An Overview of Text Sentiment Analysis: a Croatian use case

Marina Mravak[1]

***Abstract*— Sentiment analysis, also known as Opinion Mining, is a very active research field, related to a branch of computer science known as Natural Language Processing, where a piece of text is categorized based on people's opinions or attitudes. This paper presents a comparison of textual data based on representation methods, also called word embedding. The comparison is performed on a dataset in the Croatian language that represents reviews marked with sentiment labels. Five pre-trained BERT models are used for feature extraction, with one model pre-trained in Croatian and the others in English. The comparison was made based on the approach of machine learning algorithms. The experiment measures the performance of the model and gives answers to the question whether the representation methods depend on the language or whether it is possible to predict the data regardless of the language with acceptable minor errors. The results of the experiment showed that it is very important to use models pre-trained in the language on which the analysis is performed, so in this case, using the CroSloEngual model, the accuracy of the prediction performance was about 91% for the SVC algorithm, which proved to be the best. However, the other models are not far behind because the smallest difference with an accuracy of 80.13% was achieved by the Logistic Regression algorithm in the AlBERT model. However, a challenge arises, where due to limited resources, especially time and memory, the analysis can be performed with a limited dataset.**

***Index Terms*— sentiment analysis, preprocessing, multilingual, BERT, Croatian language**

I. INTRODUCTION

A driver of global conversation, social media produces a large amount of user-generated data that expands daily. Data collected on various social networks (e.g. Facebook) and commercial websites (e.g. Amazon) show the exchange of user experiences, knowledge and opinions on the latest world trends based on reviews, comments or discussions that are frequently published and shared among users. Thus, social networks become a useful and powerful source of various types of sentiment. The study of sentiment analysis (SA) studies and attempts to interpret user-generated content (UGC) using computational and statistical methods. When dealing with textual data, it can be described as sentiment analysis of user opinions with the goal of understanding the polarity of sentiment (e.g. positive, neutral, negative). In the research [1], data was collected on farmers' protests in India in order to understand the sentiments shared by the public at the international level. Also, emphasis is placed on the COVID-19 pandemic, which in recent years has been a current domain in the analysis of sentiments on social networks. Furthermore, a sentiment risk index was built for the financial sector based on textual data in [2] by sentiment analysis, thus contributing to the creation of indicators for economic and financial analysis.

In order to perform sentiment analysis, various approaches and levels of sentiment analysis are addressed, and many application areas of application. The machine learning approach is most often used, so the study [1] considers different classification algorithms of supervised machine learning, and highlights the Random Forest algorithm with the highest accuracy in an experiment with labeled data. Likewise, in the study [3] a Sentiment Analysis Engine (SAE) is applied, which evaluates user sentiments in terms of positive, negative or neutral polarity based on a machine learning classification model, in order to create a Decision Support System (DSS) tool for managing promotional and marketing campaigns on multiple social media channels. Thus, supervised machine learning methods achieve remarkable performance gains for sentiment classification using labeled data. However, obtaining labeled data is expensive and requires expert effort, so unsupervised tools

[1]Marina Mravak, 13116053, Department of Data Science and Engineering

are used to improve the performance of sentiment classification from unlabeled data. The study [4] presents SSentiA, a tool that combines a machine learning classifier with a lexicon-based method, thus emphasizing a hybrid approach.

There are several challenges that sentiment analysis experts face, and one of them is sarcasm. It is implicit and characterized by the presence of positive words in a negative sense. In the study [5], an attempt was made to build a balanced system that improved the results by 8% with sarcasm features, and achieved an accuracy of 89.24% using a machine learning approach. Also, recent research [6], [7] highlights sentiment analysis of extremism, offensive speech, and complaints [8] on social networks, especially on under-researched languages, where classification algorithms exceed the performance of 82%.

Furthermore, large unstructured text datasets that include slang, misspelled words and abbreviations are cited as a problem. In a research study [9], an automated system was developed emphasizing text pre-processing and keyword extraction.

However, sentiment analysis is considered challenging, when analyzing a very large amount of textual data, a complex mixture of natural language is usually included, entailing the elaboration of text using natural language processing (NLP) techniques for SA called multilingual sentiment analysis. With this, scientists are trying to create an SA model that would include a set of different languages that would be applied in each analysis, where there would be no need for machine translation from one language to another and access to each language would be equal regardless of language resources. For this reason, many studies investigate the portability of tools that perform sentiment analysis in different languages with datasets of similar domains [10]. This is how MultiTACOS was created, trained and tested in a multilingual environment, which is also a starting point for deeper research, but very dependent on language resources.

Due to the lack of sources and lexicons in many non-English languages, integrated word extraction frameworks and tools are proposed to reduce the high training cost, time and memory consumption, and the lack of enriched and complete lexicons [11], [12] and the feature optimization problem [13], [4]. The performances of the proposed methods are acceptable from the perspective of accuracy, F1-score and execution time.

The goal of this research is the task of sentiment analysis, which will be implemented in the insufficiently researched Croatian language. Performance based on classification algorithms will be verified using data representing user reviews on the website Booking.com (https://www.booking.com/), an online travel agency for lodging reservations and other travel products, which provides online reservation services for all people around the world.

It is important to emphasize that data extraction, features and text pre-processing steps play an important role in the large amount of data offered by social networks and websites such as Booking.com. The raw form of this data can contain noise and a very large number of spelling and grammatical errors, so it is necessary to clean and pre-process the text before analysis. Transforming input text data into a fixed-length feature vector, which actually represents a text representation vector or word embedding, is one of the essential tasks of SA. Many approaches of word embedding methods are offered, but mostly used for a specific task and purpose. However, in many recent studies, they have been reinforced by available language resources such as sentiment lexicons.

For this reason, the creation of a pre-trained deep learning multilingual model for performing NLP tasks, including SA, takes place. Many SA and NLP researchers are trying to choose a suitable text representation method, especially when dealing with multilingual data, complementing and reviewing the portability of tools and models to better perform the SA task.

The significant contribution and research questions of this study can be defined as:

- explore the important pre-processing steps that affect the SA task,
- examine text representations based on the performance of SA task models using already trained multilingual models,
- based on classification measures, check the performance of different machine learning approach algorithms,
- check whether there is a significant difference in the results obtained by classification using word embedding models trained in the Croatian language compared to word embedding models trained in the English language.

In short, two important research questions are considered:

1) Is the use of word embedding methods independent of the language, without a significant difference in the performance of the SA task?
2) Is there a problem with computer resources using a large amount of data in the performance of the SA task?

After that, the focus is on the methodology, where the steps of preprocessing the collected data are described, and the models used for data representation are defined. Furthermore, prediction based on SA task and validation of prediction performance based on metrics are defined. Afterward, the results and the discussion are presented. At the end of the paper, the overall conclusion was reached.

II. Related work

Sentiment analysis is becoming increasingly challenging due to large amounts of data and multilingualism. Researchers have conducted various analyzes of the collected text datasets and used various approaches and techniques to improve performance on the SA task. This section describes related works that have been collected based on a systematic literature review [15] using several data sources where the defined query string "multilingual sentiment analysis" and ("pre-processing" or "lexicon-based" or "machine learning") executes. The following digital libraries were used to carry out the process of this research:

- IEEEXplore (https://ieeexplore.ieee.org/),
- Science Direct(https://www.sciencedirect.com/),
- SpringerLin (https://www.springerlink.com/),
- ACM Digital Library (https://portal.acm.org/),
- Google Scholar (https://scholar.google.com/).

The initial search resulted in a total of 122 articles, which is shown in Table 1. Several filtering steps were performed as in the listed review, such as removing duplicates, applying inclusion and exclusion criteria to remove irrelevant articles, and quality assessment rules to evaluate the quality of an individual article. Furthermore, as the first inclusion/exclusion criterion, the abstract and body of each paper is reviewed based on the condition that the primary studies were published in journals using multilingual sentiment analysis in the period from 2019 to 2022. Studies that did not include multilingual analysis, published at conferences and other studies irrelevant to the research questions were excluded from this study. Based on the inclusion and exclusion criteria, 24 articles were selected, on which 6 questions for quality assessment were finally performed, and the results ranging from zero to six points are shown in Table 2.

2019-2022 articles

Digital Libraries	Search Results
IEEE Explore	33 (0)
Science Direct	29 (26)
Springer Link	71 (28)
ACM Digital Library	6 (4)
Google Scholar	774 (61)
WoScc	5 (3)
Total	918 (122)

TABLE 1

NUMBER OF THE ARTICLES PER DIGITAL LIBRARY

From the above set of articles, a few of the most similar articles answering similar research questions and implementing similar methods for performing the experiment were extracted and described in this section.

In the study [10], a set of resources on policy-

Number	QA1	QA2	QA3	QA4	QA5	QA6	Total score
A3	1	1	1	1	1	1	6
A11	1	1	1	1	1	1	6
A13	1	1	1	1	1	1	6
A19	1	1	1	1	1	1	6
A22	1	1	1	1	1	1	6
A23	1	1	1	1	1	1	6
A32	1	1	1	1	1	1	6
A36	1	1	1	1	1	1	6
A37	1	1	1	1	1	1	6
A42	1	1	1	1	1	1	6
A43	1	1	1	1	1	1	6
A46	1	1	1	1	1	1	6
A50	1	1	1	1	1	1	6
A52	1	1	1	1	1	1	6
A54	1	1	1	1	1	1	6
A58	1	1	1	1	1	1	6
A120	1	1	1	1	1	1	6
A20	1	0,5	1	1	1	1	5,5
A5	1	1	1	0,5	0,5	1	5
A18	1	0,5	0,5	1	1	1	5
A33	1	0,5	1	1	1	0,5	5
A12	1	1	1	0,5	0,5	0,5	4,5
A10	1	0,5	0,5	1	0,5	0,5	4
A122	1	0,5	0,5	0,5	0,5	1	4

TABLE 2

QUALITY ASSESSMENT PER ARTICLES

related topics is provided for English, French, Italian, Spanish and Catalan that includes new and reference corpora collected for the purpose of this study. The main objective was to apply a machine learning system in a multilingual scenario to investigate the transferability of SD techniques, an attitude detection system (i.e., MultiTACOS) to different languages. This motivated the selection of datasets characterized by domain similarity (i.e. politics, election campaigns and referenda). Two datasets were already available as benchmarks developed in the context of recent evaluation campaigns, i.e. E-USA and R*-CAT, and the other two were created specifically for this study, i.e. E-FRA and R-ITA. Two datasets refer to elections (E-USA, E-FRA), while the other two refer to referendums (R*-CAT, R-ITA).

Several experiments were conducted using different classical machine learning methods exploiting four groups of features: stylistic, structural, affective and contextual to test the transferability of these features across different languages and domains. Also, the performances of three simple deep architectures (LSTM, biLSTM and CNN) based on classification algorithms (SVM and LR) were compared. It is confirmed that neural models prove to be strong in different scenarios. Also, space is left for multilingual analysis of new languages, for example non-Indo-European languages.

Due to the limited amount of Thai language research, the study [16] proposes a framework for sentiment analysis in Thai, which uses different types of features, namely word embedding, part of speech, semantic features, and all combinations of these features together with the Thai-SenticNet5 corpus. Furthermore, deep learning algorithms — convolutional neural network (CNN) and bidirectional long-term memory (BLSTM) combined in different ways — are coming together. Experimental results show that combining all three features - word embedding, POS tags and sentic and combining deep learning algorithms was able to improve the overall performance. The best hybrid deep learning was BLSTM-CNN which achieved F1-scores around 0.74.

Furthermore, the sentiment analysis of languages with limited resources, such as the Indian language Manipuri in the study [17], is investigated. Using different types of machine learning based approaches based on a dataset collected from local daily newspapers. Text preprocessing techniques such as transliteration, negative morpheme-based lexicon building, and noisy word filtering are emphasized. Based on two types of text representation: TF-IDF and Okapi BM25, different classifiers with TF-IDF representation show better results. It highlights the problem of limited availability of good Manipuri language-specific toolkits that prevent current work from including additional language features.

In order to solve the problem of slangs, misspelled words and abbreviations in the collected unstructured text dataset, in this research study [9] a new proposed system was developed which consists of four main stages: data collection, text preprocessing, keyword extraction and classification. The input data was collected from the dataset: amazon customer review. The preprocessing phase consists of three steps: lemmatization, spam detection, and removal of stop words and URLs. Using a latent Dirichlet allocation topic modeling approach, keywords were extracted and classified into three forms (positive, negative and neutral) using an efficient machine learning classifier: a convolutional neural network based on a selective memory architecture. The system increased accuracy in sentiment analysis up to 6-20% compared to existing systems, on average it achieved about 92.83% classification accuracy.

In this paper [18], a sentiment analysis based on Catalan tweets is investigated where three main steps are: Catalan language preprocessing tool, classification model and corpus training. Five models are compared to select the best algorithm: Naive Bayes, maximum entropy, support vector machine, decision tree and neural networks. However, the researchers are expanding the study and developing a public web service with the best achieved classification model where users will be able to check its effectiveness. An accuracy of 80% is achieved, and the challenges in further improving performance are outlined, emphasizing the prioritization of preprocessing tasks and the removal of the lowest-priority ones. Including a spell checker before training the model classifier could greatly help improve performance since it is very common to find word errors or abbreviations.

III. Methodology

In this section, the methodology of this research study is described, which includes data collection, the steps of textual data preprocessing, and the techniques used to represent the textual data required for defining and executing the model are presented. The models and classification algorithms that will be used for the SA task are defined.

A. Dataset

The collection of textual dataset used for this study was done using Octoparse (https://www.octoparse.com/), a cloud-based web data extraction no-code tool that helps users extract relevant information from various types of websites. Thus, it enables users from various industries to scrape unstructured data and save it in various file formats including Excel, plain text, CSV and HTML. To retrieve data, it is necessary to create an extraction task that follows the rules that will extract the required data. The task consists of the sequence of actions shown in Figure 1.

The beginning of the extraction task is defining a list of links to the Booking.com website, which provides online lodging reservation services, where the list contains about a hundred links of all types of lodging (hotels, hostels, apartments, holiday homes, etc.) located in the destinations of Split, Zagreb and Dubrovnik, Republic of Croatia. Next step is defining a loop where for each link it shows user reviews, and filtering reviews in Croatian language. After, review text extraction is performed, where the texts of positive and negative reviews are selected for extraction in separate columns. Lastly, data extraction for each link in list is processed and resulting data are saved to Excel file.

Considering that the tool is limited to performing the task on a maximum of 10 links, the task was performed multiple times with several lists of links. A total of n = 23 Excel files containing user reviews were collected and eventually merged into one Excel file with a total of n = 11 467 records, where the columns are review containing textual data, and label containing two categories that are positive (1) or negative (0) depending on the text of the review.

Fig. 1. Extraction task for collecting data

B. Data preprocessing

Data retrieved from social networks containing user reviews, comments or opinions is usually unstructured, and its raw form may contain many grammatical and spelling errors, as well as useless words or characters. Therefore, it is very necessary to clean and preprocess the text in a machine-

understandable format before analysis. Various preprocessing steps and NLP tasks were performed on the collected dataset, including removing duplicates and undefined or missing values, setting lowercase letters, removing stop words and punctuation, lemmatization and stemming, and lastly tokenization.

In order to perform preprocessing based on the collected dataset, the Python programming language was used, as well as libraries such as pandas (https://pandas.pydata.org/docs/), string (https://docs.python.org/3/library/string.html), and nltk (https://www.nltk.org/) in the Google Collaboratory (https://colab.research.google.com/).

The first step is to remove duplicate records and missing or undefined values. It was noticed that there are a total of n = 2126 duplicate entities and a total of n = 2 undefined entities, so the shape of the dataset, or the number of total reviews, was reduced to n = 9339 unique records. Furthermore, the next preprocessing step is to remove punctuation and set the text of each review to lowercase. From the string library, the string.punctuation module is retrieved, which contains a set of all punctuation. A method is created that checks for each character in each review whether it is in the punctuation set. If it is not punctuation, each character is joined to a new review and saved in a new column called wop_review in the pandas data frame.

Then, for each record in the wop_review column, the text content is set to lowercase letters, and saved in a new column called l_review. To perform the step of removing stop words, a set of words downloaded and saved in a txt file from the web keyword analyzer tools Ranks.nl (https://www.ranks.nl/stopwords/croatian) was used. Similar to punctuation, a method was created that checks for each word in each review whether it is in the list of stop words. If it is not a stop word, the words are joined into a new review and saved in a new column called wos_review.

Stemming and lemmatization are text normalization techniques of natural language processing field, where stemming is a process that stems or removes the last character of a word, often leading to incorrect meanings and spelling, while lemmatization considers the context and converts the word to its meaningful base form, which is called lemma. The stemming task was performed by using the stemming class for the Croatian language downloaded from https://github.com/dmarsic/StemmerHr, while the lemmagen3 module, which supports the Croatian language, was used for lemmatization. For each record, the mentioned techniques were applied and the obtained data were saved in new columns s_review and lem_review.

C. Word embedding

Word embeddings are mainly used as input features for models built for NLP tasks, including SA. Each model requires input text in a specific format. This study explores context-based word embeddings that capture other forms of information resulting in more accurate feature representations. The advantage of this approach over others is that it results in better model performance. The most researched models that are also implemented and compared in this paper are: Bidirectional Encoder Representations (BERT), DistilBERT, RoBERT and ALBERT, models trained on a specific corpus with a fixed vocabulary in the English language, and CroSlo-Engular model, model trained on a specific corpus with a fixed vocabulary in the Croatian language, which are retrieved from the transformers library (https://github.com/huggingface/transformers), and download and documentation of the models is available at https://huggingface.co/models.

Each model contains its own tokenizer that divides the input text reviews, into smaller units known as tokens, and before learning the model, they are converted into IDs.

Also, the pre-trained model requires that all input text be of identical length, so the maximum length parameter is defined. In this case, the length is defined depending on the number of tokens of each review. The selected value is max_length = 64. If the review contains a length smaller than the maximum length, then padding (empty tokens) is added, but if it is longer, then the tokens are shortened. Also, it is necessary to define the batch size, i.e. the number of samples or reviews processed through the model. In this case, TPU allows batch size = 4000, i.e. a total of 4000 reviews that can be passed through the model. The output of the model is the hidden state, a vector of defined hidden size for each token in the input sequence, and the last layer is taken which gives the representation vector of the input text.

D. Model definition

The pre-processed dataset is divided into training data (80%) and testing data (20%). The sklearn (https://scikit-learn.org/) library for machine learning and modeling, including classification, is used for predictive data analysis. A model for predicting review polarity was built based on the collected data, where the performance of the pre-trained models was compared depending on the language. The following algorithms were used for classification: Logistic Regression, Linear SVC, Random Forest Classifier, KNeighbors Classifier, Gaussian NB and Decision Tree Classifier and C-Support Vector Classifier (SVC) contained in the sklearn library. Finally, the predicted polarities of the reviews generated by the models and the actual polarities of the reviews were compared. Different criteria are applied to evaluate the performance of the used models. One of the most common criteria for evaluating model performance is accuracy, which is calculated based on next equation:

$$accuracy = \frac{TN + TP}{TN + FN + TP + FP}$$

where TN is the number of opinions with negative polarity that are correctly marked with negative polarity. TP is the number of opinions with positive polarity that are correctly marked with positive polarity. FP is the number of opinions with negative polarity that are incorrectly marked with positive polarity. FN is the number of opinions with positive polarity that were incorrectly labeled as negative polarity. Also, a precision metric is used that indicates the number of reviews correctly labeled as belonging to the TP positive class divided by the total number of reviews labeled as belonging to the positive class (i.e. the sum of true positives and false positives, reviews incorrectly labeled as belonging to the class), as shown in next equation:

$$precision = \frac{TP}{TP + FP}.$$

Two other metrics used to assess performance quality are recall and F-score. Next equation defines recall:

$$recall = \frac{TP}{TP + FN}$$

where is the number of true positive reviews divided by the total number of reviews that actually belong to the positive class (ie, the sum of true positives and false negatives, reviews that were not marked as positive but should have been). The F1-score is the harmonic mean of precision and recall, and is defined by the equation:

$$F1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall}.$$

IV. Results and Discussion

A. Data exploration

Before making predictions, an exploratory analysis of the collected text dataset will be explored. So, this set contains user reviews in which users expressed their opinions, observations and comments about the lodging they booked. When researching the data, questions about the number of sentences, words and the most frequent tokens in the entire corpus were interesting. The overall corpus contains 11 818 sentences and 165 850 words. The most frequent tokens are, as usual, conjunctions, prepositions and pronouns. Furthermore, the dataset contains a total of 5909 positive polarity reviews, and 3430 negative polarity reviews, as shown by the countplot in Figure 2.

In the case of positive reviews, a few of the most common nouns can be singled out, such as: staff, location, breakfast, hotel and parking, along with the most common adjectives: pleasant, kind, tidy, comfortable and tasty. Negative reviews highlight nouns such as: staff, bathroom, bed, breakfast, isolation and smell, accompanied by adjectives: terrible, uncomfortable and unpleasant.

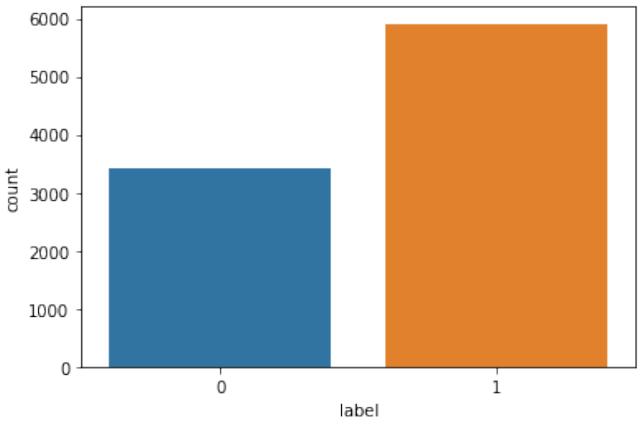

Fig. 2. Number of the reviews by polarity

B. Prediction

After text preprocessing was performed and a representative vector of reviews was obtained, the dataset was divided into training and testing, where 80% of reviews were included in the training set, and 20% of reviews were included in the testing set. After that, a prediction was made for a particular classification algorithm based on five word embedding models: BERT, DistilBERT, RoBERT, AlBERT trained in English, and CroSloEngular BERT model trained in Croatian.

As shown in Table IV-B, the best results are achieved with the CroSloEngular model where the highest accuracy is 91.25% using the SVC algorithm. Furthermore, very high accuracy is achieved with the other algorithms, where the biggest difference of 11.12%, which is obtained by the Logistic regression algorithm and AlBERTa model, stands out. Also, the accuracy of 83.37% achieved by the Decision Tree classifier, at the same time the worst using the CroSloEngular model, turns out to be better than the best result obtained using the other models. The following model with very good results stands out, where the highest accuracy of the Linear SVC algorithm was 86.38%, and the RoBERTa model is used. Other models achieved the highest accuracy using the Logistic Regression algorithm, and the highest difference of 6.12% was obtained using AlBERT. The accuracy of 80.13% is also the least successful result among the tested models.

Out of a total of 800 reviews in the test dataset, 288 reviews are marked with negative polarity and 512 with positive polarity, and looking at the precision, recall and F1-score metrics, it is obvious that the best results are achieved for reviews with positive polarity, compared to reviews with negative polarity. With the CroSloEngual BERT model and the SVC algorithm, it can be seen that of all the reviews that the algorithm predicted to be positive, a total of 92% were. Likewise, of all reviews that are truly positive, the model correctly predicted this outcome for 94% of reviews. The F1-score is 0.93, where the achieved value is very close to 1, and it says that the model is very good at predicting the polarity of reviews.

Model	Algorithm	Accuracy [%]	Precision	Recall	F1-score
CroSloEngual	SVC	91.25	0.92	0.94	0.93
BERT	Logistic regression	84.13	0.89	0.86	0.87
DistilBERT	Logistic regression	86.25	0.89	0.89	0.89
RoBERT	Linear SVC	86.38	0.89	0.89	0.89
AlBERT	Logistic regression	80.13	0.85	0.84	0.84

TABLE 3

RESULTS OBTAINED BY PREDICTION PER MODEL

Furthermore, with the CroSloEngual model, looking at the total number of predicted positive reviews that are actually positive amounts to 248 using the best SVC algorithm, while the number of predicted negative reviews that are actually negative amounts to 482. Also, there are predicted positive reviews that are actually negative in total 30, which means that the model was wrong in only 6% of observations. While in the case of predicted negative reviews, which are actually positive, the model made a mistake in only 16% of observations, so a total of 40. This error in misclassifying negative reviews can be attributed to their annotation. In fact, the accuracy of their classification by models should be additionally checked by reviewing the initial annotation.

Accordingly, the research showed that very good results are achieved using word embedding models based on BERT, and also the language in which the analysis is performed plays an important role. Furthermore, the best achieved results in the sentiment analysis of reviews in the Croatian language were obtained using the CroSloEngual BERT model. However, using models that were not pre-trained in the Croatian language but in English, the results show a small deviation of 11%, which represents their possibility of being used in sentiment analysis or some other NLP task in the Croatian language.

V. CONCLUSION

This paper describes several steps for preprocessing collected data for polarity prediction, also performing the task of sentiment analysis using different word embedding techniques and classification algorithms. The data was collected using the online travel agency Booking.com, which enables lodging reservations, using a web scraping tool used to retrieve data from websites. Data were collected on the reviews of various users of the Booking.com website, and it contains a total of 11,467 records in the Croatian language marked with polarities negative (0) or positive (1). After processing the collected data, word embedding models are defined and predictions are made using

different machine learning algorithms. The results of the research showed that a very important role in prediction is played by models that have already been trained in the language in which the analysis is performed.

Thus, this experiment showed a very high accuracy of 91.25% using the SVC algorithm, in the case when the CroSloEngual BERT model, pre-trained on the corpus of the Croatian language, was used. Also, it is important to mention that the other models pre-trained in English resulted in very good results, and the biggest difference between the best algorithms is about 11%. Therefore, other models such as BERT and DistilBERT proved to be very good, where the accuracy of the Logistic Regression classifier achieved the highest results, 84.13% and 86.25%, respectively.

Furthermore, it is important to point out that there is an insufficient number of publicly available annotated datasets in the Croatian language that are used for the SA task, and this paper highlights the contribution by collecting a new dataset that can be used for further research in the NLP field. Also, the useful tool Octoparse, which can be used for online retrieval of data from various websites, stands out.

The paper investigates important preprocessing steps that can affect prediction performance, so it is very important to perform the above operations before the SA task in order to achieve better model performance and reduce the dimensionality of the input data, which reduces the time and memory of executing the word embedding model and prediction, which was very noticeable during the experiment.

A limitation that appeared during this research is the domain of the dataset, where the data was collected from only one web source, so the data contains reviews about lodging in some cities in Croatia. In further research, it is possible to expand the dataset, for example by collecting data on service reviews from the Amazon.com commercial website. Also, due to insufficient memory and time-consuming execution of the word embedding model, a certain number of records from the dataset were used.

In future work, other languages will be included and analyzed, and the most interesting results could be achieved when comparing languages similar to Croatian, such as Serbian or Bosnian. Attempts will be made to include several more data cleaning and preprocessing steps, such as the extraction features of Part of Speech (PoS) tags, negation, opinion word or phrases that contribute to the prediction performance, and to expand the research domains and data extraction sources to further determine efficiency of the model.

ACKNOWLEDGMENT

This paper was written as a research project for the graduate course “Machine Learning” at the Faculty of Science, University of Split, Croatia.

REFERENCES

[1] Neogi, Ashwin & Garg, Kirti & Mishra, Ram Krishn & Dwivedi, Yogesh. (2021). Sentiment analysis and classification of Indian farmers' protest using twitter data. International Journal of Information Management Data Insights. 1. 100019. 10.1016/j.jjimei.2021.100019.

[2] Fernandez, Raul & Guizar, Brenda & Rho, Caterina. (2021). A sentiment-based risk indicator for the Mexican financial sector. Latin American Journal of Central Banking. 2. 100036. 10.1016/j.latcb.2021.100036.

[3] Ducange, Pietro & Fazzolari, Michela & Petrocchi, Marinella & Vecchio, Massimo. (2019). An effective Decision Support System for social media listening based on cross-source sentiment analysis models. Engineering Applications of Artificial Intelligence. 78. 71-85. 10.1016/j.engappai.2018.10.014.

[4] Sazzed, Salim & Jayarathna, Sampath. (2021). SSentiA: A Self-supervised Sentiment Analyzer for classification from unlabeled data. Machine Learning with Applications. 4. 100026. 10.1016/j.mlwa.2021.100026.

[5] Touahri, Ibtissam & Mazroui, Azzeddine. (2021). Enhancement of a multi-dialectal sentiment analysis system by the detection of the implied sarcastic features. Knowledge-Based Systems. 227. 107232. 10.1016/j.knosys.2021.107232.

[6] Asif, Muhammad & Ishtiaq, Atiab & Ahmad, Haseeb & Aljuaid, Hanan & Shah, Jalal. (2020). Sentiment Analysis of Extremism in Social Media from Textual Information. Telematics and Informatics. 48. 101345. 10.1016/j.tele.2020.101345.

[7] Kocoń, Jan & Figas, Alicja & Gruza, Marcin & Puchalska, Daria & Kajdanowicz, Tomasz & Kazienko, Przemysław. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. Information Processing & Management. 58. 102643. 10.1016/j.ipm.2021.102643.

[8] Singh, Apoorva & Saha, Sriparna & Hasanuzzaman, Md & Dey, Kuntal. (2022). Multitask Learning for Complaint Identification and Sentiment Analysis. Cognitive Computation. 14. 1-16. 10.1007/s12559-021-09844-7.

[9] Mandhula, Trupthi & Pabboju, Suresh & Gugulotu, Narsimha. (2020). Predicting the customer's opinion on amazon products using selective memory architecture-based convolutional neural network. The Journal of Supercomputing. 76. 10.1007/s11227-019-03081-4.

[10] Lai, Mirko & Cignarella, Alessandra & Hernandez Farias, Delia & Bosco, Cristina & Patti, Viviana & Rosso, Paolo. (2020). Multilingual Stance Detection in Social Media Political Debates. Computer Speech & Language. 63. 101075. 10.1016/j.csl.2020.101075.

[11] Kaity, Mohammed & Balakrishnan, Vimala. (2020). An integrated semi-automated framework for domain-based polarity words extraction from an unannotated non-English corpus. The Journal of Supercomputing. 76. 10.1007/s11227-020-03222-0.

[12] Keyvanpour, Mohammad & Zandian, Zahra & Heidarypanah, Maryam. (2020). OMLML: a helpful opinion mining method based on lexicon and machine learning in social networks. Social Network Analysis and Mining. 10. 10.1007/s13278-019-0622-6.

[13] Wang, Zhaoxia & Lin, Zhiping. (2020). Optimal Feature Selection for Learning-Based Algorithms for Sentiment Classification. Cognitive Computation. 12. 10.1007/s12559-019-09669-5.

[14] Imam Abubakar, Amina & Roko, Abubakar & Bui, Aminu & Saidu, Ibrahim. (2021). An Enhanced Feature Acquisition for Sentiment Analysis of English and Hausa Tweets. International Journal of Advanced Computer Science and Applications. 12. 10.14569/IJACSA.2021.0120913.

[15] Abdullah, Nur Atiqah Sia & Rusli, Nur Ida Aniza. (2021). Multilingual Sentiment Analysis: A Systematic Literature Review. Pertanika Journal of Science and Technology. 29. 10.47836/pjst.29.1.25.

[16] Pasupa, Kitsuchart & Ayutthaya, Thititorn. (2022). Hybrid Deep Learning Models for Thai Sentiment Analysis. Cognitive Computation. 14. 10.1007/s12559-020-09770-0.

[17] Meetei, Loitongbam & Singh, Thoudam Doren & Borgohain, Samir & Bandyopadhyay, Sivaji. (2021). Low resource language specific pre-processing and features for sentiment analysis task. Language Resources and Evaluation. 55. 10.1007/s10579-021-09541-9.

[18] Balaguer, Pau & Teixidó, Ivan & Vilaplana, Jordi & Mateo Fornes, Jordi & Rius, J. & Solsona, Francesc. (2019). CatSent: a Catalan sentiment analysis website. Multimedia Tools and Applications. 78. 1-19. 10.1007/s11042-019-07877-7.