Interdisciplinary Description of Complex Systems 21(6), 607-622, 2023

AN EXAMPLE OF THE CONSISTENCY ANALYSIS OF THE CLASSIFICATION OF TEXTUAL MATERIALS BY THE ANALYST AND USING THE NAÏVE BAYESIAN CLASSIFIER

Josip Ježovita*, Mateja Plenković and Nika Đuho

Catholic University of Croatia
Zagreb, Croatia

DOI: 10.7906/indecs.21.6.6
Regular article

Received: 10 July 2023.
Accepted: 7 December 2023.

ABSTRACT

Sentiment analysis is a particular form of content analysis, and its application has become popular with the growth of Internet platforms where a wide range of content is generated. Today, various classifiers use for sentiment analysis, and in this article, we show an example of using a Naïve Bayesian classifier. The aim is to examine the consistency of classifying textual materials into a positive, negative or neutral tone by analysts and the Bayesian algorithm. The hypotheses are that there is an increase in the agreement between the two ways of classifying textual materials as (1) the complexity of the formulations and (2) the size of the learning datasets increases. Based on the results, both hypotheses were accepted, but only on certain groups of messages. Increasing the size of the learning datasets and increasing the complexity of the formulations helped the classification accuracy for messages in a positive tone, while the classification accuracy for messages in other tones was high and equal regardless of varying the parameters. Correlation analysis showed a high positive correlation between the outcomes the Bayesian algorithm classified and the tones the analyst determined (r = 0,816).

KEY WORDS

content analysis, sentiment analysis, naïve Bayes classifier

CLASSIFICATION

APA: 2240, 2260

JEL: C38

*Corresponding author, η: josip.jezovita@unicath.hr; +385 (0) 1 370 66 06;
Catholic University of Croatia, Ilica 242, HR – 10 000 Zagreb, Croatia

INTRODUCTION

CONTENT ANALYSIS

Content analysis is one of the most widely used research methods in the social sciences. It is a process of studying and parsing verbal or non-verbal content to observe its characteristics and messages [1; p.258]. It is most often used in media research and has become especially popular with the growth and application of different types and means of Internet communication. Since a large part of internet-mediated communication is public, the opinions and information exchanged in this way have a potentially important role in shaping the public sphere. They are also a valuable source of data in social research. Finding and using an adequate way of analysing such data sources is a special challenge.

There are two basic types of content analysis: qualitative and quantitative content analysis [2; p.81]. Qualitative or non-frequency analysis is based on the subjective evaluation of the analysed content, where the most important is the existence or non-existence of specific properties (instead of the frequency of their occurrence). Quantitative analysis implies a systematic and objective procedure by which it is possible to find more precise indicators. The goal of quantitative content analysis is to determine the existence of specific properties and to express them quantitatively through the degree of their representation in the analysed content. The implementation of quantitative analysis includes several research phases: defining the subject of research; formation of aims and hypotheses; defining the research population; sample selection; definition of the unit of analysis; defining criteria for quantifying the unit of analysis; and defining a content unit and constructing an analytical matrix [3; p.172]. Defining the content unit implies the selection of criteria according to which the analysis is carried out. This phase is the most sensitive part of the analytical work. The criteria must be sensitive enough to identify essential characteristics of the content. They must also be adequate, simple, and unequivocal to ensure an objective analysis, i.e., the consistency and reliability of the analysis. Objectivity is achieved through constructing an analytical matrix, which includes a greater or lesser number of analytical criteria set in relation to the selected content, analysis procedure and method of data collection. To ensure objectivity, it is also necessary to ensure a larger number of analysts who must have a certain level of education and training for this type of analysis. It is required to conduct analyst training and use the same criteria to reduce subjectivity and increase objectivity [2; p.20]. Content analysis often requires understanding the context in which the content appears. Lack of context or insufficient understanding of the context can lead to misinterpretations or a lack of deeper understanding of the analysed contents [4; p.11]. Most researchers agree that by combining quantitative and qualitative content analysis, precision and objectivity can be achieved in measuring the observable features of the studied content, but also reveal their hidden dimensions and interrelationships [1; p.259].

SENTIMENT ANALYSIS

Sentiment analysis is a particular form of content analysis. Its application has become popular with the growth of internet platforms on which a wide range of content is generated [5; p.37]. The term “opinion mining” is also used for sentiment analysis because it is a method that deals with the analysis of people's opinions, sentiments, evaluations, assessments, attitudes, and emotions towards different products, services, organizations, individuals, events, topics, etc. It is based on finding statistical or linguistic patterns in the text that reveal an attitude about something or someone [6; p.5].

The most important indicators of sentiment analysis are sentiment words or opinion words [6; p.8]. The analysed words or textual information can be classified into two groups: facts and opinions. Facts refer to the transmission of objective data, while opinions express the author's sentiment.

Opinions are subjective expressions that describe the sentiment, assessment, or feelings that individuals have towards certain entities and their characteristics [7; p.20]. Sentiment is defined as an underlying feeling, attitude, evaluation, or emotion attached to an opinion. It can be summarized through three indicators: (1) "y" or the type of sentiment – determining whether it is an objective or subjective sentiment, (2) "o" or the orientation or polarity of the sentiment – positive, neutral, or negative, and (3) "i" or the intensity or strength of the sentiment – revealing whether the analysed unit is weakly, moderately, or strongly positive or negative. Polarities enable the detection of opinions, i.e., whether individuals express a positive, neutral, or negative sentiment (revealed through the adjectives, nouns, verbs, phrases, and idioms used) according to the analysed content. Neutral sentiment is categorized as an objective category of subjective analysis or, in other words, as an opinion or idea that does not have a clear tendency and cannot be classified as positive or negative sentiment [7; p.3].

Sentiment analysis can be conducted at different levels: (1) at the level of a specific document, whereby a conclusion whether the text of the entire document leaves a positive, neutral or negative sentiment is made with regard to the selected subject of research, (2) at the level of a sentence, whereby after the analysis of all sentences we decide whether each one documents a positive, neutral or negative sentiment, and (3) at the level of a feature, word or phrase, which includes the opinions or feelings of individuals that can be identified as positive, neutral or negative [8; p. 168].

There are several types of sentiment analysis (e.g., "aspect-based analysis", "intent-based analysis", "rule-based analysis", "lexicon-based analysis"), wherein the context of this article stands out "machine-learning analysis" as sentiment analysis in which manual data entry is not required because text input transforms into vector features [9; p.4]. This can be done in two ways: supervised and unsupervised. During machine learning in the supervised algorithm, the correct documents are given, which are positive and negative, and the algorithm learns and recognizes the difference based on this. In the unsupervised case, the algorithm finds a certain structure in the documents and divides them into two or more clusters. Unsupervised learning is more difficult to evaluate, while for supervised learning, criteria can be compared. The above is achieved through partitioning (dividing data into train and test sets), where one part of the data is used for learning and the other for validating what has been learned. A training set is a data set used to train a new classifier. The test set is a data set that is used to test the classifier on data that does not appear in the training set - it is used on unseen data [10; p.33]. In this case, it is best to do circular validation to reduce the bias when dividing the data into those two sets. The result of the classification is the mean value of the results of one group, and the more reliable the classification is, the more similar the results of each group are [11; p.9].

The advantages of sentiment analysis are that no dictionary is needed, and great precision is possible when classifying sentiments [11; p.12]. On the other hand, the limitations of sentiment analysis are the complexity of language and the use of sarcasm and irony, which can be difficult for computer algorithms [9; p.7]. Polarities can have different meanings depending on the context in which they are used. This problem is especially present while using conditional and interrogative sentences, but also when using sarcastic comments because, in this case, the polarity can have an opposite meaning than usual [12; p.47]. The problem is also the possibility of misinterpretation, which can lead to wrong and unreliable conclusions. Sentiment analysis often looks at the text from a collective perspective – it does not consider the context of individual users. That can lead to incorrect conclusions when it comes to the opinions of individuals. The quality of the conclusions also depends on the quality of the data used for analysis. Sentiment analysis most often focuses on the analysis of opinions rather than fact information by one or more persons [9; p.10]. Different individuals or groups may have different experiences, interests, and worldviews, and therefore it is necessary to provide a large amount of data for sentiment analysis to be valid [12; p.47].

Different classifiers used for sentiment analysis (e.g., decision tree, support vector machine, logistic regression, Naïve Bayes classifier) are mutually exclusive. This article will present an example of using the Naïve Bayes classifier. Its advantages are simplicity of implementation, efficiency and speed, robustness towards data forest, and tolerance of incomplete data. On the other hand, the disadvantages are sensitivity to the quality of learning data and less adaptability to complex and non-linear relationships [13; pp.49-55, 14; pp.3043-3049, 15; pp.153- 159, 16; pp.525-531].

THE USE OF NAÏVE BAYES CLASSIFIER IN SOCIAL RESEARCH

Recently, the possibility of using new statistical tools in social research has been discussed. In the methodological sense, research in the social sciences is faced with limited samples, imprecise measurements, and variables that are difficult to control. They rely on statistical conclusions, and their success depends primarily on their methods. Traditional statistical tools work well if: (1) the variables have a normal distribution, (2) there is no important prior knowledge or information about the variables used and analysed in the research, and (3) the number of data is relatively large in relation to the subject of research. If these prerequisites are not met, they become "weak" and difficult to apply tools that can lead to wrong conclusions. That is especially important for the social sciences, which face several problems: (1) the used variables rarely have a normal distribution, (2) the social sciences are interdisciplinary and numerous knowledge comes from other scientific disciplines, and (3) in some research it is not easy to satisfy a representative or sufficiently large sample because this usually requires high costs [17; pp.662-664].

The Naïve Bayes classifier is a statistical tool that overcomes the limitations of traditional statistical tools. It can be used to analyse various problems that cannot be analysed with the help of conventional statistical approaches but can also serve as a complement. Compared to traditional statistical tools, the Naïve Bayesian classifier is based on probability theory (it considers hypotheses and unknown effects). That means that it can be applied to different types of distribution, contains relevant prior information (uses previous evidence to solve problems), and can be applied to any sample size [18; p.4]. More precisely, the algorithm works on the principle of determining the probability that an individual text material, based on its specific parts, belongs to one of the predefined groups or classes. Textual materials must be separated into parts before classification using the Bayesian algorithm. These parts are called independent features and can appear as individual words or formulations in a text. When it is said that properties should be independent, it means that one word (or formulation) should not be conditioned by the presence of other words in the textual material and that all words are equally important. This assumption is often unrealistic, hence the name "naïve" Bayesian classifier, due to its oversimplification of the relationship between properties (words or formulations). Textual materials can be "cleaned" before disassembling them into independent features so that the algorithm can more clearly distinguish the unique properties of these materials. One way of "cleaning" is to remove punctuation marks from the text. To determine the probability of individual text materials belonging to a predetermined class, the Bayesian algorithm first needs to "learn" the characteristics of those texts within each class. It requires a learning dataset with already pre-classified text materials by class. The "learning" process takes place by observing how many times each word or formulation is repeated within a class, and based on this information, their conditional probability of appearing in a particular class is calculated. After the calculation, the probabilities obtained for each word or formulation are multiplied, which is repeated for each class. The value of the product determines the probability of the observed text material belonging to one of the classes.

In the recent literature, many examples exist of using the Naïve Bayesian classifier. We will mention some of them. Tago and Jin used Naïve Bayes in their research "Analyzing Influence of Emotional Tweets on User Relationships by Naïve Bayes Classification and Statistical

Tests" [19]. They investigated whether positive users construct their relationships actively. Words that were not in the dictionary were excluded. To solve their problem, the authors used the Naïve Bayes classification. They obtained almost the same result, and a significant difference was confirmed for the followee fluctuation, follower fluctuation, and mutual follow fluctuation. Naïve Bayes classification confirmed the results of their previous study that positive users connect with other users not one-sided but bilaterally. Chaabi et al. used Naïve Bayes in their research "Determination of Distant Learner's Sociological Profile Based on Fuzzy Logic and Naïve Bayes Techniques" [20]. Their research is based on automatic analysis of asynchronous textual conversations. Their analysis consists of four stages: recovery, filtering, lemmatization, and message classification. The Naïve Bayes has proved to be helpful in practice because it is suited to problems of message categorization and has the advantage of being efficient in terms of processing power in the absence of standardization of speech acts and for determining the social behaviours of learners. Shaziya used the Naïve Bayes model in her research "Prediction of Students Performance in Semester Exams Using Naïve Bayes Classifier" [21] to analyse the impact of education on improving students' performance. Data Mining was used to analyse vast amounts of data from many domains. The educational data mining area is being explored, and its impact in improving the quality of education. Ernawati et al. used Naïve Bayes in their research "Implementation of the Naïve Bayes Algorithm with Feature Selection using Genetic Algorithm for Sentiment Review Analysis of Fashion Online Companies" [22]. The authors concluded that the Bayes algorithm could improve accuracy based on provided data. The accuracy of the Naïve Bayes algorithm before using feature selection was 68,5 %, while the accuracy after using genetic algorithm feature selection was 87,5 %. Jing et al. used the Naïve Bayes model in their research, "Information credibility evaluation in online professional social network using tree augmented Naïve Bayes classifier" [23]. The authors proposed an approach using Tree Augmented Naïve Bayes Classifier and PageRank algorithm to evaluate the information credibility of the user profile in online professional social networks. Bayes Classifier was used to calculate the trustworthiness probability of a user's profile based on selected components in that profile and calculate the authority of user profile information by PageRank algorithm based on other users' recommendations and endorsements. The comparison between the two classification approaches shows that the integrated approach performs better than using only Bayes classifiers. In other words, the PageRank algorithm effectively improves the performance of the Bayes Classifier. Mihaljević used the Naïve Bayes model in his research "Analysis and Creation of Free Sentiment Analysis Programs" [24]. The author concluded that programs have mostly rich options for displaying results through tables or lists containing keywords of analysis, many charts, etc. However, most programs work only in English (there is insufficient support for other languages and sentiment analysis) and are still not accurate enough to replace human interpretation, especially while analysing idioms, sarcasm, and slang. Boulitsakis-Logothetis used the Naïve Bayes model in his research "Fairness-aware Naïve Bayes Classifier for Data with Multiple Sensitive Features" [25]. The author concluded that some considerations should be considered while using the Naïve Bayes model. First, the balance between statistical parity and the accuracy of the classifier should be pursued. The author also recommends further reading on the advantages and disadvantages of group fairness in general, as well as parity, so the users could decide whether to use Naïve Bayes in their research model. There are also some limitations of the Naïve Bayes model. The algorithm of Naïve Bayes does not automatically make a classification task fair when applied (it is only possible by doing extensive domain-specific investigation). Furthermore, the author also recommends reading sociological researchers where Naïve Bayes is used. Finally, he recommends identifying groups in the data using a set of observable qualities.

Despite its advantages, the Naïve Bayes classifier is still not used often enough in the social sciences for several reasons: (1) aversion to mathematics, (2) fear of writing computer syntax

(code), and (3) fear of leaving the comfort zone. However, due to the perceived advantages, scientists have developed several tools that could encourage its more frequent use, such as: replacing mathematical formulas with graphs (relationship trees), automatically generating syntax, and providing graphical visualization of models, results, and diagnostic tests [17; p.666]. Thus, the Naïve Bayesian classifier can give an understanding of certain phenomena studied by social sciences while ensuring the validity of statistical conclusions. This article aims to provide an example of classifying different textual materials by introducing a Naïve Bayesian classifier. This algorithm can help classify different textual materials in a shorter time and with fewer resources. It also gives analysts the possibility of additional verification of their conclusions about the category into which they have classified a text.

RESEARCH AIMS AND HYPOTHESES

In this article, the authors provide an example of an analysis of compatibility between two approaches in text classification: classification by the analyst and by using the naïve Bayesian classifier. The research aim is to examine the consistency of the classification of textual materials into positive, negative, or neutral tones by analysts and by using the Bayesian algorithm while varying (1) the complexity of the formulations (independent properties in each message) and (2) the size of the learning datasets on which the Bayesian algorithm can "learn" how to classify text. The hypotheses are that there is an increase in the agreement between the two ways of classifying textual materials as (1) the complexity of the formulations and (2) the size of the learning datasets increases.

METHODOLOGY

POPULATION

This article aimed to examine the tone of the textual materials on Forum.hr, a public online platform where users exchanged opinions about the Covid-19 pandemic. The data source consisted of the messages posted on the "Coronavirus" block of the "Society" sub-forum, which contained various topics related to the pandemic. The data were collected using a free online tool called Web Scraper, which enabled web scraping of the messages from the Internet. The following variables were extracted for each message: (1) the topic title, (2) the publication time, (3) the author's name, and (4) the message content. That information was entered into an analytical matrix in which each separate message (post) was considered a single unit of analysis. The data set comprised 112 314 messages from 2 583 authors and 169 topics, posted from March 2020 to May 2021.

Due to the large volume of data, the data set was reduced by applying a filter based on the presence of two keywords: vaccine/vaccination and/or headquarters (civil protection). These keywords were selected based on the assumption that they would elicit diverse opinions and tones among the forum users. The filtered data set included 3 277 messages from 590 authors and 98 topics. Two analysts manually coded these messages using three tone categories (positive, neutral, and negative). The inter-coder reliability was measured by Cohen's Kappa coefficient, which yielded a value of 0,801.

The Bayesian algorithm operates on the assumption of independence among the features in a text. Therefore, the first step of the analysis was to decompose all the downloaded messages (Figure 1 – step 1) into individual words. However, this assumption might not hold for more complex expressions (combinations of several words) that could also be considered independent features (the first part of the research objective). Unlike the Bayesian algorithm, human analysts consider the broader context and dependencies among the expressions in a text.

To approximate the human perspective, the messages were decomposed into independent features using one of five methods: unigrams, bigrams, trigrams, four-grams, and five-grams. For instance, the sentence: "I don't know what to think about the vaccine that everyone keeps talking about." was decomposed as follows: (1) "I"; "don't"; "know"; "what"; etc.; (2) "I don't"; "don't know"; "know what"; etc.; (3) "I don't know"; "know what to"; etc.; (4) "I don't know what"; "what to think about"; etc.; (5) "I don't know what to"; "to think about the vaccine"; etc. (Figure 1 – step 2). The five decomposition methods resulted in five separate databases containing the population of all the analysed messages decomposed according to different criteria. Three of these databases are illustrated in Figure 1 – step 2.

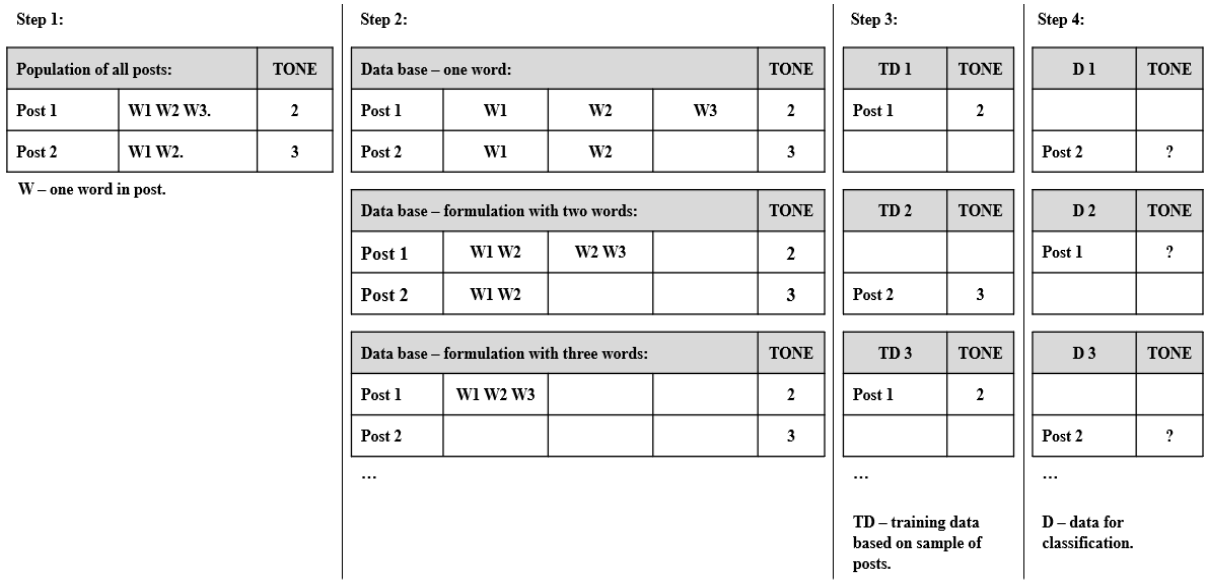

Figure 1. Scheme of the database preparation procedure for analysis.

DEFINING SAMPLES (LEARNING DATASETS)

The next step of the analysis was to define the training data sets for the Bayesian algorithm. The research aimed to test the agreement between the analyst's classification and the Bayesian algorithm's classification. Therefore, we decided to split the five previously described databases into smaller segments that would serve as the training data for the algorithm (Figure 1 – step 3). The messages in these segments were selected using a simple random sampling method but with different sample sizes to address the second part of the research objective. Specifically, nine samples were drawn from each of the five databases, ranging from 10 % to 90 % of the messages from the original databases (increasing by ten percentage points each time). Considering that we first created five databases with different levels of feature complexity and then created nine additional databases with different sample sizes from each of them using the sampling method, we obtained 45 data sets (training data sets) that served as inputs for the algorithm to "learn" how to distinguish messages based on their tone.

ANALYSIS AND PRESENTATION OF DATA

Human analysts and the Bayesian algorithm performed the tone classification of the textual materials using a combination of Excel (part of the Office 365 software package) and Rstudio (2023.03.0). The data were summarized using basic descriptive statistics, and the association between variables was measured using correlation analysis, namely, the contingency coefficient, which is appropriate for nominal-scale variables. The results were displayed in tables and graphs using line plots and contingency tables. The data analysis was conducted using the statistical program SPSS (v21).

RESULTS & DISCUSION

STATISTICS ON THE REPRESENTATION OF MESSAGES CLASSIFIED BY ANALYSTS

In the study, two analysts classified the 3 277 messages according to their tone: positive, neutral, or negative. The result was that the most common messages were of a neutral tone (71,3 %), followed by negatively toned messages with a share of 25,1 % of messages and positively toned messages with a share of 3,7 %. The fact that there is no equal representation of tones later in this article explains certain results.

STATISTICS ON THE REPRESENTATION OF MESSAGES CLASSIFIED USING THE BAYESIAN ALGORITHM

The classification of messages using the Bayesian algorithm took place in such a way that the conditional probability of its belonging to one of three tones (positive, neutral, negative) was calculated for each message. For example, if a message received the highest probability of a positive tone, it was also classified that way. However, in the classification process, there were also situations where individual messages were determined to have a zero per cent probability of belonging to any of the three tones or to have an equal probability of belonging to all tones. In these situations, it was not possible to determine how to classify the messages, and they were not included in the further analysis. As for the first results, Figure 2 shows the shares of classified messages using the Bayesian algorithm, regardless of the agreement of that classification with the analyst's classification. It can be seen that the shares of classified messages varied depending on the differences in (1) the complexity of word formulations and (2) the size of the learning datasets based on which the Bayesian algorithm learned to classify messages. For example, if the success of the classification is observed concerning the criterion of the share of messages for which the algorithm was able to determine their tone, the most successful scenario occurred on the learning datasets of the largest size (right side of Figure 2). However, there is a more pronounced difference between different scenarios if they are also observed concerning the formulation complexity parameter. More specifically, by "learning" on a learning dataset consisting of 90 % of all messages divided into three-word formulations, the algorithm classified as much as 84 % of all messages. Also, it is visible that the proportion of messages for which the algorithm succeeded in setting the tone increased continuously as the learning datasets increased, but again except for the classification modality in which the Bayesian algorithm learned on five-word formulations. Although these results still do not correspond directly to the set research aims, they indicate that the success of using the Bayesian algorithm is conditioned concerning the two observed parameters.

The following figures show even more evident patterns in changes in the share of classified messages. Results in Figures 3 to 5 are visibly similar to that in Figure 2, but the shares of messages are divided by individual tones. Several general conclusions can be drawn by comparing all three figures in parallel. Before interpreting the results in figures, an example of reading the results can be taken with modality in which the algorithm learned to classify messages on the smallest learning datasets (left side) and only based on one word (full black line). Shares of messages are as follows: neutral tone – 48,4 % (Figure 3), negative tone – 51,1 % (Figure 4), and positive tone – 0,5 % (Figure 5). Cumulatively, it is 100 % of the messages concerning the previously described classification modality.

Returning to the general conclusions, the first concerns the modality in which the algorithm learned to classify messages based on a single word only (solid black line). This classification method is the least dependent on the size of the learning dataset. In other words, no significant changes were recorded in the shares of classified messages by increasing the size of the learning datasets. Secondly, classification modalities in which the algorithm learned based on more complex formulations (two or more words) showed a greater dependence on the size of the learning datasets.

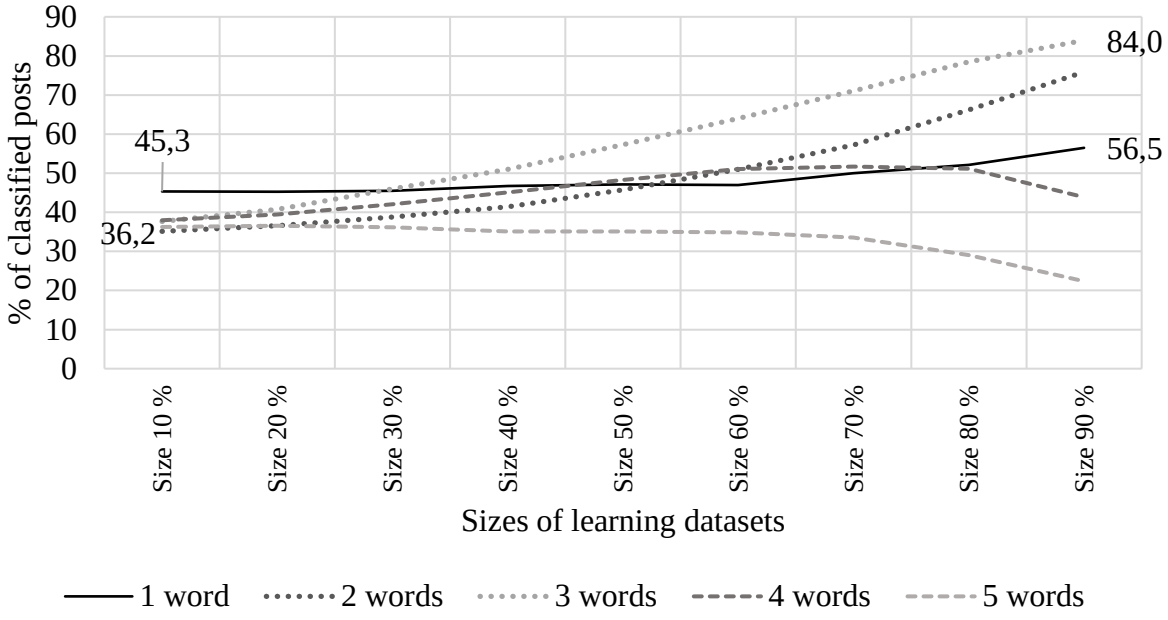

Figure 2. The percentage of all classified posts regarding the complexity of the formulations and the size of the learning datasets.

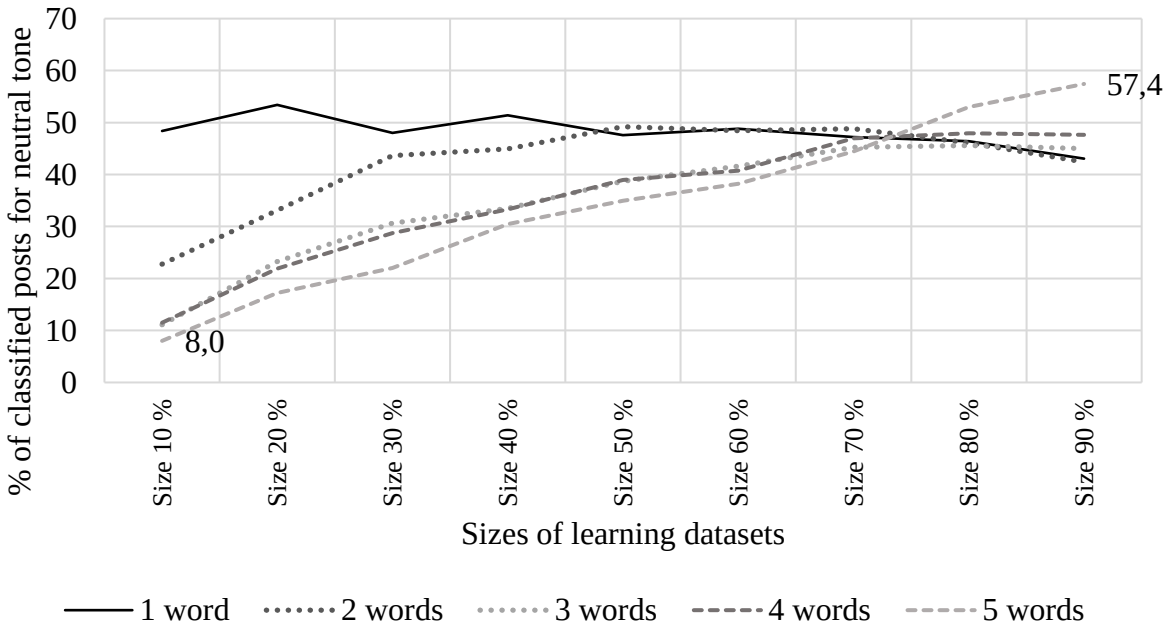

Figure 3. The percentage of classified posts regarding the complexity of the formulations and the size of the learning datasets (neutral tone).

More specifically, for neutral and negative tones, a stable increase in the share of classified messages was recorded by increasing the size of learning datasets, while in parallel, a continuous decrease in the share of positively toned messages was recorded. Thirdly, by increasing the size of learning datasets, proportions of messages per tone increasingly began to reach proportions of tones determined by analysts and which were described in previous sections. Based on the results presented so far, it can be concluded that more evident patterns are obtained for the description of changes in the share of classified messages if they are not viewed as a whole but by individual tones.

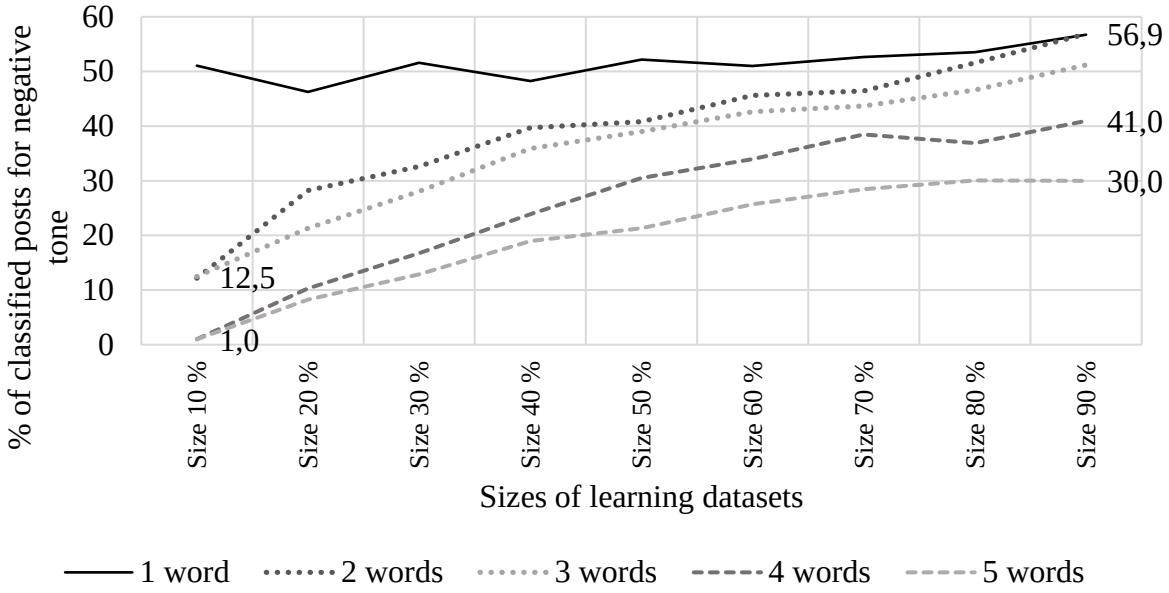

Figure 4. The percentage of classified posts regarding the complexity of the formulations and the size of the learning datasets (negative tone).

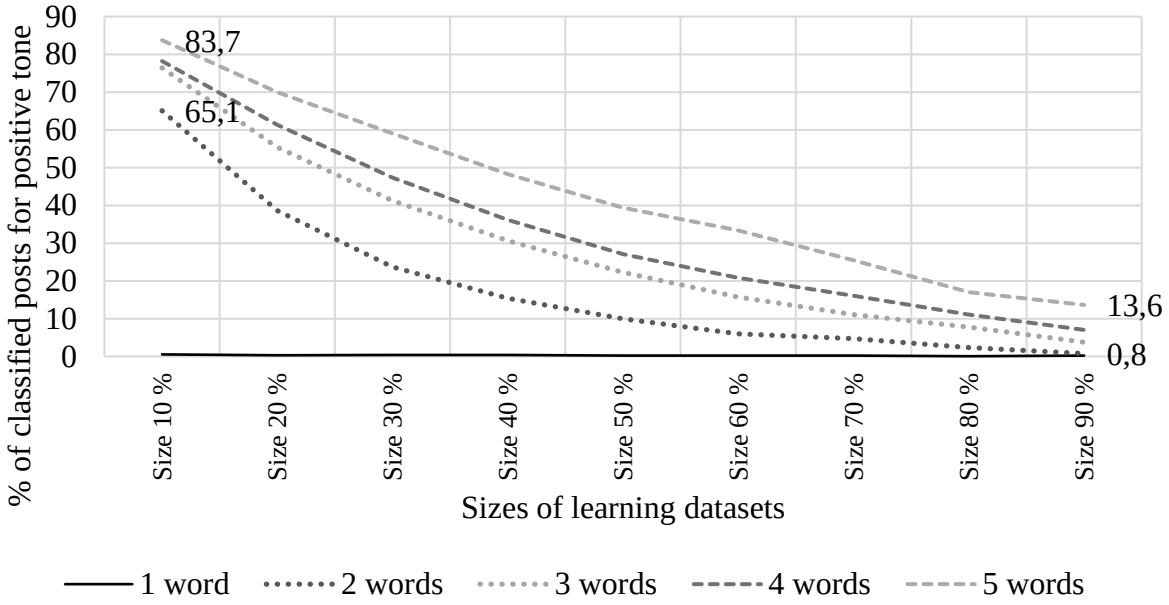

Figure 5. The percentage of classified posts regarding the complexity of the formulations and the size of the learning datasets (positive tone).

By observing Figures 3 to 5, a fourth conclusion can be drawn. There is a certain level of consistency in changes in the share of classified messages, observing these changes concerning different database sizes and formulation complexity (45 classification modalities). To gain a more detailed insight into described conclusion, the following tables show correlations between classifications made with different sizes of learning datasets (Table 1) and different levels of complexity of formulations (Table 2). It can be seen in Table 1 that the results of message classification become more and more similar as the learning datasets increase. For example, the two modalities with the most similar classification results are those that contain 70 % and 80 % of the total number of messages. Similar conclusions can be drawn by observing Table 2. It can be seen that the strongest correlations are between those classifications that were based on more complex formulations (lower right part of Table 2).

Table 1. Correlations between different size learning datasets.

	Sizes of learning datasets								
	Size 10 %	**Size 20 %**	**Size 30 %**	**Size 40 %**	**Size 50 %**	**Size 60 %**	**Size 70 %**	**Size 80 %**	**Size 90 %**
Size 10 %	1,000	0,554	0,531	0,492	0,474	0,452	0,431	0,387	0,370
Size 20 %		1,000	0,559	0,543	0,543	0,518	0,506	0,481	0,450
Size 30 %			1,000	0,580	0,582	0,582	0,563	0,529	0,510
Size 40 %				1,000	0,602	0,600	0,595	0,583	0,540
Size 50 %					1,000	0,605	0,625	0,592	0,560
Size 60 %						1,000	0,626	0,625	0,590
Size 70 %							1,000	0,627	0,590
Size 80 %								1,000	0,621
Size 90 %									1,000

Table 2. Correlations between learning datasets with different levels of complexity of formulations.

	Complexity of formulations				
	1 word	**2 words**	**3 words**	**4 words**	**5 words**
1 word	1,000	0,700	0,644	0,616	0,619
2 words		1,000	0,744	0,688	0,638
3 words			1,000	0,783	0,720
4 words				1,000	0,768
5 words					1,000

CONCORDANCE OF ANALYST-MADE CLASSIFICATION WITH BAYESIAN CLASSIFICATION

So far, results about the share of messages classified by the Bayesian algorithm have been presented. However, it has not been observed whether this classification agrees with the analyst's classification. In this part, this factor is also taken into account to be able to answer the research hypotheses. Figure 6 shows the differences between these shares. But before interpreting the results, it is necessary to explain the meaning of individual numerical values in the graph. If we recall that the Bayesian algorithm could learn how to classify messages based on 45 different learning datasets, individual values in Figure 6 represent the average or median value of the share of classified messages from those sets. For example, the value 50,7 % is interpreted as the proportion of messages for a neutral tone, which is obtained as the average of all proportions in the 45 datasets for that tone. Similarly, the value 33,0 % represents the average proportion of classified messages for that same tone but, in this case, only correctly classified messages. Figure 6 shows that the shares of correctly classified messages using the

Bayesian algorithm are lower than the share of the total number of classified messages regardless of tone. The biggest difference between the described shares was recorded for the positive tone. In that tone, the share of correctly classified messages is lower by 64 % compared to the share of the total number of classified messages. The smallest differences between shares of classified and correctly classified messages are for those in neutral tone.

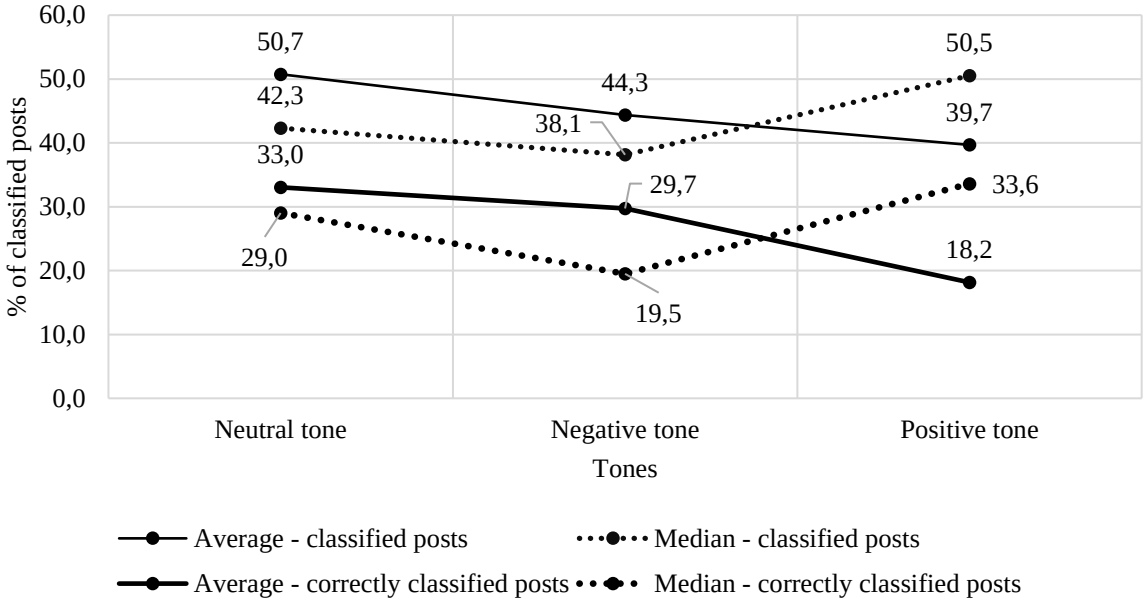

Figure 6. Percentage of classified and correctly classified posts in regards to their tone.

Finally, more detailed results exist about the share of correctly classified messages depending on different classification modalities (Figures 7 to 9). First, the shares of correctly classified messages in different classification modalities are more similar than the shares of the total number of classified messages (Figures 3 to 5). The exception is the shares of messages classified in a positive tone (Figure 9), which are similar only on larger learning datasets (right side of the display). Second, messages in neutral and positive tones remained relatively robust to changes in the size of the learning datasets (Figures 7 to 8). More specifically, no significant changes were recorded in the shares of these messages as the set sizes increased. The same conclusions are not valid for positively toned messages (Figure 9), where a decrease in their shares was recorded by increasing the size of the sets. Thirdly, the shares of messages classified in different tones follow the ratios of tones determined by the analysts for the same messages relatively well. For example, in the modality in which the algorithm learned on the largest learning dataset and five-word formulations, the following results were obtained compared to what was determined by the analysts: (1) neutral-toned messages are underrepresented by only 9,7 percentage points, (2) negatively toned messages are more prevalent by 11,4 percentage points, and (3) positive messages are underrepresented by 1,9 percentage points. Following on, correlation analysis indicated a high positive correlation (r = 0,816) between all variables describing the shares of tones classified by the analyst and the algorithm.

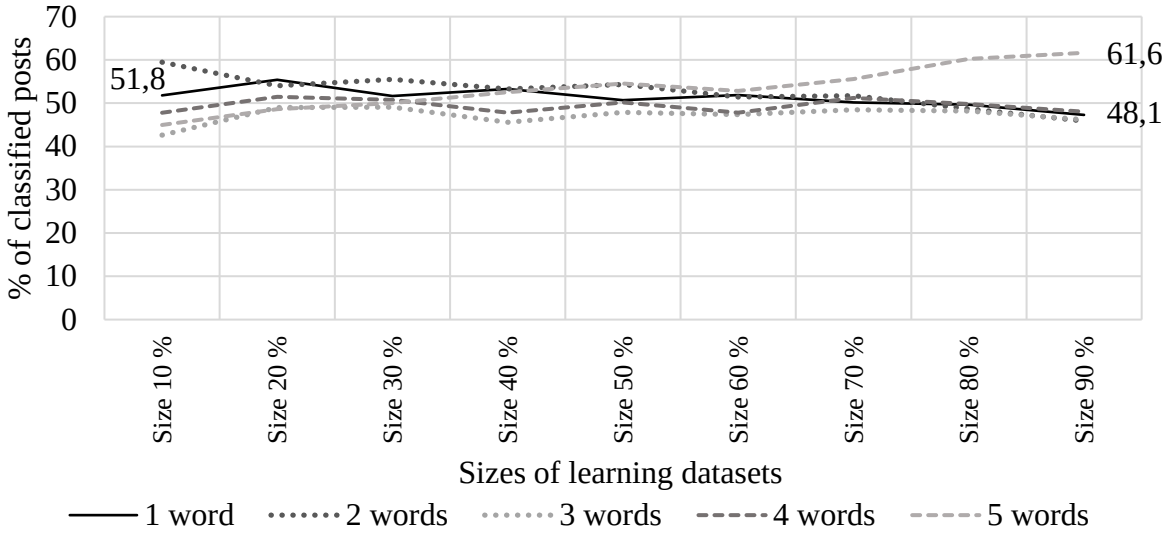

Figure 7. The percentage of correctly classified posts regarding the complexity of the formulations and the size of the learning datasets (neutral tone).

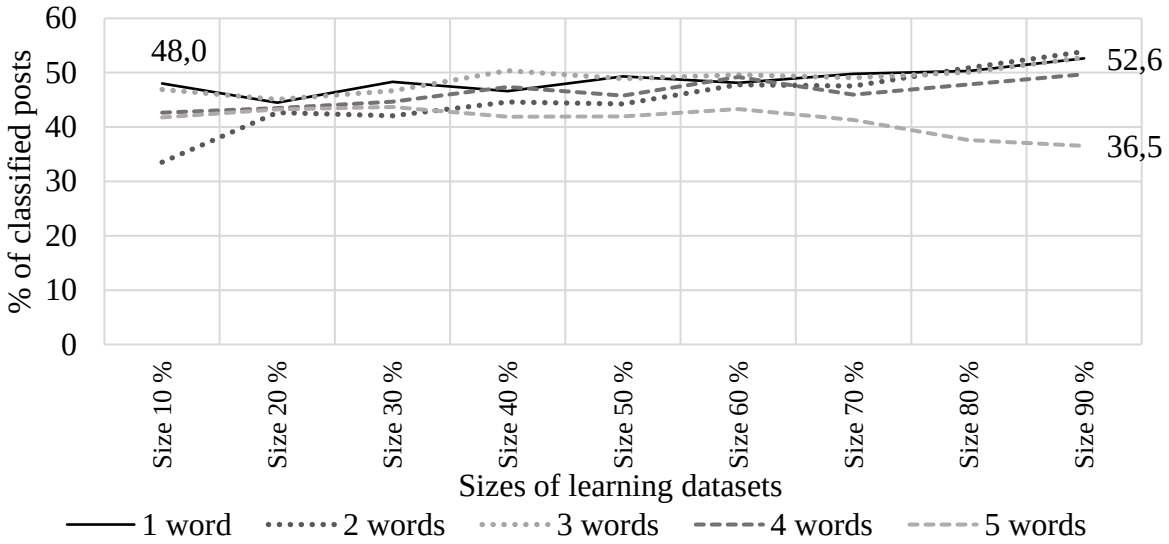

Figure 8. The percentage of correctly classified posts regarding the complexity of the formulations and the size of the learning datasets (negative tone).

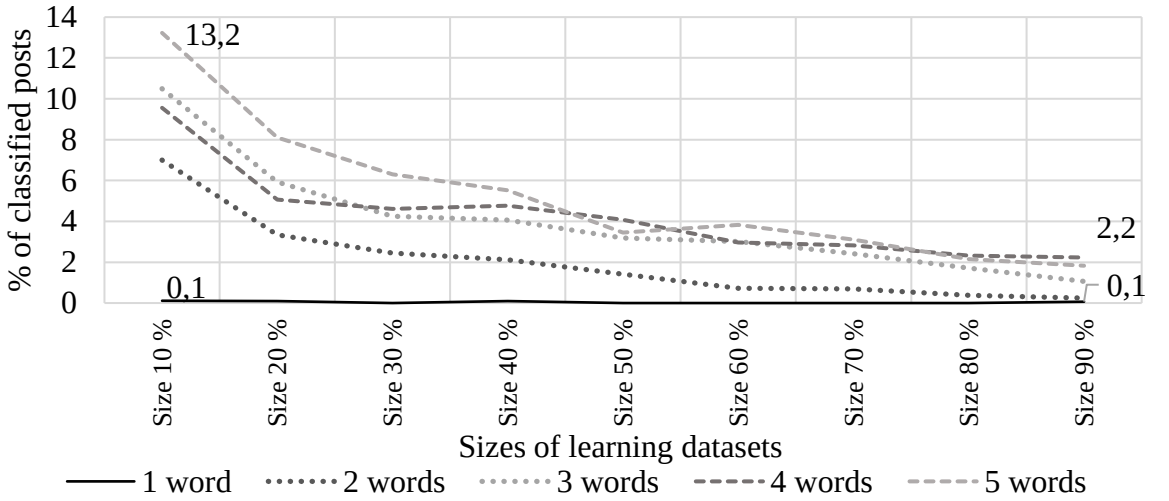

Figure 9. The percentage of correctly classified posts regarding the complexity of the formulations and the size of the learning datasets (positive tone).

CONCLUSION

This article aimed to examine the consistency of the classification of textual materials into positive, negative or neutral tones by analysts and using the Bayesian algorithm. In doing so, two parameters were varied based on which the algorithm learned how to classify messages: (1) complexity of formulations, and (2) size of learning datasets. The results first show data on the representation of messages in relation to the tone assigned to them by the analyst. Out of 3 277 messages, most of them are classified in a neutral tone (71,3 %), followed by messages in a negative tone (25,1 %), while the fewest messages are in a positive tone (3,7 %). The database of tone-specific messages served as the starting point for the application of the Naïve Bayesian classifier. Learning datasets of different sizes and with different levels of complexity of word formulations were obtained from it. For example, the simplest dataset contained only 10 % of the messages from the original database, and the Bayesian algorithm learned how to classify messages based on individual words. In contrast, the most complex dataset contained 90 % of the messages from the original database, and the algorithm learned how to classify based on five-word formulations. Based on all combinations, 45 different learning datasets were created.

After the application of the Bayesian algorithm, it was shown that there are more pronounced differences in the representation and variation of classified messages by tones if the shares of "all" classified messages are compared, regardless of the accuracy of that classification (the first group) and the share of "correctly" classified messages (second group). It turned out that the representation of classified messages from the first group, looking at them by tone, varied greatly depending on the size of the learning datasets and the complexity of the formulations. This conclusion is valid for messages for all three tones (Figures 3 to 5). The representation of messages from the second group showed a much lower level of variation compared to the previously described parameters, and this conclusion is especially valid for messages that were toned as neutral or negative (Figures 7 and 8). More precisely, the analysis showed that correctly classified messages in a neutral tone were represented by about 50 % of all messages in all classification modalities, and negatively toned messages were also represented to a similar extent. In other words, no significant changes were recorded for the described shares by varying the size of the learning datasets or the complexity of the formulations. The share of correctly classified neutrally toned messages proved to be the most stable in relation to various statistical indicators (median and arithmetic mean) (Figure 6), but this should not be surprising if we refer to the theoretical part of the article, which states how neutral sentiment is categorized as an objective category of subjective analysis. Furthermore, as for correctly classified messages in a positive tone, it was shown that their representation changed depending on the previously described parameters. More precisely, the share of these messages began to approach the share determined by the analyst (up to 3,7 %) only when the learning datasets began to increase and when the Bayesian algorithm learned to classify based on more complex formulations (Figure 9).

Based on these results, both research hypotheses can be accepted, but only on certain groups of messages. Increasing the size of the learning datasets and increasing the complexity of the formulations helped the classification accuracy for messages in a positive tone, while the classification accuracy for messages in other tones was high and equal regardless of varying the parameters. Also, the correlation analysis showed a high positive correlation between the outcomes classified using the Bayesian algorithm and the tones determined by the analyst (r = 0,816). Considering the potential reasons for the recorded differences between classes (tones), positively toned messages were represented by less than 5 % of all messages, which could have influenced their greater susceptibility to varying parameters. One of the factors that can influence the success of the classification using the Bayesian algorithm is the "quality" of independent properties in the textual materials. In classification classes with a smaller number

of textual materials, various specificities or irregularities in the text can come to the fore much more easily, ultimately affecting the algorithm's classification power. In the introductory part of the article, it was pointed out that different polarities can have different meanings depending on their context in textual material. In other words, analysts may find themselves in the problem of applying equally objective and consistent text classification criteria for all types of specific tones, which cannot be ruled out as a scenario that also happened in our analysis of individual forum posts. Nevertheless, looking at most of the posts we classified, the Bayesian algorithm confirmed our conclusions, which demonstrated the potential of applying that algorithm as an additional help or confirmation of the conclusions that analysts make by applying the classical approach to the classification of textual material.

In the end, two potential directions for further research arise from the above, which concern the issue of determining the adequate relative size of individual classes in learning datasets and the adequate quality of independent features in the observed text materials.

REFERENCES

[1] Lamza Posavec, V.: *Social Research Methodology: Basic Insights*. Institute for Social Sciences Ivo Pilar, Zagreb, 2021,

[2] Krippendorff, K.: *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Philadelphia, 2018,

[3] Lamza Posavec, V.: *Social Research Methods*. University of Zagreb – Faculty of Croatian Studies, 2004,

[4] Riffe, D.; Lacy, S. and Fico, F.: *Analyzing Media Messages: Using Quantitative Analysis in Research*. Lawrence Erlbaum Associates, New York, 2005,

[5] Johannson, M.: *Everyday opinions in news discussion forums: Public vernacular discourse*. Discourse, Context & Media **19**(1), 5-12, 2017, http://dx.doi.org/10.1016/j.dcm.2017.03.001,

[6] Liu, B.: *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, Cham, 2012,

[7] Liu, B..: *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, Chicago, 2015,

[8] Kovačević, A. and Kovačević, Ž.: *Sentiment Analysis Tools*. Polytechnic and Design **9**(3), 167-174, 2021, http://dx.doi.org/10.19279/TVZ.PD.2021-9-3-02,

[9] Sudhir. P. and Suresh V.D.: *Comparative study of various approaches, applications and classifiers for sentiment analysis*. Global Transitions Proceedings **2**(2), 205-2011, 2021, http://dx.doi.org/10.1016/j.gltp.2021.08.004,

[10] Krstić, Ž.: *Big Data and semantic analysis: Exploiting the value of unstructured data in business*. B.Sc. Thesis. University of Split – Faculty of Economics, 2014,

[11] Yassenov. K. and Misailovič, S.: *Sentiment Analysis of Movie Review Comments*. International Conference on Data Mining Workshops. 2009,

[12] Raguzin, A.: *Sentiment Analysis of Texts and Tweets Related to War and Immigrant Crises*. University of Rijeka – Faculty of Informatics and Digital Technologies, 2018,

[13] Lewis. D.D. and Ringuette, M.A.: *The Naïve Bayes Classifier: Maximum-likelihood vs. MAP estimation*. AAAI-94 Workshop on Empirical Methods in Natural Language Processing, 1994,

[14] Mahesh, P. and Mather, P.: *Support Vector classifiers for Land Cover Classification.* International Journal of Remote Sensing **29**(10), 3043-3049, 2008, http://dx.doi.org/10.1080%2F01431160802007624,

[15] Alanezi, M., et al.: *Comparing Naïve Bayes, Decision Tree and Logistic Regression Methods in Fraudulent Credit Card Transactions*.
International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy. Sakheer, 2020,

[16] Guia, M., et al.: *Comparison of Naïve Bayes, Support Vector Machine, Decision Trees, and Sentiment Analysis*.
11th International Conference on Knowledge Discovery and Information Retrieval. Vienna, 2019,

[17] Lee, M.D. and Wagenmakers, E.J.: *Bayesian statistical inference in psychology: Comment on Trafimow (2003)*.
Psychological Review **112**(3), 662-668, 2005,
http://dx.doi.org/10.1037/0033-295X.112.3.662,

[18] Jackman, S.: *Bayesian Modelling in the Social Sciences: an introduction to Markov-Chain Monte Carlo*.
Stanford University – Department of Political Science, 2000,

[19] Tago, K. and Jin, Q.: *Analyzing Influence of Emotional Tweets on User Relationships by Naïve Bayes Classification and Statistical Tests*.
10th International Conference on Service-Oriented Computing and Applications, 2017,

[20] Chaabi, Y.; Lekdioui, K. and Messoussi, R.: *Determination of Distant Learner's Sociological Profile Based on Fuzzy Logic and Naïve Bayes Techniques*.
International Journal of Emerging Technologies in Learning **12**(10), 56-75, 2017,
http://dx.doi.org/10.3991/ijet.v12i10.6727,

[21] Shaziya, H.: *Prediction of Students Performance in Semester Exams Using Naïve Bayes Classifier*.
International Journal of Innovative Research in Science, Engineering and Technology **4**(10), 9824-9829, 2015,
http://dx.doi.org/10.15680/IJIRSET.2015.0410072,

[22] Ernawati, S., et al.: *Implementation of the Naïve Bayes Algorithm with Feature Selection using Genetic Algorithm for Sentiment Review Analysis of Fashion Online Companies*.
6th International Conference on Cyber and IT Service Management, Pittsburgh, 2018,

[23] Jing, N.; Wu, Z.; Lyu, S. and Sugumaran, V.: *Information credibility evaluation in online professional social network using tree augmented Naïve Bayes classifier*.
Electronic Commerce Research **21**(6), 645-669, 2021,
http://dx.doi.org/10.1007/s10660-019-09387-y,

[24] Mihaljević, J.: *Analysis and Creation of Free Sentiment Analysis Programs*.
Media Research **25**(1), 83-105, 2019,
http://dx.doi.org/10.22572/mi.25.1.4,

[25] Boulitsakis-Logothetis, S.: *Fairness-aware Naive Bayes Classifier for Data with Multiple Sensitive Features*.
Proceedings of the AAAI Spring Symposium on Achieving Wellbeing in AI. Standford University, Palo Alto, 2022.