The Polarity of Croatian Online News Related to COVID-19: A First Insight

Ilić, Anton; Beliga, Slobodan

The Polarity of Croatian Online News Related to COVID-19: A First Insight

Anton Ilić
University of Rijeka
Department of Informatics
Radmile Matejčić 2, 51 000 Rijeka, Croatia
ailic@studnent.uniri.hr

Slobodan Beliga
University of Rijeka
Department of Informatics
Center for Artificial Intelligence and Cybersecurity
Radmile Matejčić 2, 51 000 Rijeka, Croatia
sbeliga@inf.uniri.hr

Abstract. *The polarity of online news publications thematically related to COVID-19 is analysed. A collection of sentiment annotations for news articles written in the Croatian language was created and compose a new Cro-CoV-Senti-articles-2020 dataset. The news article's sentiment is derived from the reactions of portal readers. In addition, well-known sentiment analysis approaches that use lexicons and machine learning algorithms have been implemented to automatically determine the sentiment of online news. Besides, the VADER framework was used in parallel. It has been found that for the purposes of crisis communication analysis when rapid analysis solutions are needed, existing tools can be used for preliminary sentiment analysis despite some technical shortcomings. However, for a more extensive analysis of the media space and highly valuable insights, some refinements are needed. This preliminary analysis, on a sample of approximately 3,400 newspaper articles related to COVID-19, finds that readers perceive as many as two-thirds of articles negatively rather than positively.*

Keywords. sentiment analysis, COVID-19, online news articles, sentiment lexicon, machine learning, VADER

1 Introduction

Sentiment analysis or opinion mining is one of the most popular tasks in the field of natural language processing. It is a text analysis that aims to detect or predict sentiment (positive, negative, and often neutral) within the text, the entire document, or just some parts of the text such as paragraphs or sentences. By analysing sentiment based on the subjectivity present in the text, we aim to measure the attitudes, feelings, and opinions (of the author of the text) using computational methods and tools (Birjali et al., 2021; Liu, 2012).

There are two fundamental approaches to sentiment analysis (Lane et al., 2019): lexicon-based, and machine learning-based (ML-based). Moreover, there are also hybrid-based that combines the previous two (Alamoodi et al., 2021), and the most recent deep learning-based approach.

In this study, we constructed two different lexicon-based approaches for sentiment analysis, but also ML-based approaches. In parallel, we analyse the sentiment results using a popular open-source tool, called VADER, also lexicon-based and fully unsupervised – does not require annotations in advance. For the experiments, we constructed a dataset, which consists of suitable COVID-19 related news articles and their label of sentiment class (positive or negative).

The main goal of this paper is to analyse the sentiment of crisis communication in the news published on the Croatian online portal related to COVID-19 topics during the pandemic. There are three main contributions of this paper: (1) a preliminary insight into the polarity of online news publications during the coronavirus epidemic in Croatia, (2) a new dataset suitable for sentiment analysis which contains news articles related to the coronavirus together with sentiment labels derived from the news portal reader's opinion, and (3) the success of sentiment predicting on COVID-19 related articles using VADER, the existing sentiment lexicon and several different lexicon-based, as well as machine learning-based approaches, was examined. To the best of our knowledge, this is the first sentiment analysis research that explores COVID-19 related articles of Croatian online news.

2 Background and Motivation

Previous research related to COVID-19 (Melo & Figueiredo, 2021) in the Croatian media space and social networks for some aspects of online communication was conducted. Twitter social

network communication was analysed by sentiment analysis of Twitter messages (Babić et al., 2021) and COVID-19 information spreading (Babić et al., 2021). Crisis communication of Croatian online portals was explored by topic modeling of COVID-19 related articles (Bogović et al., 2021), crisis discourse analysis (Beliga et al., 2021) related to pandemic on news portal texts. With this study on sentiment analysis of newspaper articles, we would explore another still unexplored aspect of pandemic communication on Croatian news portals.

The first problem in sentiment analysis is manifested in the lack of subjectivity in the text. Therefore, the analysis of factual texts (whose content does not express attitudes and feelings) is more demanding. We often call them objective texts. On the other hand, subjective texts like product reviews or comments on blogs and social networks that express feelings, mood, emotions, and opinions are more appreciative of sentiment analysis (Sarkar, 2019).

In this paper, we use various sentiment analysis models to analyse the polarity of news published in the Croatian online media space during the pandemic caused by the SARS-CoV-2 virus. In particular, the task is very challenging because it requires analysis of news articles that belong rather to the objective, than to the subjective type of text. Nevertheless, sentiment analysis could support crisis communication analysis and investigate the polarity of news content as well as the news attitudes that the media communicated to the public during the pandemic.

Opinions are central to almost all human activities because they are key influencers of our behaviours (Liu, 2012). Therefore, the assessment of experiences and emotions that portal readers have when reading COVID-19 news articles plays an important role in the analysis of online media communication. The findings of this study, related to feelings that pervade COVID-19 news readers, may be useful for future crisis communication and possible adjustments of the directives of publishing news on portals, but also the awareness of the population about the patterns of media houses and their publication habits in the pandemic period.

Apart from the problem related to the lack of subjectivity in the text, another major problem is the lack of linguistic resources for the analysis of sentiment on texts written in the Croatian language. Sentiment analysis challenges in the multilingual scenario related to both low-resource and high-resource languages are discussed in (Nankani et al., 2020). A robust, reliable, and publicly available tool for analysing the sentiment of texts written in the Croatian language has not yet been developed. Especially not for the analysis of newspaper articles related to the pandemic or coronavirus theme. Therefore, our challenge lies in making the best and rapid solution for crisis communication analysis using available resources for the Croatian language.

There are several scientific studies related to the sentiment analysis of texts written in Croatian. The research by Glavaš et al. (2012) deals with the semi-supervised acquisition of the sentiment lexicon. They developed a Croatian sentiment lexicon called CroSentiLex. It consists of positive and negative lists of words, each containing 37K Croatian lemmas ranked by positivity and negativity, respectively, with the corresponding PageRank scores. The rankings were created automatically based on small positive and negative seed sets and co-occurrence frequencies, using the PageRank algorithm. The approach that Glavaš et al. (2012) used for sentiment analysis is corpus-based. They construct sentiment-annotated lexicon using Croatian newswire corpus *Vijesnik* which contains approximately 230K documents. Considering that we also want to analyse newspaper articles, there is an indication that such a lexicon of sentiment could be useful. The lexicon is a publicly available resource, and we used it in this research.

Sentiment analysis on COVID-19 related content has so far been most often researched on texts written in English, such as the research presented in (Shofiya & Abidi, 2021) where the SentiStrength tool was used to extricate sentiment polarity of tweet texts, and support vector machine (SVM) algorithm was employed for sentiment classification. There is also somewhat more challenging research on texts written in other languages such as Brazilian Portuguese (Melo & Figueiredo, 2021), which due to the lack of language resources translated all original texts from Portuguese to English and then used available resources for English, or a deep learning sentiment analyser for low-resource languages such as Albanian (Kastrati et al., 2021). For Spanish, a multi-layered perceptron (MLP) was used (Heras-Pedrosa et al., 2020), while for Greek simple lexicons were used (Kydros et al., 2021). Due to the lack of tools for low-resourced languages, in the multilingual scenario (Kruspe et al., 2020) used the pre-trained word- and sentence-level embeddings with a skip-gram version of word2vec, and a multilingual version of Bidirectional Encoder Representations from Transformers (BERT) trained on Wikipedia data or tweets containing COVID-19 keywords, the Multilingual Universal Sentence Encoder (MUSE), etc.

3 Methods

The main idea of our approach is to combine existing (off-the-shelf) NLP resources and adapt them to perform the rapid study. Due to the specifics of crisis communication, it is very important to get information about the polarity of communication messages that the media communicate in the online media space to the public, as soon as possible. Constructing new language resources and sophisticated methods can be demanding and time-

consuming. In order to be as efficient as possible, we decided to use the existing language reserves that are available for sentiment analysis for Croatian texts.

The following subsections describe the process of data collection – i.e., newspaper articles that communicate crisis topics, the process of annotating the sentiment of such articles and the processing tools and methods we used to analyse natural language and automatically determine the sentiment of newspaper articles.

3.1 Data

Newspaper articles published on the Croatian online portal *dalmacijadanas.hr* were analysed. *Dalmacijadanas* is an online portal that covers the media space of the largest Croatian county, in which Split, other coastal cities, Zagora and the islands are primarily represented. In the publications, journalists emphasize positive stories from the lives of "ordinary" people, follow the Croatian local economy, entrepreneurs, students, associations, etc. They combine the characteristics of a news portal and a portal that follows the stories, problems, and successes of Dalmatians. It covers a wide range of topics and is suitable for crisis communication analysis in the area of southern Croatia.

Data acquisition lasted from January 1, 2020 to April 30, 2021. We tracked all articles published during that period. Basic statistics of collected articles in the observed period were presented in Table 1.

Table 1. Statistics of news articles

	No.
All published articles	34,332
COVID-19 related articles	12,717
Useful articles	3,392

In the period of 16 months covered by this study, the number of published articles on the portal reaches more than 34 thousand. All articles that contain in their heading, subheadings, or body text any of the words from the COVID-19 thesaurus[1] (such as coronavirus, pandemic, etc.) are classified as COVID-19 related. The words in the COVID-19 thesaurus are written in their canonical form, but all their morphological variations are also used in the filtering.

It is important to note that 37% of published articles are related to COVID-19. Previous studies have shown that this number for the entire 2020 year on Croatian mainstream news portals reaches almost 45% (Beliga et al., 2021). This may be an indication that in this regional area (covered by *dalmacijadanas.hr*) is less occupancy of the media space with COVID-19 topics than in Croatian mainstream portals in general. Notwithstanding this incidental finding, from 37% of COVID-19 related articles, only 26.7% of them left an impression on readers that is strong enough to leave their reaction to the read text in the form of emotion (emoticon). Based on such insight, we can say that the overall readers' interest in expressing emotions to texts related to COVID-19 is relatively small.

Furthermore, in experiments, we use only articles related to COVID-19 for which a minimum of three readers left their reaction on the portal in the form of an emoticon. In this way, we wanted to ensure high annotation quality given the possibility of divergent opinions of readers. All COVID-19 articles that had only one or two reader reactions were not included in the experiments. Constructed *Cro-CoV-Senti-articles-2020*[2] dataset contains 3,392 news articles with a corresponding sentiment label (positive or negative) which was derived from at least 3 reactions of readers. In the next subsection sentiment annotations on COVID-19 related articles are explained in detail.

3.2 Sentiment Annotations

The specificity of portal *dalmacijadanas.hr* is that it leaves readers space to express their emotional reaction to the read news. They can do it by selecting just one of the seven offered emoticons. Emoticons are positioned at the bottom of the news article, and readers are not required to react or use them if they do not want to. In this way, readers can express their reaction, i.e. their sentiment to the read text. Emotions are presented to readers in the form of emoticons shown in Figure 1.

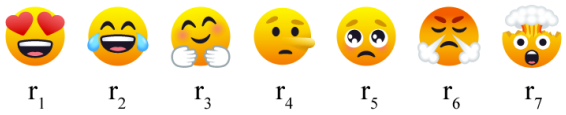

Figure 1. Reactions presented in the form of emoticons (simply marked as: r_1-love, r_2-laugh, r_3-hug, r_4-lie, r_5-sad, r_6-mad, r_7-mind blown)

We did not independently select the defined emoticons during the design of this study, they were determined by the editorial board of the portal. However, we decided to use them because we think they represent a good sample by which basic or primary emotions (such as fear, sadness, love, anger, joy, etc.) can be designated or described. Table 2 contains the text alias, the corresponding reaction expressed by each emoticon, the polarity (P), and weight factor (w_i) for each emoticon marked on Figure 1 from r_1 to r_7. According to the

[1] COVID-19 thesaurus used for articles filtering: https://github.com/sbeliga/InfoCoV/tree/main/CECIIS2021

[2] The dataset contains links to newspaper articles with corresponding sentiment labels. It is publicly available from: https://github.com/sbeliga/InfoCoV/tree/main/CECIIS2021. In addition, data generated and/or analysed during the current study are available from the corresponding author on reasonable request.

recommendations of the cognitive linguist, positive (+) and negative (-) polarity labels are assigned to the emoticons, and a bipolar label (+/-) to those that are ambiguous (they can belong to a positive but also to a negative polarity class depending on the situation, i.e. context). In addition, we defined positivity/negativity in terms of weight factor for all emoticons on a scale from strictly positive value 1 to strictly negative value -1. The mean of the scale (value 0) represented a neutral reaction.

Table 2. Description of emoticons[3], associated descriptions, polarities and assigned weights

r_i	Text alias	Represented reaction	P	w_i
r_1	*wow; heart-face; heart-eyes*	love; infatuation; adoration	+	1
r_2	*face with-tears-of-joy; laughing-crying; LOL; laughing-tears*	joy; joking; fun; teasing; irony; sarcasm	+/-	-0.25
r_3	*hug; hugging; hugging-face*	love; affection; gratitude; support; consolation; care; enthusiasm; accomplishment	+	1
r_4	*hmm; liar; lying-face; long-nose*	lying; deceit; dishonesty, disbelief or feeling abashed	-	-0.25
r_5	*sad; cry; crying -face; worried-face*	sadness; pain; crying; tense emotions, concern, anxiety, disappointment, unhappiness	-	-1
r_6	*frustrated; mad-face; steaming; face-with-steam-from-nose*	irritation, anger, contempt, dominance, empowerment; frustration; pumped-up passion	-	-1
r_7	*mind-blown; exploding-head*	shock, awe, amazement, disbelief; upset; unacceptable or unpleasant emotion	+/-	-0.5

The two emoticons belong to a strictly positive class (marked with r_1 and r_3), and they express strictly positive reactions to the read textual content such as love or support. We assigned their weight factor to a strictly positive value of 1. The other two emoticons represent strictly negative reactions (marked with r_5 and r_6) such as sadness and anger. We assigned them strictly negative weights with a value of -1. The next 3 emoticons have a bipolar label (+/-). Based on their expertise and insight into the sentiment of newspaper articles that readers most often associate with such bipolar emoticons, linguists decided that such emoticons belong more to the negative than to the positive class. Two bipolar emoticons (marked with r_2 and r_4) were rated as slightly negative (joy and lying). Reaction *joy* is considered rather be used in a negative connotation such as sarcasm, than in the context of positive reaction such as *laughter*, *fun*, or *joke*. Reaction *lying* is used in confusing situations that are not necessarily confirmed as negative. Therefore, the weight factor for both emoticons is slightly negative and set up to -0.25. The last reaction, called *mind-blown* (marked with r_7) can indicate amazement which can be both positive and negative. However, it is more commonly used to express unacceptable or unpleasant reactions or emotions such as upset, so it is weighted with a negative value of -0.5.

Final article polarity – i.e. sentiment expressed with a numerical value of each newspaper article was defined by the sum of the reactions assigned by the readers for all 7 defined reactions that were additionally weighted with the previously described weight values. Such article polarity $P_{article}$ is calculated by the expression (1) according to the number of times readers have chosen reaction r_i multiplied with the corresponding weight w_i as defined in Table 2:

$$P_{article} = \sum_{i=1}^{7} Count(r_i)\, w_i \qquad (1)$$

In addition, based on reader reactions and values calculated by equation (1), each article is assigned to a negative or non-negative (positive) class. We defined mapping as an ordered triple (*P*, *S*, *f*) that in *P* contains a set of polarity values determined by readers for all articles in the collection, in *S* contains a set of two possible classes (negative and positive), and the rule $f{:}P \rightarrow S$ according to which each article $x \in P$ is associated with a unique class $y \in S$ in such a way that $y = f(x)$. Such a binary classifier or function $f{:}P \rightarrow S$ is defined by the expression:

$$f(x) = \begin{cases} non-negative, x \geq 0 \\ negative, x < 0 \end{cases} \qquad (2)$$

In this study, the class positive is also called non-negative because the polarity 0 is associated with the positive class.

Representation of emoticons in COVID-19 related articles that meet the criterion of minimally 3 readers reactions are presented in Figure 2. Each bar on the chart represents the total count of particular emoji reaction sorted in descending order. The columns on the graph that are coloured with red represent positive, with blue negative, and with the green bipolar, i.e. slightly negative reactions. The most common reaction on COVID-19 related articles is strictly negative and is manifested in the emoticon r_5–*sad*. The most commonly used strictly positive reactions, those manifested by emoticons of love and support (r_1-*love* and r_3-*hug*) are ranked slightly lower than strictly negative reactions (r5-*sad* and r_6-*mad*). Bipolar, i.e. slightly negative reactions (green bars)

[3] According to: https://emojipedia.org/

are at the bottom of the chart scale. Furthermore, it is evident that the total sum of strictly negative reactions is significantly higher than the sum of strictly positive reactions (blue columns are ranked higher than red ones). With a higher amount of negative reactions to the content of news articles, readers express their negative opinion and consequently suggest a larger number of negatively polarized articles than positive ones.

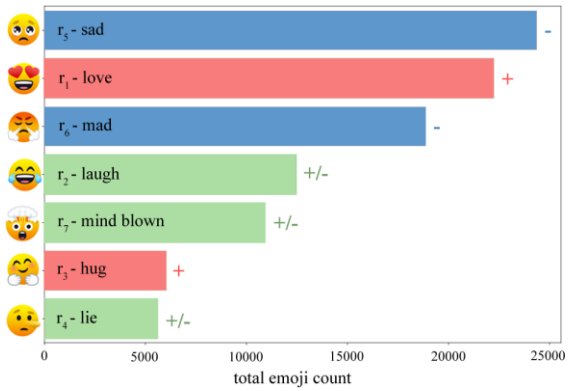

Figure 2. Representation of emoticons in articles that meet the criterion of a minimum number of reader reactions (emoji marked as: r_1-love, r_2-laugh, r_3-hug, r_4-lie, r_5-sad, r_6-mad, r_7-mind blown)

This was confirmed by the ratio of articles representing the non-negative and negative class. In the *Cro-CoV-Senti-articles-2020* dataset, according to the classification defined by formula (2), 2271 articles belong to the negative class, while the remaining 1121 to non-negative. Hence, we find that the ratio of negative to non-negative class is 2:1. Approximately, third (33.05%) of articles fold to the non-negative class, while the remaining two-thirds (66.95%) belong to samples of articles that are classified as negative. Nevertheless, the general sentiment of people publicly expressing reactions to such publications is three times more often in negative than in non-negative polarity.

Table 3. Representation of emoticons (readers reactions) depending on the sentiment class

Reaction	Class	
	-	+
r_1-love (+)	4568	17637
r_2-laugh (+/-)	10167	2267
r_3-hug (+)	2342	3681
r_4-lie (+/-)	4205	1376
r_5-sad (-)	23283	1063
r_6-mad (-)	17077	1791
r_7-mind blown (+/-)	9123	1767
TOTAL	61642	27815

Representation of emoticons depending on the class is shown in Table 3. As expected, predominantly positive emoticons (reactions) are more represented in the non-negative and predominantly negative emoticons in the negative class. Higher values are highlighted in grey.

Figure 3 shows the representation of articles in 6 selected newspaper categories, grouped by class (negative and non-negative). It is important to note that the category of News (i.e. general daily news) is also the most numerous category in terms of the total number of articles, unlike all other categories. The Relax category, which writes about casual topics such as music, gastronomy, fashion, lifestyle, and health, has an almost equal number of articles in both classes. The dominance of the non-negative (positive) class is evident in the category of columns, i.e. the permanent newspaper columns of individual journalists, but also in articles related to sports.

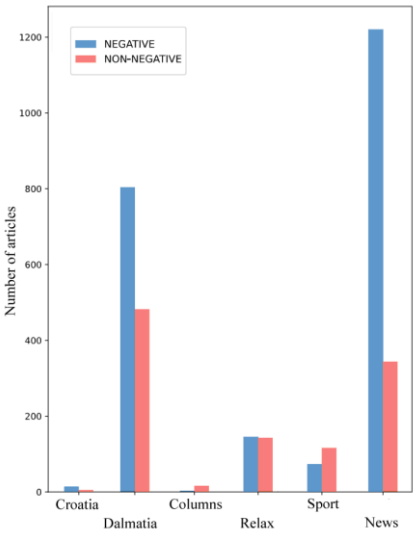

Figure 3. Positivity of articles by news category

3.3 Automatic Sentiment Analysis

In the experiments of this study, we perform unsupervised sentiment analysis based on lexicons. In parallel, we used some available and well-known rule-based open-source solution. Additionally, to predict the sentiment of the text we also use some traditional supervised machine learning models.

Two unsupervised lexicon-based approaches were implemented as follows. The first implemented approach is called **Average Sentiment of Sentences** (**ASS**). This approach follows the principle of averaging sentiment at the sentence level. The sentiment of articles was determined by the following procedure:

- The text is segmented at the sentence level.
- Sentences are lemmatized and stopwords are removed.

- The occurrences from the sentence were assigned with numerical polarity values based on the sentiment lexicon.
- The polarity of the individual sentence was calculated by summing the values of all the assigned polarities to the individual words in a sentence. Then, the polarity of the individual sentence is expressed as a value normalized on the sentence level.
- Finally, the average polarity value of all sentences is calculated.
- According to the average value of the sentence polarity for the article and equation (2), class affiliation is determined.

The second implemented approach is called **Sentiment of Full Text** (**SFT**). This approach is based on the principle of simple sentiment accumulation at the occurrence level. The sentiment of articles is in this case determined by the following procedure:

- The text is lemmatized and stopwords are removed.
- The words were assigned with numerical polarity values based on the sentiment lexicon.
- The polarity of text was calculated by summing the values of all the assigned polarities to the individual words. Then, the polarity of the text is expressed as a value normalized on the document level.
- According to full-text polarity and equation (2), class affiliation is determined.

It is important to note that in both approaches, words in a sentence or text are neglected in the calculation if the corresponding polarity value for them is not defined in the sentiment lexicon. In addition, we paid attention to words used for negation. With the words behind the negations, corresponding polarity values are shifted to opposite polarity (numeric value remains the same while the sign in front of the value is changed). Article headings are treated as sentences added to the beginning of the article text. Croatian sentiment lexicon CroSentiLex[4] constructed by Glavaš et al. are used in both, ASS and SFT approaches.

The next implemented approach is machine learning-based. It combines well-known TF-IDF (Term Frequency - Inverse Document Frequency) model for document representation and three different machine learning classifiers Naïve Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF). The sentiment of articles is determined by the following procedure:

- The text is lemmatized and stopwords are removed.
- Using the TF-IDF model, the text is converted to features.
- A binary classifier model is induced and trained to predict a negative and non-negative class using the NB, SVM, or RF algorithm.
- A classifier was trained and tested.

In the experiment, 5-fold cross-validation is used to estimate the sentiment classification skill of the model on new data. The problem of unbalanced classes was observed. The negative class has more than 2.000 samples, while the non-negative one has slightly more than 1.000. The difference in the number of samples between classes is large.

In order to achieve a balance between classes that is more suitable for machine learning, resampling which combines up-sampled minority class (non-negative) with down-sampled majority class (negative), was done. Machine learning was performed on a balanced set of samples (2.000 samples per class).

The last approach that was used is one of the popular lexicon-based sentiment model called **VADER** (Valence Aware Dictionary and sEntiment Reasoner). It is developed by Hutto and Gilbert (2014) and is based on a rule-based sentiment analysis framework. Initially, it was tuned to analyse sentiments in social media, requires no training data, but is constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon (Hutto & Gilbert, 2014). It is fully unsupervised and can be applied directly to unlabelled data.

VADER takes care of both polarity and intensity of emotions. Polarity determines whether the sentiment is positive or negative while the intensity indicates how positive or negative the sentiment is. In terms of *F*1 score, VADER achieves a score of 0.96, outperforming individual human raters who achieve a score of 0.84 when classifying the sentiment of tweets into positive, neutral, or negative classes. In determining the sentiment of a text, valence scores are assigned to the words. The score is measured on a scale from -4 (for the most negative sentiment) to +4 (for the most positive sentiment). The computation of the valence score of an input sentence is realised by a heuristic. It pays particular attention to punctuation (.!?), capitalization (e.g., ALL-CAPS), degree modifiers (e.g. “extremely sad” or “slightly mad”), polarity shift (conjunctions, e.g. "but"), and polarity negation. (e.g. "isn't really nice"). The final compound score is expressed as a normalized value on a scale of - 1 to 1. It is calculated as the sum of the valence scores given in the lexicon for all words, adjusted according to the rules (heuristics) mentioned before. The sentiment lexicon used in this model contains sentiment scores associated with words, emoticons, and slang. It contains over 7,500 lexical features with validated valence scores. Since it was originally developed for English, in this study we adapted it for use in Croatian texts. First, we translated the texts from Croatian to English using the Google Translate

[4] https://takelab.fer.hr/sentilex/

API through the *googletrans*[5] Python library. Then, we used the English translation as input to VADER. In the experiments, we used the open-sourced version of VADER[6] that comes under the MIT License.

All previously mentioned approaches were used in two different experimental settings. The first, which analyses only the sentiment of the headlines of newspaper articles, and the second, which analyses the entire texts of newspaper articles.

3.4 Evaluation Methodology

The dataset used in the experiment contains news article text with the corresponding class label (negative or non-negative sentiment) – derived from reactions provided by at least three authors who read particular news on the portal and chose one of the seven available emotions presented as emoticons. Such sentiment annotations are taken as ground truth in the evaluation procedure. Since the standard binary classification task is evaluated, standard information retrieval and machine learning measures were used for evaluation. For all supervised and unsupervised sentiment analysis approaches, accuracy (ACC), precision (P), recall (R), and the *F*1 score are used for the evaluation. In a binary classification task, the classifier's prediction is defined in terms of "positive" or "negative". Moreover, the terms "true" and "false" refer to whether this prediction is consistent with the external judgment. Considering these definitions, Table 4 can be formulated.

Table 4. Actual vs. predicted conditions.

	Actual class (observation)	
Predicted class (expectation)	**TP** (true positive) correct result	**FP** (false positive) unexpected result
	FN (false negative) missing result	**TN** (true negative) correct absence of result

Accuracy, precision, recall, and *F*1-measure are defined with equations:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$P = \frac{TP}{TP + FP} \quad (4)$$

$$R = \frac{TP}{TP + FN} \quad (5)$$

$$F1 = \frac{2PR}{P + R} \quad (6)$$

In experiments with unsupervised models, all available articles, i.e. the complete dataset (3,392), were used in the evaluation process. In machine learning-based models, after class balancing (resampling), a 5-fold cross-validation procedure was performed.

[5] https://github.com/ssut/py-googletrans
[6] https://github.com/cjhutto/vaderSentiment

4 Results

In this section, we present the results obtained by sentiment analysis in terms of accuracy (*ACC*), precision (*P*), recall (*R*), and *F*1 score. Results are presented separately for those achieved by standard lexicon-based models, and those achieved by hybrid-based models that combine TF-IDF features and machine learning algorithms. Models are abbreviated as follows: ASS – Average sentiment of sentences, SFT – Sentiment of full text, VADER - Valence Aware Dictionary and sEntiment Reasoner, NB - Naïve Bayes, SVM - Support Vector Machine, RF – Random Forest.

Firstly, in Table 5, results are presented for sentiment obtained exclusively from the headlines of newspaper articles. Lexicon-based models perform worse than machine learning ones. Simple ASS and SFT models achieve slightly lower *F*1 scores than the more sophisticated VADER model. Although VADER is only 2% better than SFT, all models of this lexicon-based group do not achieve enviable results. Their results are only slightly better than those that would be achieved by a random classifier. The group of ML-based models achieves significantly higher values for all observed measures from accuracy, precision, recall, to *F*1 scores. In terms of *F*1 score, the Random Forest model achieves the highest value (0.83).

Table 5. Comparison of models for sentiment analysis of newspaper article headlines

	ACC	***P***	***R***	***F1***
	Lexicon-based models			
ASS	0.48	0.48	0.48	0.47
SFT	0.53	0.53	0.53	0.53
VADER	0.62	0.62	0.63	**0.55**
	Machine learning-based models			
NB	0.82	0.82	0.82	0.76
SVM	0.82	0.82	0.82	0.82
RF	0.84	0.84	0.83	**0.83**

Secondly, in Table 6, results are presented for sentiment obtained from news articles, including headlines, subheadings, and the full text of the news article. In this case, both groups of models (lexicon- and ML-based) for all observed measures, achieve very similar results as those shown in Table 1 for models that determine sentiment only on the article's headings. However, in terms of *F*1 score, from the group of lexicon-based models, VADER is again the best and achieves 0.57, while in the group of hybrid-based models Random Forest stood out with 0.89.

Based on the presented results, we can conclude that ML-based models are more successful in the cases of sentiment analysis we have considered in this study. The success of the automatic sentiment analysis in terms of the *F*1 score based just on the headline of a news article reaches values that are slightly higher than 80%, while for the full text of a news article this value reaches almost 90%.

The results achieved by the VADER model are certainly affected by the noise entered into the data during the automatic machine translation of Croatian texts into their English language pair. The lower success rate of the model is the penalty we get due to the use of a model that was initially designed for texts written in English and then refitted for the Croatian language. Apart from machine translation, the valence scores of words in English are certainly not completely identical to the scores in Croatian. Therefore, the polarities of words in the sentiment lexicon used by VADER are also possible causes of poorer results.

Table 6. Comparison of models for sentiment analysis of newspaper articles

	ACC	***P***	***R***	***F*1**
	Lexicon-based models			
ASS	0.42	0.47	0.47	0.43
SFT	0.52	0.54	0.53	0.53
VADER	0.57	0.63	0.63	**0.57**
	Machine learning-based models			
NB	0.77	0.77	0.77	0.77
SVM	0.84	0.84	0.84	0.84
RF	0.89	0.89	0.89	**0.89**

Table 7. Sentiment results presented by class for Random Forest model

	P	***R***	***F*1**
	article headlines		
negative	0.89	0.77	0.82
non-negative	0.79	0.91	0.85
	article full text		
negative	0.91	0.87	0.89
non-negative	0.87	0.92	0.89

Table 7 presents classification performance for the Random Forest algorithm, which proved to be the best. The results are shown separately by class. In both cases, when we analyse only the headlines and when we analyse the whole text, the values of *R* are slightly higher than the values of *P* in the non-negative class. In contrast to the negative class where *P* dominates in front of *R*.

Finally, it is important to note that techniques used in this study (primarily sentiment lexicon), due to the purpose of rapid analysis of crisis communication, were taken as resources from previous research and were not specifically designed for COVID-19 (domain or corpora) adapted analysis. Therefore, we can be satisfied with the success of the quickly adapted models and consider them sufficiently reliable for the preliminary analysis that gives the first insight into the sentiment of the online media news space related to the COVID-19 topic. Especially those ML-based which shows promising results.

5 Discussion and Conclusion

This paper provides the first insight into the polarity of Croatian online news related to COVID-19. Natural language processing techniques were used to analyse the sentiment of news articles.

The paper examines several different methods that could be easily performed for Croatian based on the available tools and resources. The open-source VADER method was used. Several simple lexicon-based methods have been constructed that use the available sentiment lexicon for Croatian. Some simple machine-learning-based methods with NB, SVM, and RF algorithms which are commonly used in sentiment analysis were also examined. Finally, a dataset suitable for coronavirus-related sentiment analysis was constructed. This preliminary results on a sample of approximately 3,400 newspaper articles related to COVID-19, finds that readers perceive as many as two-thirds of articles negatively rather than positively.

A lack of language resources, a sentiment lexicon appropriate for a particular corpus or domain, and especially, robust sentiment analysis system, do not provide much opportunity for sentiment analysis on a specific task. Another aggravating circumstance is the time needed to develop the tools and for the analysis itself, which is important if not crucial in crisis situations. Conducted experiments have shown that we can easily construct lexicon-based approaches if we have at least some sentiment lexicon. Our results on the article headlines or on the complete news articles conducted by methods that accumulate sentiment at the level of a single sentence, or the entire text, prove to be insufficient. They achieve results only slightly better than random classifiers.

Another solution is to use ready-made tools such as VADER. This solution may be deficient due to application in a language different from English. The reason for this is machine translation, which introduces noise into the data. Consequently, the quality of sentiment analysis is impaired.

The third solution is to use standard machine learning-based approaches. They have proven to be the most cost-effective in our experiments. The TF-IDF model in combination with the RF algorithm proved to be successful enough. Based on such an approach, sufficiently reliable preliminary results can be obtained. The advantage of this approach is that in

crisis situations, results can be obtained quickly and even without a sentiment lexicon. However, a sufficient amount of human sentiment annotations is required.

However, for more serious and extensive research in future work, it is advisable to consider using more advanced methods. For example, using the corpus-based dependency graph to create a more reliable vocabulary of sentiment, as suggested in recent work by Ban Kirigin et al. (2021). Another way to circumvent the lack of resources for Croatian could be to develop a deep-learning-based model. For example, cross-lingual sentiment analysis using a multi-task learning approach, as proposed by Thakkar et al. (2021), could incorporate better quality language resources from another language in the same (sub)family of Slavic languages into more reliable sentiment analysis solutions for Croatian.

Acknowledgments

We kindly thank the portal *dalmacijadanas.hr* for agreeing to participate in this research. This work has been supported in part by the Croatian Science Foundation under the project IP-CORONA-04-2061, "Multilayer Framework for the Information Spreading Characterization in Social Media during the COVID-19 Crisis" (InfoCoV).

References

Alamoodi, A. H., Zaidan, B. B., Zaidan, A. A., Albahri, O. S., Mohammed, K. I., Malik, R. Q., Almahdi, E. M., Chyad, M. A., Tareq, Z., Albahri, A. S., Hameed, H., & Alaa, M. (2021). Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert Systems with Applications*, *167*. https://doi.org/10.1016/ j.eswa.2020.114155

Babić, K., Petrović, M., Beliga, S., Martinčić-Ipšić, S., Jarynowski, A., & Meštrović, A. (2021). In: *Yang XS., Sherratt S., Dey N., Joshi A. (eds) Proceedings of Sixth International Congress on Information and Communication Technology*. Lecture Notes in Networks and Systems, vol 216. Springer, Singapore. https://doi.org/10.1007/978-981-16-1781-2_35

Babić, K., Petrović, M., Beliga, S., Martinčić-Ipšić, S., Pranjić, M., & Meštrović, A. (2021). Prediction of COVID-19 related information spreading on Twitter. *Proceedings of the IEEE International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2021)*, accepted for publication.

Ban Kirigin, T., Bujačić Babić, S., & Perak, B. (2021). Lexical Sense Labeling and Sentiment Potential Analysis Using Corpus-Based Dependency Graph. *Mathematics*, *9*(12), 1–22. https://doi.org/10.3390/math9121449

Beliga, S., Martinčić-Ipšić, S., Matešić, M., & Meštrović, A. (2021). Natural Language Processing and Statistic: The First Six Months of the COVID-19 Infodemic in Croatia. In *The Covid-19 Pandemic as a Challenge for Media and Communication Studies*. Routledge, Taylor & Francis Group, accepted for publication.

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*. https://doi.org/10.1016/j.knosys.2021.107134

Bogović, P. K., Meštrović, A., Beliga, S., & Martinčić-Ipšić, S. (2021). Topic Modelling of Croatian News During COVID-19 Pandemic. *Proceedings of the IEEE International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2021)*, accepted for publication.

Glavaš, G., Šnajder, J., & Dalbelo Bašić, B. (2012). Semi-Supervised Acquisition of Croatian Sentiment Lexicon. *Proceedings of 15th International Conference on Text, Speech and Dialogue, TSD 2012, Brno*, 166–173.

Heras-Pedrosa, C. de las, Sánchez-Núñez, P., & Peláez, J. I. (2020). Sentiment Analysis and Emotion Understanding during the COVID-19 Pandemic in Spain and Its Impact on Digital Ecosystems. *International Journal of Environmental Research and Public Health*, *17*(15), 1–22. https://doi.org/10.3390/IJERPH17155542

Hutto, C. J., & Gilbert, E. E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. https://github.com/cjhutto/vaderSentiment

Kastrati, Z., Ahmedi, L., Kurti, A., Kadriu, F., Murtezaj, D., & Gashi, F. (2021). A Deep Learning Sentiment Analyser for Social Media Comments in Low-Resource Languages. *Electronics*, *10*(10), 1133. https://doi.org/10.3390/ELECTRONICS10101133

Kruspe, A., Häberle, M., Kuhn, I., & Zhu, X. X. (2020). *Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic*. https://arxiv.org/abs/2008.12172v1

Kydros, D., Argyropoulou, M., & Vrana, V. (2021). A Content and Sentiment Analysis of Greek Tweets during the Pandemic. *Sustainability*, *13*(11), 6150. https://doi.org/10.3390/SU13116150

Lane, H., Howard, C., & Hapke, H. M. (2019). *Natural Language Processing in Action*. Manning. https://www.oreilly.com/library/view/natural-language-processing/9781617294631/

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool. https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.html

Melo, T. de, & Figueiredo, C. M. S. (2021). Comparing News Articles and Tweets About COVID-19 in Brazil: Sentiment Analysis and Topic Modeling Approach. *JMIR Public Health and Surveillance*, *7*(2). https://doi.org/10.2196/24585

Nankani, H., Dutta, H., Shrivastava, H., Rama Krishna, P. V. N. S., Mahata, D., & Shah, R. R. (2020). Multilingual Sentiment Analysis. In *Deep Learning-Based Approaches for Sentiment Analysis* (pp. 193–236). Springer, Singapore. https://doi.org/10.1007/978-981-15-1216-2_8

Sarkar, D. (2019). Text Analytics with Python: A Practitioner's Guide to Natural Language Processing. In *Text Analytics with Python* (2nd ed.). Apress. https://doi.org/10.1007/978-1-4842-4354-1

Shofiya, C., & Abidi, S. (2021). Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data. *International Journal of Environmental Research and Public Health 2021, Vol. 18, Page 5993*, *18*(11), 1–10. https://doi.org/10.3390/IJERPH18115993

Thakkar, G., Mikelić Preradović, N., & Tadić, Ma. (2021). Multi-task Learning for Cross-Lingual Sentiment Analysis. *2nd International Workshop on Cross-Lingual Event-Centric Open Analytics*, 1–9. http://ceur-ws.org/Vol-2829/short1.pdf