy, -lev, -nin, -voy, -rev, -skyy, -lov, -lui, -lyy. Then the trained classifier carries out gender classification of new surnames. The findings are recorded in two files according to female and male surnames. Next step is checking the results by an expert. The efficiency of the classifier is evaluated according to the Accuracy metric. The numerical values of the Accuracy metrics are provided in the Results and Discussions section. The obtained results are used on the next stage, which is designated on the BPMN- diagram on the Figure 1 as subprocess with the label: "Conduct gender classification of surnames without gender features".

The second stage shows the process of surnames' classification without clearly expressed gender features according to authors' texts. The functional model of it is presented in Figure 2. Input data is a file with mistaken defined surnames from the first stage. To use the classifier, it is necessary to form training corpus of texts from GRAC according to the available features [27]: one subcorpus with texts written by women and another one by men. The set of texts have been formed based on the subset of well-defined surnames from the first stage.

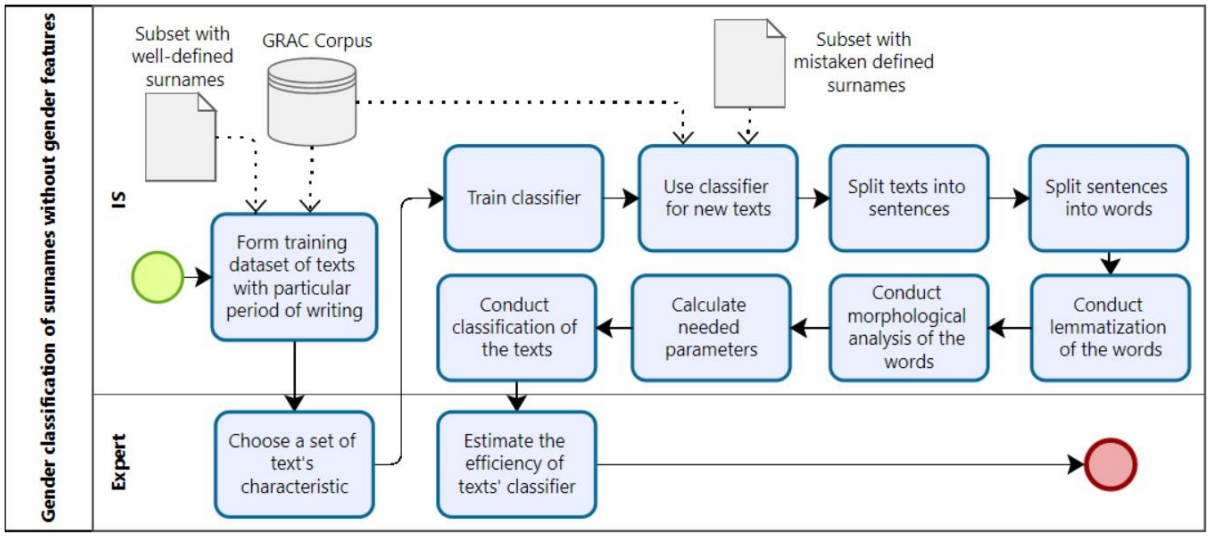

Figure 2: BPMN-diagram of gender classification of surnames without gender features

The user can use the classification in one or two stages, depending on obtained results. If a surname has clear gender characteristics, then the result after the first stage will be satisfactory. If the surname does not have clear gender characteristics and it was classified inaccurately, the user can use the function of analysis of a person's text. If the first stage do not give the desired result, then the second stage will clearly reveal the gender identity of the author. Due to the fact that the classification is carried out in two stages, the analysis of efficiency is also carried out in two stages. The efficiency of the first stage, which classifies only surnames, is higher.

Furthermore, the time of writing of the texts should be taken into account. To define gender identity of authors of some period, it is necessary to use a base of texts with such period of writing. Usage of texts from another time-period leads to unreliable classification results, since the linguistic characteristics of the text significantly depend on the writing period.

When the training corpus is created, the classifier is trained and used for new unknown texts. An analyzed text is divided into sentences. There are many markers of the end of a sentence: a period, an exclamation mark, a question mark, a newline, a tab. Then classifier conducts the following actions for each sentence: calculation of the number of punctuation marks, dividing the sentence into words, lemmatization and morphological analysis, calculation of the number of different parts of speech, counting the number of idioms and phraseologies (additional dictionary is also available in GRAC [2]), counting the number of words with errors. The obtained information is the base for the calculating of the classification indicators of the whole text.

According to the functional model of resolving the problem of gender determination, this task is a classification task. It is proposed to use naïve Bayesian classifier as the classification method, because it has showed the best results [12]. Both stages of functional model of gender classification task of surnames use classifier. Input data for the first stage is gender features of surnames. If we get dataset of surnames without clearly expressed gender features, it is necessary to resolve classification task, where the output classes should be two groups of texts depending on the gender of an author. The training and using the classifier for the both stages is similar to each other. Consider the process of train and use the naïve Bayesian classifier on the example of resolving the text classification task in more detailed way.

Let designate the following notation:

- C is a set of indicators that describe the text;
- $D_c, c \in C$ is a set of values of the c-th indicator;
- T is a set of texts;
- f_{cd} is d-th value of c-th indicator, $c \in C, d \in D_c$;
- f_{cd}^t is d-th value of c-th indicator in t-th text, $c \in C, d \in D_c, t \in T$.

It is necessary to determine the gender of the author for each text, so designate G as output class, $G = \{g_l\}, l = \overline{1,2}$, where g_l – l-th value of G-th class. The set of texts is divided into two sets, so $T = \cup_{l=1}^{2} T_l$, where T_l is a set of texts of l-th class.

Let introduce notation for the number of occurrences of certain values of indicators:

- x_{cd} is the number of occurrences d-th value of the indicator f_{cd}, where $x_{cd} = \sum_{t \in T} f_{cd}^t$;
- x_{cd}^t is the number of occurrences d-th value of the indicator f_{cd}^t;
- y_l is number of occurrences g_l as value of the output class G.

Taking into account the aforementioned notations, the general view of the algorithm for using the Bayesian classifier to determine the gender of texts is presented in the form of an activity diagram in Figure **1**3.

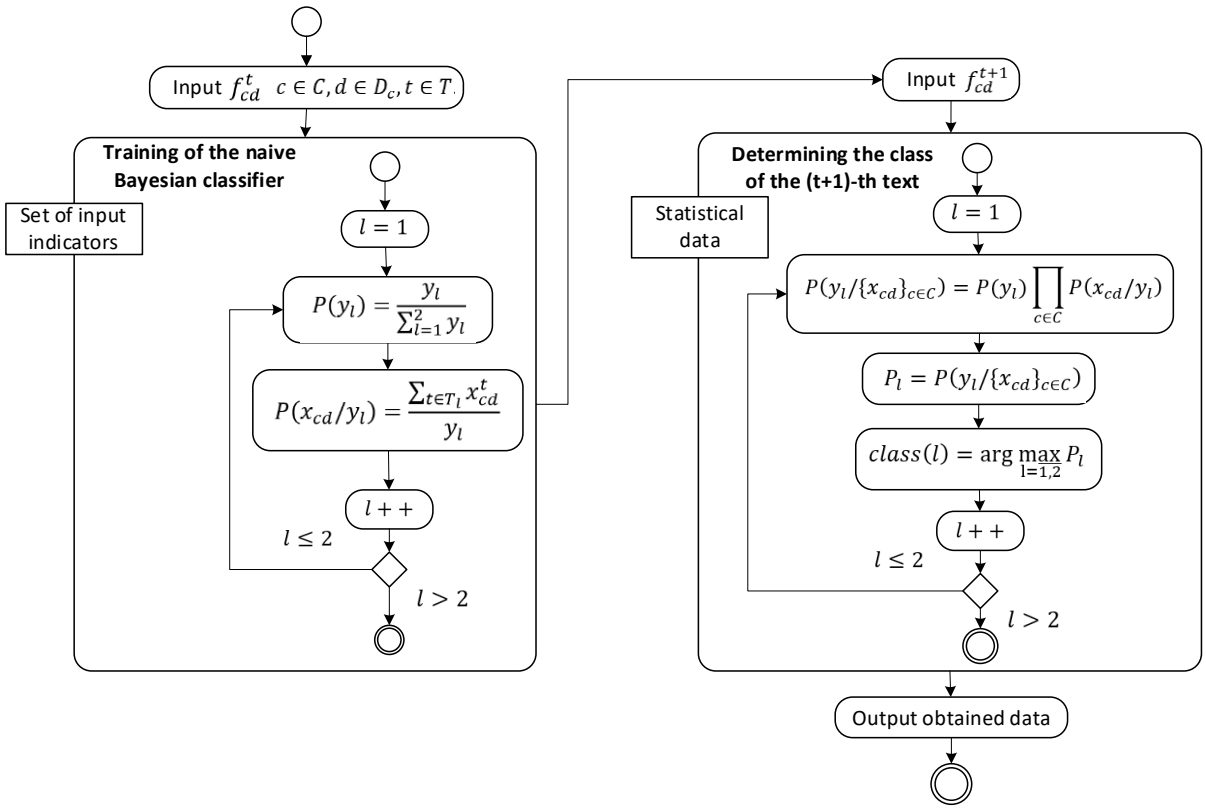

Figure 3: Model of determining gender based on the use of the Bayesian classifier

The algorithm consists of two stages: the first stage is responsible for training the Bayesian classifier, the second one describes the process of using the Bayesian classifier, which was trained for the previous steps. Thus, the mathematical model for solving the given problem is presented.

5. Results and discussions

Aforementioned research has proposed to use different metrics for estimating the efficiency of the developed classifier. It depends on the problem being solved. The Accuracy metric has been chosen for efficiency estimation of the task of classification of surnames with gender features. The Precision and Recall metrics were chosen for efficiency estimation of the task of classification of surnames without gender features.

The calculated values of chosen metrics for preresearch are presented in Table 1. The classifier used the dataset for the first stage, which contains 500 male and 500 female surnames. The structure of training set and result of using the set are the following: 1163 surnames from the training dataset were identified as male (1143 of them were identified correctly) and 1156 surnames were identified as female (1005 of them were identified correctly). For the second stage, classifier was trained on subcorpuses consisting of text written by 30 female authors and 30 male authors. 850 texts were taken for analysis.

Table 1
Metrics values for preresearch

Metrics	Task of classification of surnames with gender features		Task of classification of surnames without gender features	
	Male surnames	Female surnames	Male texts	Female texts
Accuracy	0,98	0,87	-	-
Precision	-	-	0,87	0,92
Recall	-	-	0,88	0,91

The values of efficiency metrics for main research are shown in Table 2. At the first stage of the main research all available surnames were analyzed. The result of the second stage is the analysis of all texts of necessary authors. The distinctive feature of the second stage is the values of Precision and Recall metrics are calculated only for the texts written by authors with known gender. It is strict rule, because only in this case it is possible to determine the values of needed parameters used to calculate these metrics.

Table 2
Metrics values for main research

Metrics	Task of classification of surnames with gender features		Task of classification of surnames without gender features	
	Male surnames	Female surnames	Male texts	Female texts
Accuracy	0,93	0,83	-	-
Precision	-	-	0,84	0,89
Recall	-	-	0,83	0,88

The comparative analysis of the obtained data from Tables 1 and 2 has showed, that the values of all metrics for main research has been decreased. It can be explained by the fact that the classifier makes more errors when analyzing a larger amount of data. However, the metrics values remained quite high.

6. Conclusions

The problem of gender classification of surnames is quite large-scale. It is out of linguistic boundary and requires the use of automated natural language processing methods. There is no doubt that the given task is relevant, since it requires a solution in various areas of human activity. A specialist linguist can resolve this task manually, but the opportunities of modern information technologies and big amount of information in electronic form can decrease the processing time and increase the quality of obtained data. The task of determining gender is a classification task. Many machine-learning methods have demonstrated their efficiency for resolving the given task. It was confirmed by numerous studies for different languages.

Thus, in this study, an approach for solving the problem of the gender determination of Ukrainian surnames has been proposed . The main point of the given task is undertaking the classification in two stages. The first stage has been based only on gender's features. The second stage allows conducting classification of the authors' surnames according to their texts. The set of texts with known authors and the various suitable literature sources have been analyzed for creating the set of texts' characteristics. The features from this set were divided into five groups according to meaning of different calculated numbers. The conducted research and analysis of the efficiency of the classifier showed the possibility of using the proposed approach to determine gender of surname to improve the process of determining the author's expertise.

7. Acknowledgment

The authors are sincerely grateful to the developer of the GRAC corpus Maria Shvedova for the data provided for analysis.

8. References

[1] GRAC, 2022. URL: http://uacorpus.org/
[2] M. Shvedova. (2020, Apr.) "The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorpus.org): Architecture and Functionality," in Proc. of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS'2020), vol. I: Main Conference, Lviv, Ukraine, Apr. 2020. pp. 489–506. URL: http://ceur-ws.org/Vol-2604/paper36.pdf
[3] A. V. Kirilina, Gender: linguistic aspects. Moscow (1999) (in Russian).
[4] E .I. Goroshko, A. V. Kirilina, "Gender Research in Linguistics Today", Gender Research, no. 2, pp. 234-241, Kharkiv (1999) (in Russian).
[5] E. A. Zemskaya, M. A. Kitaygorodskaya, N. N. Rozanova, "Features of male and female speech", Russian language and its functioning, Moscow, Science (1999), pp. 90-136, (in Russian).
[6] A. V. Anishchenko, "On the gender characteristics of the realization of emotional reactions", Gender: Language, Culture, Communication, Moscow (2003), pp. 18-19 (in Russian).
[7] E. G. Borisova, "The use of interjections in the speech of women and men", Gender: Language, Culture, Communication, Moscow (2003), pp. 28-29 (in Russian).
[8] Gender determination by name – when accuracy really matters, 2016. URL: https://habr.com/ru/post/274499/, (in Russian).
[9] MISexDetector. Neural network for detecting user's sex by name, 2019. URL: https://github.com/Rai220/MlSexDetector
[10] Gender determination by name, 2017. URL: https://ru.stackoverflow.com/questions/655179, (in Russian).
[11] Gender API, 2022. URL: https://gender-api.com/en/
[12] A. Shleiko, N. Borysova, Z. Kochuieva and K. Melnyk, "An overview of existing machine learning methods for gender classification of names", in Proc. of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS'2021), vol. II, Lviv, Ukraine, Apr. 2021, pp. 91-92. URL: http://web.kpi.kharkov.ua/iks/wp-content/uploads/sites/113/2021/10/CoLInS_Volume2_2021.pdf
[13] Supervised Learning, 2020. URL: https://www.ibm.com/cloud/learn/supervised-learning
[14] T. B. Kryuchkova, "Some research of the features of the use of the Russian language by men and women", Problems of psycholinguistics, Moscow, 1995, pp. 186-199 (in Russian).
[15] I. N. Kavinkina, "Diminutives as markers of the linguistic consciousness of men and women", Word formation and nominative derivation in Slavic languages, Part 1, pp. 25-31, Grodno (1998) (in Russian).
[16] N. Shkvorchenko, "Internet discourse as a linguistic category", Current issues of the humanities, V. 3, no 23 (2019), pp. 62-72(in Ukrainian).
[17] N. L. Pushkareva, "Gender Linguistics and Historical Sciences", Ethnographic Review (2001), no. 2, pp. 31-40.
[18] E. I. Goroshko, "On the question of the correlation of quantitative and qualitative methods of data analysis in linguistic gender studies", Gender: Language, Culture, Communication, Moscow, 2003, pp, 35-36 (in Russian).
[19] Yu. P. Maslova, "Features of the development of gender linguistic research in Ukraine and abroad", Scientific notes of the National University of Ostroh Academy. Series: Philological, no. 57 (2015), pp. 100-105 (in Ukrainian).
[20] N. V. Vyazigina, "Gender lingustics and diagnostics of the sex as a problem of authorship's expertise," Legal Linguistics, no. 2(13), pp. 48–53, (2013). DOI: https://doi.org/10.14258/leglin(2013)%25x. , (in Russian).

[21] E. I. Goroshko, “Features of male and female verbal behavior: (psycholinguistic analysis,” Ph.D. dissertation, the Russian Academy of Sciences, Institute of Linguistics, Moscow (1996) (in Russian).
[22] O. Goroshko, Differentiation in Male and Female Speech Styles. Budapest, Hungary: Open Society Institute Center for Publishing Development Electronic Publishing Program (1999).
[23] E. S. Oshchepkova, “Gender identification of the author from the written text: Lexical and grammatical aspect,” Ph.D. dissertation, Moscow State Linguistic University, Moscow (2003) (in Russian).
[24] A. V. Plusnina, “Characteristics of Man and Female Written Speech in Gender Consciousness of Communicators,” Yaroslavl Pedagogical Bulletin, no. 1, pp 184–188, 2012. URL: http://vestnik.yspu.org/releases/2012_1g/41.pdf, (in Russian).
[25] T. A. Litvinova, “Written text author’s characteristics ascertainment (profiling),” Philology. Theory and Practice, no. 2(13), pp. 90–94, 2012. URL: https://www.gramota.net/articles/issn_1997-2911_2012_2_29.pdf, (in Russian).
[26] T. A. Litvinova, O. V. Zagorovskaya, V. A. Chervaneva and O. A. Litvinova, “The problem of author gender attribution impact of genre,” Russian Journal of Education and Psychology, no. 1(33), 2014. DOI: http://dx.doi.org/10.12731/2218-7405-2014-1-4, (in Russian).
[27] Sketch Engine. Concordance GRAC v.11. Advanced. URL: https://parasol.vmguest.uni-jena.de/grac_crystal/#concordance?corpname=grac11