

1

2

3

4 Identification, analysis and prediction of valid and false information related to
5 vaccines from Romanian tweets

6

7

8 Andrei Valeanu, Dragos Paul Mihai*, Corina Andrei, Ciprian Puscasu, Alexandra Mihaela

9 Ionica, Miruna Ioana Hinoveanu, Valentina Patricia Predoi, Ema Bulancea, Cornel Chirita,

10 Simona Negres, Cristian Daniel Marineci

11

12

13

14 Carol Davila University of Medicine and Pharmacy, Faculty of Pharmacy, 6 Traian Vuia St.,

15 020956 Bucharest, Romania

16

17

18 *Corresponding author

19 E-mail: dragos_mihai@umfcd.ro (DPM)

20 **Abstract**

21 The online misinformation might undermine the vaccination efforts. Therefore, given the fact that
22 no study specifically analyzed online vaccine related content written in Romanian, the main
23 objective of the study was to detect and evaluate tweets related to vaccines and written in
24 Romanian language. 1400 Romanian vaccine related tweets were manually classified in true,
25 neutral and fake information and analyzed based on wordcloud representations, a correlation
26 analysis between the three classes and specific tweet characteristics and the validation of several
27 predictive machine learning algorithms. The tweets annotated as misinformation showed specific
28 word patterns and were liked and reshared more often as compared to the true and neutral ones.
29 The validation of the machine learning algorithms yielded enhanced results in terms of Area Under
30 the Receiver Operating Characteristic Curve Score (0.744-0.843) when evaluating the Support
31 Vector Classifier. The predictive model estimates in a well calibrated manner the probability that
32 a specific Twitter post is true, neutral or fake. The current study offers important insights regarding
33 vaccine related online content written in an Eastern European language. Future studies must aim
34 at building an online platform for rapid identification of vaccine misinformation and raising
35 awareness for the general population.

36 **Keywords:** vaccines, public health, misinformation, wordcloud, machine learning, support vectors

37

38

39

40

41 **1. Introduction**

42 Vaccines are among the most important medications worldwide. It is estimated that they have
43 saved millions of lives and that they will continue to do so [1]. Vaccines had a crucial role in the
44 eradication of smallpox in 1980 and in bringing poliomyelitis very close to eradication [2,3]. In
45 addition, a report found that as of November 2021, the Covid-19 vaccines saved nearly half a
46 million lives in less than a year in the over 60 years old group across the WHO European Region
47 [4].

48 However, despite their essential therapeutic effect and good safety profile, various disinformation
49 articles, news and social media posts have emerged in the last decades, leading to the anti-vaccine
50 movement. Even though the facts behind such information were proven to be false, the vaccine
51 fake news phenomenon has led in many countries to a reduction of the vaccination rates, both in
52 the adult and the pediatric population [5]. Low vaccination rates pose the risk of diseases that
53 currently have a low impact in the population to return with a higher impact, with an additional
54 burden on the healthcare system [1].

55 Numerous fake news related to the Covid-19 vaccines have also emerged and spread during the
56 pandemic [5,6]. The online disinformation, in combination with other social and economic factors
57 (such as media usage, educational background, health literacy, public trust in the government and
58 health system) have been hypothesized to influence a person's decision of getting the Covid-19
59 vaccine [7–9].

60 Social media platforms (such as Facebook and Twitter) are among the most important tools for
61 spreading information about vaccines, whether it is valid information or fake news [5]. Therefore,

62 the analysis of the content distributed through such platforms might be of an utmost importance in
63 order to inform the general population and the health policy makers.

64 With regards to Twitter content, several studies have evaluated vaccine related posts (whether
65 Covid or non-Covid), with regards to identifying and predicting disinformation, analyzing vaccine
66 hesitancy, performing sentiment classification or other relevant analyses [10–19]. The majority of
67 the studies were based on tweets written in English. Other analyzed languages were Dutch,
68 Moroccan and Turkish, while one study involved a multi-language approach for detection and
69 classification of tweets related to Covid-19 [12,14,15,18]. However, to our knowledge, no such
70 study specifically analyzed vaccine related content based on Romanian tweets.

71 Therefore, the objective of the present study was to analyze vaccine related content, with the main
72 goal of developing specific machine learning models for predicting disinformation from tweets
73 written in Romanian.

74 **2. Materials and Methods**

75 **2.1. Data collection**

76 The vaccine related tweets were automatically extracted by using snsrape package developed in
77 Python programming language [20,21]. The Twitter API was queried by using all the Romanian
78 forms of the noun “vaccine”, the verb “to vaccinate”, as well as other ironical related terms (such
79 as “vax”, “vaxin” or “vaxxin”) [20]. All tweets (both original posts and replies, both Covid and
80 non-Covid vaccine information) from 4 relevant 4-week periods during the Covid-19 pandemic
81 were initially collected (First period: March 16, 2020 – April 12, 2020; second period: December
82 27, 2020 – January 23, 2021; third period: May 3, 2021 – May 30, 2021; fourth period: October
83 18, 2021 – November 14, 2021). Each period was considered suggestive for the aim of extracting

84 relevant batches of tweets. March 16, 2020 was the date in which the Emergency State was
85 declared in Romania due to the Covid-19 pandemic; December 27, 2020 was the first day of the
86 Covid-19 vaccination campaign in Romania; May 2021 was the month in which the highest
87 average number of Covid-19 vaccine doses were administered and October-November 2021 was
88 the period with the highest number of deaths due to Covid-19 in Romania [22].

89 After the initial collection, for each of the 4 periods, the tweets from the 7-day period with the
90 highest number of tweets were considered and represented the internal dataset (1300 tweets). The
91 final collection stage also included selecting the tweets with at least one retweet. The two filters
92 were applied in order to obtain a relevant batch of posts related to vaccines, feasible for manual
93 annotation [20,21]. In addition, a random batch of 100 tweets from April 2021 were collected,
94 which represented the external validation dataset.

95 In order to obtain relevant information for the data analysis phase, the following parameters were
96 extracted for each tweet: date and time, tweet ID, tweet content, number of likes, number of
97 retweets, number of replies. All the information was anonymously collected through snsrape
98 package, which is based on the Twitter API [20].

99 **2.2. Manual annotation**

100 In order to analyze the collected Twitter posts in a relevant manner, all the tweets had to be
101 manually annotated. The tweets were classified in true, neutral or fake based on their text content.
102 The true classification (class 0) meant valid scientific information related to vaccines (whether it
103 was about Covid-19 vaccines or other vaccine types) or true general information regarding the
104 Romanian vaccination campaign. The neutral classification (class 1) regarded irrelevant, ironical
105 or other vaccine related comments, without manipulative or misleading intent. The fake

106 classification (class 2) referred to false or misleading information related to vaccines (both Covid-
107 19 and other types) or the Romanian vaccination campaign. The scientific validity of the posts,
108 when appropriate, was assessed in relation to the official sources of health information (such as
109 the European Centre for Disease Prevention and Control, the Summaries of Product Characteristics
110 of the vaccines approved in Romania or trusted health fact check websites) [23–27]. It should also
111 be noted that the majority of the tweets were related to Covid vaccines. However, the Twitter posts
112 related to other types of vaccines were not eliminated, in order to increase the variability and
113 complexity of the obtained dataset.

114 A total number of 9 annotators participated in the task. The external validation dataset was assessed
115 by all 9 annotators and the final classification of each tweet was obtained by a majority vote. The
116 internal dataset was annotated in a similar manner; however, due to the larger number of posts, the
117 internal data was split into 3 parts of similar number of tweets and each part was annotated by 3
118 different annotators and the final classification was established by a majority vote. Hence, all 9
119 annotators took part both in the external validation data and in the internal data annotation. In
120 addition, it should be mentioned that when a majority vote could not be applied (due to an equal
121 distribution of votes among the three classes), the tweet was annotated as neutral, in order to ensure
122 an unbiased data analysis. No tweet was eliminated when a majority vote could not be applied, in
123 order to enhance the variability of the processed vaccine data. The agreement between annotators
124 was established on the internal and external data by using Krippendorff's alpha coefficient. The
125 computation of the metric was considered relevant since it provides an ordinal option when
126 assessing the agreement [28]. Therefore, the differences between true and neutral annotations are
127 not penalized as hard as the differences between true and fake annotations.

128

129 **2.3. Text preprocessing**

130 In order to accurately analyze the annotated tweets, the text content had to be preprocessed. The
131 text preprocessing and machine learning development and validation were performed by using
132 Python Programming Language, version 3.9.2 [21].

133 In order to curate the text and obtain a simplified version, all special characters and stop words
134 were removed from the tweets and all letters were converted to lowercase. The standard stop word
135 list for Romanian provided by spacy was used. In addition, with the aim of providing a bias
136 reduction for the development of the machine learning algorithm, all hyperlinks and words starting
137 with the '@' symbol (with which the content of tweet replies begins) were also eliminated.
138 However, it should be noted that no lemmatization was performed on the selected tweets, since,
139 taking into consideration practical reasons, it was considered that different word forms might
140 provide different meaning and intent to specific phrases; moreover, as an example, as opposed to
141 English language, the Romanian language has a higher number of forms for the noun “vaccine”
142 and the verb “to vaccinate” [29,30].

143 **2.4. Preliminary analysis**

144 In order to characterize and extract relevant characteristics from the obtained dataset, a preliminary
145 analysis was performed, based on two important methods. The first one implied extracting the
146 most frequent single words and word combinations based on a wordcloud technique, in order to
147 offer a simplified and relevant visualization of the dataset The words were obtained for each of the
148 3 classes (true, neutral and fake) from the 1300 vaccine tweets. The second method was applied in
149 order to evaluate the relationships between the manual classification of the Twitter posts and other
150 characteristics. Hence the Spearman’s correlation coefficient, along with the p value for statistical

151 significance were computed between the manual classification (true – class 0, neutral – class 1,
152 fake – class 2) and each of the following characteristics of the 1300 tweets: number of replies,
153 number of retweets, number of likes and the length of the post, quantified by the number of
154 words [31].

155 **2.5. Building and validating the machine learning algorithm**

156 The machine learning algorithm was developed by using Python’s scikit-learn package (four
157 classical machine learning models: Support Vector Machines Classifier (SVM), Multilayer
158 Perceptron (MLP, a type of neural networks), Random Forest Classifier (RFC) and an ensemble
159 model (scikit-learn Voting Classifier), developed by averaging the probabilities which were
160 predicted by the SVM and the MLP), as well as Tensorflow (for two specific deep learning models:
161 recurrent convolutional neural networks (RCNN) – Tensorflow implementation and BERT – based
162 on a model which was pretrained on a Romanian 15 GB uncased text corpus, downloaded from
163 Huggingface (dumitrescustefan/bert-base-romanian-uncased-v1 model) and then executed
164 through Tensorflow) [32–34]. With regards to the classical machine learning models since scikit-
165 learn does not accept string data as input, the text content had to be converted to numerical data,
166 by using the TfidfVectorizer function. No words were eliminated from the text corpus when
167 performing the string-to-float conversion [34]. On the other hand, the deep learning models which
168 were implemented required specific word tokenizers. The RCNN model was built after using the
169 specific Tensorflow tokenizer, while the BERT model implemented the specific Romanian based
170 AutoTokenizer downloaded from the huggingface website [32,33].

171 The six machine learning algorithms were validated and compared on the obtained data. They were
172 tested based on their ability of estimating the probability that a specific tweet is true, neutral or
173 fake, as well as of correctly classifying a tweet as being true, neutral or fake. The Area Under the

174 Receiver Operating Characteristic Curve Score (ROC AUC Score, both a One-Versus-One (OVO)
175 strategy and a One-Versus-Rest (OVR) strategy) was used for testing the probability prediction
176 ability of the algorithms and was the most important overall measure for evaluating the machine
177 learning models: the higher the ROC AUC Score is, the better are the probabilities calibrated. In
178 addition, the accuracy, precision, recall, F1 Score and Matthews Correlation Coefficient were used
179 to test the classification ability of the developed models. The Matthews Correlation Coefficient
180 was considered the most important global classification measure, since it provides a relevant bias
181 reduction approach and takes into consideration class imbalance [34–36].

182 Both an internal and an external validation were performed for the machine learning algorithms.
183 The internal validation was performed on the 1300 tweets (internal dataset) and aimed at evaluating
184 the internal consistency of the model combined with the ability of perform on unseen data. Hence,
185 the dataset was split into 4 parts based on the 4 pandemic periods for which the posts were collected
186 (internal period validation). The predictive algorithms were validated 4 times: each time, the
187 training set included the tweets from 3 of the periods; the model was trained on the 3 periods and
188 was evaluated based on the unseen data from the 4th period. Therefore, the model was trained and
189 validated until all the 4 periods represented in turn the test set. In addition, a repeated 5-fold cross-
190 validation (with 10 iterations) was also performed [32–34]. However, the internal period validation
191 strategy was considered much more relevant than the cross-validation, since all tweets from a
192 specific period were either in the training set or the test set and the risk that the model was evaluated
193 on similar tweets was significantly reduced.

194 The external validation of the algorithms implied training the models on the internal data and
195 evaluating their performance on the external dataset represented by the 100 tweets from April
196 2021 [32–34].

197 Figure 1 briefly presents the three main strategies within the validation process of the machine
198 learning algorithms.

199 **Fig 1. The validation process of the machine learning algorithms**

200 The final model (SVM) for future identification of specific vaccine tweets was chosen based on
201 the best results obtained in terms of OVO and OVR ROC AUC Scores and was built by taking
202 into consideration the internal dataset (1300 tweets). The model was implemented based on a
203 probabilistic approach (useful for reliable probability estimation), a radial basis function (RBF)
204 kernel, a penalty parameter of the error term (C value) set to 1, while reducing bias caused by class
205 imbalance and breaking ties according to the confidence values of the RBF. In addition, a detailed
206 analysis was undertaken based on the probability predictions of the final model on 3 tweets from
207 the external data (one true post, one neutral post and one fake post) [34,37].

208 **3. Results**

209 **3.1. Data collection**

210 A total number of 1344 tweets were obtained, of which 44 were eliminated due to content in other
211 languages. An additional 100 tweets were randomly selected from another period (April 2021,
212 from the 7-day period with the highest number of tweets, independent on the number of retweets)
213 and represented the dataset for external validation. Table 1 presents the final 7-day periods from
214 which the posts were collected, as well as the number of tweets for each weekly time interval.

215

216

217

218 **Table 1. The time periods corresponding to the collected vaccine related tweets**

Period	Number of tweets
Internal data	
March 20, 2020 – March 26, 2020	48
January 15, 2021 – January 21, 2021	491
May 4, 2021 – May 10, 2021	322
October 19, 2021 – October 25, 2021	439
	Total: 1300 tweets
External validation data	
April 10, 2021 – April 16, 2021	100

219

220 **3.2. Manual annotation**

221 The manual annotation yielded an average inter-agreement Krippendorff's alpha of 0.64 for the
222 internal dataset (0.69 for Team 1, 0.58 for Team 2 and 0.64 for Team 3) and of 0.7 for the external
223 dataset. After applying the majority vote rule, from the 1300 tweets (internal dataset), a total
224 number of 488 (37.5%) were classified as true, 373 (28.7%) as neutral and 439 (33.8%) as fake.
225 From the 100 tweets representing the external dataset, 53 (53%) were classified as true, 24 (24%)
226 as neutral and 23 (23%) as fake.

227 In terms of overall inter-annotator agreement, from the 1300 tweets, 686 (52.8%) reached perfect
228 agreement between the 3 annotators. From the 100 tweets from the external dataset, 15 (15%)
229 reached perfect agreement between all 9 annotators.

231 **3.3. Preliminary analysis**

232 Table 2 presents the most relevant words and word combinations for the true, neutral and fake
233 tweets within the internal dataset. The most relevant 7 words and word combinations (as
234 considered by the annotators) of the most frequent 30 are presented. The words were translated
235 from Romanian to English and the original Romanian version is also presented in parenthesis,
236 when appropriate. Table 3 summarizes the results obtained by computing the Spearman's
237 correlation coefficient. The p values are not given, since all pairs yielded statistically significant
238 results ($p < 0.05$). Wordcloud representations of the most relevant words written in Romanian and
239 graphical illustration of correlation analysis are shown in Figure 2.

240 **Table 2. Most relevant words and word combinations identified through a wordcloud**
241 **model for each of the 3 classes**

Class	Most relevant words/word combinations
True	Covid-19 “News” AND “Romania” Against Covid (împotriva Covid) Vaccinated persons (personae vaccinate) Vaccine dose (doze de vaccin) Vaccination campaign (campania de vaccinare) Vaccination centers (centre de vaccinare)
Neutral	Vaccinated (vaccinat) Against Covid (împotriva Covid)

	<p>Covid-19</p> <p>“News” AND “Romania”</p> <p>Vaccination campaign (campania de vaccinare)</p> <p>Klaus Iohannis</p> <p>Vaccination certificate (certificat de vaccinare)</p>
<p>Fake</p>	<p>Vaccinated (vaccinat)</p> <p>Against Covid (împotriva Covid)</p> <p>To give birth (adus pe lume) (aggressive connotation)</p> <p>“Persons” AND “died” (“persoane” ȘI “murit”</p> <p>Adverse reactions (reacții adverse)</p> <p>Experimental vaccine (vaccin experimental)</p> <p>Mandatory vaccination (vaccinarea obligatorie)</p>

242

243

244

245

246

247

Table 3. Correlation analysis results (Spearman's correlation coefficient)

Parameter	Class	Number of replies	Number of retweets	Number of likes	Number of words
Class	1.000	0.198	0.190	0.282	0.222
Number of replies	0.198	1.000	0.453	0.654	0.117
Number of retweets	0.190	0.453	1.000	0.601	0.099
Number of likes	0.282	0.654	0.601	1.000	0.150
Number of words	0.222	0.117	0.099	0.150	1.000

248

249 **Fig 2. Wordcloud representation (the most relevant 30 words written in Romanian) for the**
250 **tweets labelled as true (A), neutral (B) and fake (C); Correlation analysis results (D)**

251 **3.4. The validation of the machine learning algorithms**

252 Performance metrics obtained after validating the 6 predictive algorithms (SVM, MLP, RF, the
253 ensemble model – SVM + MLP, RCNN and BERT) are shown in Table 4. All 3 validation types
254 are presented – repeated cross-validation, internal period validation and external validation.

255

256

Table 4. Validation results for the machine learning algorithms

Cross-validation						
Metric	SVM	MLP	RF	Ensemble (SVM + MLP)	RCNN	BERT
Classification validation						
Accuracy	0.657	0.603	0.595	0.614	0.623	0.689
Precision	0.641	0.587	0.600	0.598	0.609	0.693
Recall	0.634	0.588	0.593	0.599	0.602	0.667
F1 Score	0.632	0.585	0.583	0.596	0.599	0.658
Matthews Correlation Coefficient	0.480	0.400	0.401	0.416	0.425	0.535
Probability prediction validation						
ROC AUC Score (OVO)	0.813	0.782	0.779	0.797	0.770	0.849
ROC AUC Score (OVR)	0.825	0.788	0.783	0.803	0.779	0.858
Internal period validation						
Metric	SVM	MLP	RF	Ensemble (SVM + MLP)	RCNN	BERT
Classification validation						
Accuracy	0.567	0.546	0.551	0.576	0.537	0.601
Precision	0.572	0.539	0.554	0.562	0.525	0.632
Recall	0.552	0.536	0.526	0.557	0.519	0.573
F1 Score	0.532	0.520	0.515	0.542	0.512	0.539

Matthews Correlation Coefficient	0.352	0.317	0.308	0.354	0.291	0.416
Probability prediction validation						
ROC AUC Score (OVO)	0.744	0.738	0.727	0.745	0.702	0.787
ROC AUC Score (OVR)	0.756	0.743	0.732	0.754	0.710	0.797
External validation						
Metric	SVM	MLP	RF	Ensemble (SVM + MLP)	RCNN	BERT
Classification validation						
Accuracy	0.680	0.648	0.490	0.688	0.480	0.670
Precision	0.661	0.581	0.480	0.636	0.529	0.622
Recall	0.691	0.595	0.465	0.649	0.499	0.630
F1 Score	0.655	0.583	0.456	0.638	0.470	0.606
Matthews Correlation Coefficient	0.530	0.435	0.231	0.504	0.271	0.490
Probability prediction validation						
ROC AUC Score (OVO)	0.818	0.772	0.736	0.800	0.718	0.806
ROC AUC Score (OVR)	0.843	0.796	0.756	0.826	0.727	0.829

258

259 The distribution of predicted probabilities generated with all 6 predictive models for the external

260 dataset is illustrated in Figure 3.

261 **Fig 3. Boxplot representation of the predicted probabilities for the machine learning**
262 **algorithms obtained on the external dataset: SVM algorithm (A), MLP algorithm (B), RF**
263 **algorithm (C), Ensemble algorithm (SVM + MLP) (D), RCNN algorithm (E), BERT**
264 **algorithm (F)**

265 **3.5. Implementation example of the SVM algorithm**

266 The final algorithm chosen for implementation was the Support Vector Classifier, due to its
267 enhanced predictions quantified through the ROC AUC Score (Table 4). Table 5 presents the
268 probabilities returned by the algorithm for 3 tweets from the external dataset (one true tweet, one
269 neutral tweet and one fake tweet). In order to comply with the General Data Protection Regulation,
270 the exemplified tweets were translated and partially reformulated. In addition, in order to allow a
271 better understanding and exemplification of tweet structure and machine learning predictive
272 abilities, the probabilities for nine extra tweets are presented in supplementary S1 Table (three true
273 tweets, three neutral tweets and three fake tweets).

274

275

276

277

278

279

280

281 **Table 5. Detailed example of implementation of SVM algorithm on 3 tweets from the**
 282 **external dataset**

Reformulated tweet content	Predicted probability (true)	Predicted probability (neutral)	Predicted probability (fake)	Predicted class	Annotated class
Tweet A: At 10 days after the second Covid vaccine shot, the risk of getting infected is very low.	72.56%	6.79%	20.66%	Class 0 (true)	Class 0 (true)
Tweet B: I got the vaccine, mind your own business.	9.94%	43.38%	46.68%	Class 2 (fake) erroneous prediction	Class 1 (neutral)
Tweet C: Mass vaccination would be catastrophic for humankind.	11.60%	23.02%	65.37%	Class 2 (fake)	Class 2 (fake)

283
 284 The validation of the machine learning predictive algorithms yielded modest results in terms of
 285 classification evaluation (Table 4). The Matthews Correlation Coefficient, considered the most
 286 important metric used to assess the discriminative power of the implemented models, yielded

287 values ranging from 0.4 to 0.535 for the cross-validation technique, from 0.308 to 0.416 for the
288 internal period validation, as well as from 0.231 to 0.53 for the external validation of the developed
289 algorithms. Overall, by averaging the 2 types of internal validation, BERT resulted in the highest
290 Matthews Correlation Coefficient, of 0.535 for the cross-validation and 0.416 for the internal
291 period validation, with a 5.5% increase for cross-validation, as well as a 6.2% increase for internal
292 period validation as compared to SVM (which yielded a 0.48 Matthews Correlation Coefficient
293 for the cross-validation and a 0.416 value for the internal period validation). However, it should
294 be noted that on the external validation, the SVM algorithm outperformed BERT in terms of raw
295 classification ability, with a 0.53 Matthews Correlation Coefficient, while BERT yielded a value
296 of 0.49 for this validation metric. Since the Matthews Correlation Coefficient is a particular case
297 of the Pearson product-moment correlation coefficient, its values have the same interpretation and
298 hence it can be stated that in most cases, the validation of the BERT and SVM algorithms (the best
299 models with regards to the internal and external validation respectively) yielded moderate to
300 moderately high positive correlations between the true and predicted labels [35].

301 Nevertheless, the most important evaluation of the machine learning models was represented by
302 the probability prediction evaluation, which tested the ability of the algorithms of estimating well
303 calibrated probabilities, as quantified through the ROC AUC Score (using both an OVO and an
304 OVR approach). As for the Matthews Correlation Coefficient, the ROC AUC Score yielded the
305 lowest results for the internal period validation (OVO ROC AUC Score ranged from 0.702 to
306 0.787; OVR ROC AUC Score ranged from 0.71 to 0.797), followed by the external validation
307 (OVO ROC AUC Score ranged from 0.718 to 0.818; OVR ROC AUC Score ranged from 0.727
308 to 0.843), while the cross-validation resulted in the highest ROC AUC values (OVO ROC AUC
309 Score ranged from 0.77 to 0.849; OVR ROC AUC Score ranged from 0.779 to 0.858). Similar to

310 the results obtained in terms of Matthews Correlation Coefficient, the highest ROC AUC Scores
311 were obtained in case of BERT for both internal validation strategies (for cross-validation: OVO
312 ROC AUC = 0.849, OVR ROC AUC = 0.858; for internal period validation: OVO ROC AUC =
313 0.787, OVR ROC AUC = 0.797), followed by the SVM algorithm. On the other hand, similar to
314 the raw classification validation, SVM resulted in improved results for the ROC AUC Scores when
315 taking into consideration the external validation (OVO ROC AUC = 0.818, OVR ROC AUC =
316 0.843, as opposed to a 0.806 value in case of OVO ROC AUC and a 0.829 OVR ROC AUC for
317 the BERT model). The enhanced results which were obtained for internal validation in case of
318 BERT might be explained that the current study implemented a pre-trained BERT model based on
319 a large Romanian text corpus of 15 GB. Nevertheless, BERT validation yielded less accurate
320 results than the SVM when taking the external tweets dataset into consideration, which might have
321 been caused by a moderate amount of overfitting on the internal data (1300 tweets), as well as by
322 the low level of complexity of the processed tweets. In addition, the RCNN model implemented
323 through the Tensorflow library provided poor results both in terms of raw classification and
324 probability estimation (0.702-0.710 ROC AUC Scores for internal period validation and 0.718-
325 0.727 for external validation), which were in most cases the lowest of all 6 implemented machine
326 learning models. These results were obtained despite the high complexity of RCNN and its ability
327 to memorize both temporal and spatial relationships from texts. One reason for the poor results
328 might be related to the relatively short posts which are usually distributed through the Twitter
329 platform and to the fact that the RCNN, in contrast to the implemented BERT model, lacked a
330 specific Romanian based text corpus and didn't include any pretrained algorithm. Moreover, we
331 argue that a complex model architecture (with both recurrent and convolutional layers), without

332 any predefined recommendations, is difficult to model so that it reaches optimal results on a text
333 corpus which contains posts in a narrowly spoken language, such as Romanian [32,33].

334 In terms of analysis of predicted probabilities (for the external dataset) quantified through the
335 boxplot representations (Figure 3), both SVM (Figure 3(A)) and BERT (Figure 3(F)) offered good
336 discrimination when comparing the estimated probabilities with the true (annotated) class.
337 However, the main difference in the performance of the two models can be seen in the probability
338 estimation for the tweets labelled as neutral. More specifically, the SVM offered a more accurate
339 discrimination when predicting the probabilities that the neutral tweets from the external dataset
340 are true, neutral or fake, the probability of being neutral being higher on average than the
341 probability that the tweet was true or fake, which was also reflected in the lower ROC AUC Scores
342 for BERT, when compared to SVM. By contrast, the BERT model returned on average a higher
343 probability that the neutral tweets are fake, as compared to neutral. However, the BERT model
344 discriminated more accurately between the true tweets, as well as the fake tweets and the rest
345 (neutral/fake and true/neutral, respectively), while the SVM offered a more close, but still valid
346 discrimination.

347 **4. Discussion**

348 A detailed analysis of a batch of relevant vaccine tweets from several periods within the Covid-19
349 pandemic was undertaken.

350 The preliminary analysis implied the manual annotation of a total number of 1400 tweets, as well
351 as a preliminary analysis for establishing specific word patterns within the posts and the
352 correlations between the manual annotation and other tweet characteristics. The supervised
353 analysis consisted of building and validating several machine learning prediction models based on

354 their ability of estimating the probabilities that a specific Twitter post related to vaccines is true,
355 neutral or fake.

356 The manual annotation of the collected Twitter posts yielded good results in terms of inter-
357 agreement evaluation based on Krippendorff's alpha [38]. The inter-agreement was better for the
358 external dataset (100 tweets, Krippendorff = 0.7) than for the internal one (1300 tweets, average
359 Krippendorff = 0.64), partly due to the fact that the tweets from the external dataset were annotated
360 by all 9 annotators. Moreover, the Krippendorff obtained for each of the 3 subsets within the
361 internal data showed a certain degree of variability, with its values ranging from 0.58 to 0.69.
362 Indeed, as with other social media posts, the ones from Twitter, even when relating to health issues,
363 are written in a free, subjective manner, since they are mostly written by individual persons which
364 are granted the freedom of expression [39]. Therefore, there is a high probability that the annotators
365 ran into several ambiguous tweets and hence the interpretation of such content could have been
366 made different depending on the content and the annotator's subjective interpretation.

367 In addition, the subjective and diverse ways in which the vaccines posts were written are
368 emphasized in Table 2, where the most relevant word patterns within the 3 classes (true, neutral
369 and fake) are given. Interestingly, the true posts contained most often different forms of the noun
370 „vaccine” and the verb „to vaccinate”, which could be explained by the fact that the true posts,
371 when compared to the neutral and fake ones, contained the most references to news articles and to
372 official data related to the Romanian vaccination campaign, such as the number of persons which
373 were partially and fully vaccinated within a specific time period, the number of administered
374 vaccine doses or the updated vaccine supply. In contrast, the tweets which were labelled as fake
375 (false or manipulative information) also contained many references to the various forms of
376 „vaccine” and „to vaccinate”, but quite often they were referenced in a subjective and manipulative

377 manner. As an example, the word combination „experimental vaccine” was identified as one of
378 the most relevant patterns of the fake tweets, suggesting that the Covid vaccines were not tested
379 enough before being administered in the general population, information which is false, according
380 to various health authorities and fact check websites. One of the main reasons of propagating such
381 misinformation would be to make the population believe that the vaccines are dangerous for health
382 and cause many severe adverse reactions [23,24]. In addition, words with aggressive connotation
383 were more frequent within the tweets labelled as fake, as compared to the true and neutral ones.

384 In terms of correlation analysis results, the majority of obtained Spearman’s coefficients showed
385 moderate, but statistically significant correlations (Table 3, Figure 2(D)). The manual
386 classification (considered as class 0 – true, class 1 - neutral, class 2 – fake) was positively
387 correlated with all of the three tweet characteristics: number of replies ($r = 0.198$), number of
388 retweets ($r = 0.190$) and number of likes ($r = 0.282$). These results, even though suggesting a
389 modest positive correlation, imply that the fake vaccine tweets have a higher impact on social
390 media, tending to be retweeted and liked more often than the true and neutral ones (this might in
391 turn prioritize vaccine false information even more because of the Twitter algorithm) [40]. The
392 results are similar to the ones reflected in other studies by analyzing the online spread of
393 misinformation [41–43]. For example, the work conducted by Vosoughi et al found among 126000
394 stories related to various topics that the ones labelled as false misinformation had a more
395 pronounced spread on Twitter as compared to the valid ones [42]. Even though the effect was more
396 pronounced for information about politics, the study raises important awareness, especially
397 considering the fact that several studies show that the online spread of health misinformation may
398 be, at least partially, politically driven [43]. These observations, combined with the fact that online
399 anti-vaccine groups and accounts are more strongly connected and more likely to influence those

400 with neutral views, make enforcing policies on limiting the spread of vaccine misinformation of
401 an utmost importance [5,6].

402 Based on the OVO and OVR ROC AUC Score values which were computed for the external
403 validation of the machine learning models (Table 4, Figure 3), the SVM algorithm was chosen for
404 building the final predictive model [34]. In addition, all ROC AUC Scores which were obtained
405 when validating the SVM algorithm were above 0.74, which proved the well calibrated
406 probabilities returned by the model. Hence, from a practical point of view, the final SVM model
407 could be used for future identification of the most relevant vaccine related Twitter posts, by sorting
408 the automatically collected large tweet lists based on the predicted probabilities that the specific
409 posts represent true, neutral or fake content. The obtained information, after manually analyzed, if
410 presented through a web platform, could further aid in raising awareness regarding valid
411 information, fake news content, as well as irrelevant information related to vaccines and shared
412 through Twitter platform [20,34]. With regards to the machine learning validation results obtained
413 in other studies, a relevant comparison with the ones from the current studies would be difficult,
414 since the majority of the studies which used vaccine related Twitter content reported the F1 Score
415 as the most important classification evaluation metric [12–17,36,44,45]. The F1 Score was
416 computed in the current study as well and can be regarded as an acceptable balanced measure
417 between precision and recall. However, as highlighted by Chicco and Jurman, F1 Score can
418 provide overoptimistic results when evaluating the performance of a predictive model [36]. That
419 is the main reason for which the focus in the current study, when evaluating the classification
420 performance of the 6 algorithms, was put on the Matthews Correlation Coefficient [35,36].
421 Moreover, the reported F1 Scores showed a high degree of variability, some studies reported F1
422 Scores of under 0.6, while others reported enhanced values, of 0.7-0.8 and others obtained almost

423 perfect values, of over 0.95, while the implemented machine learning algorithms included both
424 classical (such as Random Forest and SVM) and newer model types (such as deep learning and
425 BERT) applied on various languages, such as English, French, Dutch or Moroccan [12–17,44,45].
426 As a comparison, the maximum F1 Score which was obtained in the current study SVM ranged
427 from 0.542 (internal period validation – SVM+MLP Ensemble) to 0.658 (cross-validation –
428 BERT) and 0.655 (external validation - SVM). The obtained F1 Score is therefore smaller than
429 that reported by most of the studies; however, as was already mentioned, the most important
430 validation metric in our study, the ROC AUC Score, yielded maximum values of over 0.8, which
431 translates into well calibrated probabilities [34].

432 Another relevant example is a study which implemented an algorithm based on recurrent
433 convolutional neural networks, with BERT as a word embedding model [46]. Even though it did
434 not use the F1 Score as evaluation metric, it achieved superior accuracy, of 0.989, when tested on
435 a real-world dataset, which contains real and fake news propagated during the US General
436 Presidential Election from 2016, with over 20000 instances, both Twitter and Facebook being
437 widely used for disinformation purpose [46,47]. Other research which aimed at detecting social
438 media non-vaccine related disinformation implemented a hybrid deep learning model (based on
439 recurrent neural networks) and achieved a F1 Score of 0.894, lower than in the study which
440 specifically used BERT [46,48]. As a comparison, in our study, we obtained an accuracy of 0.601-
441 0.689 when evaluating the BERT model; however, our dataset was much smaller and was
442 specifically related to vaccine information distributed through Twitter.

443 In terms of studies which performed unsupervised analysis on specific disinformation propagated
444 through Twitter, the research performed by Kobayashi et al is worth mentioning. It included 100
445 million vaccine related Japanese tweets, on which a topic, as well as a time series analysis were

446 performed [49]. In addition, with respect to other studies which specifically evaluated social media
447 disinformation, a study, conducted by De Clerck analyzed the general spread of disinformation
448 through Twitter platform by taking into consideration numerous countries included in the Twitter
449 information operations report. It proposed maximum entropy networks for identifying and
450 quantifying specific patterns in the interactions between numerous Twitter users which might have
451 had an important impact on spread of disinformation (whether or not health related). The analysis
452 had the advantage of applying various algorithms and including a large number of tweets from
453 different countries (e.g. Armenia, China, Russia, Serbia, Turkey) [50]. While our study did not
454 implement any form of unsupervised analysis, we argue that the wordcloud representation and
455 correlation analysis which were undertaken give context to the implemented and publicly available
456 machine learning model.

457 Regarding the practical implementation of the SVM model (Table 5), the given examples provide
458 relevant insights regarding the Romanian tweets structure, as well as the predictive algorithm use
459 case. The first tweet (Tweet A) refers to a valid scientific information – indeed, especially
460 considering the fact that the post was written in April 2021, when the highly contagious Omicron
461 variant and its subvariants were not circulating, two doses of either the RNA or the viral vector
462 vaccine (the ones which were available within the Romanian Vaccination Campaign) significantly
463 reduced the risk of symptomatic Covid-19 [24,25]. The predictive model accurately estimated a
464 72.56% probability that the content is true, with only 20.66% chance of being misinformation and
465 6.79% of being neutral. The second tweet (Tweet B) was manually labelled as neutral, being an
466 irrelevant statement regarding someone who got the Covid vaccine. However, the SVM algorithm
467 erroneously classified the tweet as being fake, possibly due to the fact that it was written in a
468 slightly aggressive manner. Nonetheless, when analyzing the predicted probabilities, the model

469 returned a 43.38% risk that the content is neutral and a 46.68% risk that the tweet refers to false
470 information, with only 3.3% higher than the probability of containing neutral information. The
471 third tweet given as a practical example (Tweet C) was manually labelled as false information
472 (fake). The algorithm returned the same classification, with a 65.37% probability that the content
473 is fake, a 23.02% probability that it is neutral and a 11.60% chance of being true. The content of
474 the tweet is a classical conspiracy theory, which tries to suggest that mass vaccination is not only
475 unnecessary, but detrimental. The information is obviously false: the essential role of vaccines in
476 leading to herd immunity and controlling infectious diseases is well established [24].

477 The current study has a few important advantages. First of all, to our knowledge, this is the first
478 study analyzing vaccine fake news written in Romanian from social media posts. While Romanian
479 is a narrowly spoken language, limited to Romania and Republic of Moldova, we argue that by
480 providing the detailed Python code which includes the specified analyses and the developed
481 predictive machine learning algorithm, as well as the processed (annotated, vectorized and
482 anonymized) internal and external data, our work could be used by other researchers in future
483 studies, with easy translation to other languages [21,51].

484 Secondly, as a difference from other similar studies, which used two classes (such as general
485 information and misinformation) during the data labelling process, the current work used for
486 manual annotation three classes (true, neutral and fake) [12,13,15–17]. It can be argued that this
487 approach enhances the complexity of machine learning models and provides context to the social
488 media analysis. In addition, besides the raw classification, the machine learning models which
489 were developed provide probability estimates, a relevant feature which may aid in future selection
490 of relevant vaccine tweets based on approaches which imply sorting the predicted probabilities,
491 such as the ones presented in Table 5 and Supplementary Table 1. The predictive algorithms were

492 validated in a consistent manner, both for classification and probability estimation. Relevant
493 validation strategies were implemented: the internal period validation ensured the internal
494 consistency of the models with regards to performing on tweets from different pandemic periods,
495 while the external validation ensured the evaluation of the algorithms on unseen data (Table 4,
496 Figure 4) [32–34].

497 Nevertheless, the current work has a series of limitations. First of all, the number of collected and
498 annotated tweets (1300 – internal dataset, 100 – external dataset) can be regarded as very low when
499 compared to other studies (therefore, the variability and complexity of the developed SVM
500 algorithm could have been negatively impacted) [12–17,37]. For example, Kunneman et al
501 conducted a study for measuring the stance towards vaccination (non-Covid vaccines: the
502 messages were extracted prior to the pandemic period), based on a total number of 8259 annotated
503 tweets written in Dutch; however, the study only achieved a Krippendorff’s alpha between 0.27
504 and 0.35, significantly lower than that from the current study [14]. However, Hayawi et al
505 undertook a vaccine misinformation analysis based on 15073 annotated English tweets; the
506 annotation process had the advantage of being further validated by health experts and also lead to
507 very good machine learning validation metrics (0.97 precision, 0.98 recall, 0.98 F1 Score) [17].
508 Other studies focused on Covid-19 vaccine hesitancy; while they initially automatically collected
509 large numbers of vaccine related tweets (for example, written in English, Turkish or French), the
510 manual analysis of the content implied, as in our study, a small number of tweets (approximately
511 1000-2000) [18,19,44,45]. Therefore, it should be noted that while our study comprised indeed in
512 a small dataset chosen for annotation, the fact that the tweets were chosen and annotated following
513 a standardized methodology (selecting 4 relevant pandemic periods and eliminating the tweets
514 with no retweets, as well as the fact that each Twitter post was classified by at least 3 annotators)

515 could ensure reproducibility, especially considering the fact that the Python code for data
516 preprocessing, wordcloud representation, correlation analysis and the development and validation
517 of the machine learning predictive models, as well as the Tfidf vectorized dataset and the final
518 SVM algorithm are publicly available at [https://github.com/valeanuandrei/vaccine-tweets-ro-](https://github.com/valeanuandrei/vaccine-tweets-ro-research)
519 [research](https://github.com/valeanuandrei/vaccine-tweets-ro-research) [51].

520 Secondly, even though the results of the probability validation were satisfactory, the evaluation of
521 the classification ability of the machine learning algorithms, especially for the internal period
522 validation (a maximum Matthews Correlation Coefficient of under 0.42 and a maximum F1 Score
523 of under 0.55), yielded modest results [35].

524 Therefore, the implemented natural language processing and data mining techniques, combined
525 with the 12 practical examples of tweet classification and probability prediction, provide relevant
526 insights regarding vaccine general information and misinformation spread through Twitter
527 platform and written in Romanian. Future studies must aim at collecting a large number of tweets
528 and classifying them based on a semi-supervised approach, in order to enhance the variability,
529 complexity and predictive ability of the machine learning algorithm. After these steps are
530 undertaken, an online platform might be developed, based on identifying new vaccine related
531 Twitter content, to aid in raising awareness regarding the vaccine misinformation shared through
532 social media and consequently reduce vaccine hesitancy [52,53].

533 **5. Conclusions**

534 A study aiming at analyzing and automatically classifying relevant vaccine related posts from
535 Twitter content was undertaken. A total number of 1400 tweets from relevant pandemic periods
536 were collected and manually classified as true information, neutral information or fake information

537 related to vaccines. Both an unsupervised analysis (consisting of a wordcloud evaluation and a
538 correlation analysis) and a supervised analysis (based on building several predictive machine
539 learning algorithms – SVM, MLP, RF, an ensemble voting classifier: SVM + MLP, as well as
540 complex deep learning models: RCNN and BERT) were implemented. The correlation analysis
541 yielded moderate, but significant positive correlations between the tweets labelled as
542 misinformation and the tweet engagement metrics, quantified through the number of replies,
543 retweets and likes. The machine learning algorithms were mainly validated based on their ability
544 of estimating the probability that a specific tweet is true, neutral or fake. The optimal results were
545 obtained for the Support Vector Classifier, with a ROC AUC Score ranging from 0.744 to 0.843
546 and BERT, with a ROC AUC Score ranging from 0.787 to 0.858. Future studies must aim in
547 enlarging the vaccine tweets database and optimizing the machine learning predictive abilities, in
548 order to automatically identify and classify new vaccine related valid, neutral and false information
549 distributed through Twitter platform.

550 **References**

- 551 1. Li X, Mukandavire C, Cucunubá ZM, Echeverria Londono S, Abbas K, Clapham HE, et
552 al. Estimating the health impact of vaccination against ten pathogens in 98 low-income
553 and middle-income countries from 2000 to 2030: a modelling study. *The Lancet*.
554 2021;397: 398–408. doi:10.1016/S0140-6736(20)32657-X
- 555 2. Garon JR, Cochi SL, Orenstein WA. The Challenge of Global Poliomyelitis Eradication.
556 *Infect Dis Clin North Am*. 2015;29: 651–665. doi:10.1016/j.idc.2015.07.003
- 557 3. Centers for Disease Control and Prevention. Smallpox. 2022 [cited 14 Jul 2023].
558 Available: <https://www.cdc.gov/smallpox/index.html>

- 559 4. Meslé MM, Brown J, Mook P, Hagan J, Pastore R, Bundle N, et al. Estimated number of
560 deaths directly averted in people 60 years and older as a result of COVID-19 vaccination
561 in the WHO European Region, December 2020 to November 2021. *Eurosurveillance*.
562 2021;26. doi:10.2807/1560-7917.ES.2021.26.47.2101021
- 563 5. Johnson NF, Velásquez N, Restrepo NJ, Leahy R, Gabriel N, El Oud S, et al. The online
564 competition between pro- and anti-vaccination views. *Nature*. 2020;582: 230–233.
565 doi:10.1038/s41586-020-2281-1
- 566 6. Center for Countering Digital Hate. The Disinformation Dozen. 2021 [cited 14 Jul 2023].
567 Available: <https://counterhate.com/research/the-disinformation-dozen/>
- 568 7. Allington D, McAndrew S, Moxham-Hall VL, Duffy B. Media usage predicts intention to
569 be vaccinated against SARS-CoV-2 in the US and the UK. *Vaccine*. 2021;39: 2595–2603.
570 doi:10.1016/j.vaccine.2021.02.054
- 571 8. Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. Measuring the impact of
572 COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat Hum*
573 *Behav*. 2021;5: 337–348. doi:10.1038/s41562-021-01056-1
- 574 9. Popa AD, Enache AI, Popa IV, Antoniu SA, Dragomir RA, Burlacu A. Determinants of
575 the Hesitancy toward COVID-19 Vaccination in Eastern European Countries and the
576 Relationship with Health and Vaccine Literacy: A Literature Review. *Vaccines (Basel)*.
577 2022;10: 672. doi:10.3390/vaccines10050672
- 578 10. Zhou X, Coiera E, Tsafnat G, Arachi D, Ong M-S, Dunn AG. Using social connection
579 information to improve opinion mining: Identifying negative sentiment about HPV
580 vaccines on Twitter. *Stud Health Technol Inform*. 2015;216: 761–5.

- 581 11. Shapiro GK, Surian D, Dunn AG, Perry R, Kelaher M. Comparing human papillomavirus
582 vaccine concerns on Twitter: a cross-sectional study of users in Australia, Canada and the
583 UK. *BMJ Open*. 2017;7: e016869. doi:10.1136/bmjopen-2017-016869
- 584 12. Abdul-Mageed M, Elmandany AR, Pabbi D, Verma K, Lin R. Mega-COV: A billion-scale
585 dataset of 100+ languages for COVID-19. Preprint at <https://arxiv.org/abs/200506012>.
586 2020.
- 587 13. Cui L, Lee D. CoAID: COVID-19 Healthcare Misinformation Dataset. Preprint at
588 <https://arxiv.org/abs/200600885>. 2020.
- 589 14. Kunneman F, Lambooi M, Wong A, Bosch A van den, Mollema L. Monitoring stance
590 towards vaccination in twitter messages. *BMC Med Inform Decis Mak*. 2020;20: 33.
591 doi:10.1186/s12911-020-1046-y
- 592 15. Madani Y, Erritali M, Bouikhalene B. Using artificial intelligence techniques for detecting
593 Covid-19 epidemic fake news in Moroccan tweets. *Results Phys*. 2021;25: 104266.
594 doi:10.1016/j.rinp.2021.104266
- 595 16. To QG, To KG, Huynh V-AN, Nguyen NTQ, Ngo DTN, Alley SJ, et al. Applying
596 Machine Learning to Identify Anti-Vaccination Tweets during the COVID-19 Pandemic.
597 *Int J Environ Res Public Health*. 2021;18: 4069. doi:10.3390/ijerph18084069
- 598 17. Hayawi K, Shahriar S, Serhani MA, Taleb I, Mathew SS. ANTi-Vax: a novel Twitter
599 dataset for COVID-19 vaccine misinformation detection. *Public Health*. 2022;203: 23–30.
600 doi:10.1016/j.puhe.2021.11.022

- 601 18. Küçükali H, Ataç Ö, Palteki AS, Tokaç AZ, Hayran O. Vaccine Hesitancy and Anti-
602 Vaccination Attitudes during the Start of COVID-19 Vaccination Program: A Content
603 Analysis on Twitter Data. *Vaccines (Basel)*. 2022;10: 161. doi:10.3390/vaccines10020161
- 604 19. Lanyi K, Green R, Craig D, Marshall C. COVID-19 Vaccine Hesitancy: Analysing
605 Twitter to Identify Barriers to Vaccination in a Low Uptake Region of the UK. *Front Digit*
606 *Health*. 2022;3. doi:10.3389/fdgth.2021.804855
- 607 20. JustAnotherArchivist (GitHub Repository). Snsrape. 2022 [cited 17 Jul 2023]. Available:
608 <https://github.com/JustAnotherArchivist/snsrape>
- 609 21. Python Software Foundation. Python Language Reference, version 3.9.2. 2021 [cited 17
610 Jul 2023]. Available: <http://www.python.org>
- 611 22. Ritchie H. Coronavirus Pandemic (COVID-19). In: *Our World in Data* [Internet]. 2020
612 [cited 17 Jul 2023]. Available: <https://ourworldindata.org/coronavirus>
- 613 23. European Centre for Disease Prevention and Control. Covid-19. 2022 [cited 14 Jul 2023].
614 Available: <https://www.ecdc.europa.eu/en/covid-19>
- 615 24. The Annenberg Public Policy Center. COVID-19 Misconceptions. 2022 [cited 14 Jul
616 2023]. Available: <https://www.factcheck.org/covid-misconceptions/>
- 617 25. European Medicines Agency. Treatments and vaccines for COVID-19. 2022 [cited 17 Jul
618 2023]. Available: [https://www.ema.europa.eu/en/human-regulatory/overview/public-
619 health-threats/coronavirus-disease-covid-19/treatments-vaccines-covid-19](https://www.ema.europa.eu/en/human-regulatory/overview/public-health-threats/coronavirus-disease-covid-19/treatments-vaccines-covid-19)
- 620 26. Elhadad MK, Li KF, Gebali F. Detecting Misleading Information on COVID-19. *IEEE*
621 *Access*. 2020;8: 165201–165215. doi:10.1109/ACCESS.2020.3022867

- 622 27. Shu K, Mahudeswaran D, Wang S, Lee D, Liu H. FakeNewsNet: A Data Repository with
623 News Content, Social Context, and Spatiotemporal Information for Studying Fake News
624 on Social Media. *Big Data*. 2020;8: 171–188. doi:10.1089/big.2020.0062
- 625 28. Python Software Foundation. krippendorff 0.5.1. 2022 [cited 17 Jul 2023]. Available:
626 <https://pypi.org/project/krippendorff/>
- 627 29. Romanian Academy (“Iorgu Iordan” Linguistic Institute). The Romanian Explanatory
628 Dictionary. ed. Univers Enciclopedic Gold; 2016.
- 629 30. Merriam-Webster Inc. Merriam-Webster Online Dictionary. 2022 [cited 17 Jul 2023].
630 Available: <https://www.merriam-webster.com/>
- 631 31. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy
632 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17:
633 261–272. doi:10.1038/s41592-019-0686-2
- 634 32. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-
635 Scale Machine Learning on Heterogeneous Distributed Systems. 2016.
- 636 33. Dumitrescu SD, Avram A-M, Pyysalo S. The birth of Romanian BERT. 2020.
- 637 34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn:
638 Machine Learning in Python. 2012.
- 639 35. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness,
640 markedness and correlation. 2020.

- 641 36. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over
642 F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21: 6.
643 doi:10.1186/s12864-019-6413-7
- 644 37. Scikit-learn developers. Sklearn.svm.SVC. 2022 [cited 17 Jul 2023]. Available:
645 <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVM.html>
- 646 38. Zapf A, Castell S, Morawietz L, Karch A. Measuring inter-rater reliability for nominal
647 data – which coefficients and confidence intervals are appropriate? *BMC Med Res*
648 *Methodol*. 2016;16: 93. doi:10.1186/s12874-016-0200-9
- 649 39. Twitter Inc. Defending and respecting the rights of people using our service. 2022 [cited
650 14 Jul 2023]. Available: [https://help.twitter.com/en/rules-and-policies/defending-and-](https://help.twitter.com/en/rules-and-policies/defending-and-respecting-our-users-voice)
651 [respecting-our-users-voice](https://help.twitter.com/en/rules-and-policies/defending-and-respecting-our-users-voice)
- 652 40. Twitter Inc. About your Home timeline on Twitter. 2022 [cited 14 Jul 2023]. Available:
653 <https://help.twitter.com/en/using-twitter/twitter-timeline>
- 654 41. Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, et al. The spreading of
655 misinformation online. *Proceedings of the National Academy of Sciences*. 2016;113: 554–
656 559. doi:10.1073/pnas.1517441113
- 657 42. Vosoughi S, Roy D, Aral S. The spread of true and false news online. *Science* (1979).
658 2018;359: 1146–1151. doi:10.1126/science.aap9559
- 659 43. Wang X, Zhang M, Fan W, Zhao K. Understanding the spread of COVID-19
660 misinformation on social media: The effects of topics and a political leader’s nudge. *J*
661 *Assoc Inf Sci Technol*. 2022;73: 726–737. doi:10.1002/asi.24576

- 662 44. Sauvayre R, Vernier J, Chauvière C. An Analysis of French-Language Tweets About
663 COVID-19 Vaccines: Supervised Learning Approach. *JMIR Med Inform.* 2022;10:
664 e37831. doi:10.2196/37831
- 665 45. Qorib M, Oladunni T, Denis M, Ososanya E, Cotae P. Covid-19 vaccine hesitancy: Text
666 mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter
667 dataset. *Expert Syst Appl.* 2023;212: 118715. doi:10.1016/j.eswa.2022.118715
- 668 46. Kaliyar RK, Goswami A, Narang P. FakeBERT: Fake news detection in social media with
669 a BERT-based deep learning approach. *Multimed Tools Appl.* 2021;80: 11765–11788.
670 doi:10.1007/s11042-020-10183-2
- 671 47. Sharma K, Qian F, Jiang H, Ruchansky N, Zhang M, Liu Y. Combating Fake News: A
672 Survey on Identification and Mitigation Techniques. 2019.
- 673 48. Ruchansky N, Seo S, Liu Y. CSI: A Hybrid Deep Model for Fake News Detection. 2017.
674 doi:10.1145/3132847.3132877
- 675 49. Kobayashi R, Takedomi Y, Nakayama Y, Suda T, Uno T, Hashimoto T, et al. Evolution
676 of Public Opinion on COVID-19 Vaccination in Japan: Large-Scale Twitter Data
677 Analysis. *J Med Internet Res.* 2022;24: e41928. doi:10.2196/41928
- 678 50. De Clerck B, Rocha LEC, Van Utterbeeck F. Maximum entropy networks for large scale
679 social network node analysis. *Appl Netw Sci.* 2022;7: 68. doi:10.1007/s41109-022-00506-
680 7
- 681 51. Valeanu A (GitHub Repository). Vaccine-tweets-ro-research. 2022 [cited 17 Jul 2023].
682 Available: <https://github.com/valeanuandrei/vaccine-tweets-ro-research>

683 52. Wilson SL, Wiysonge C. Social media and vaccine hesitancy. *BMJ Glob Health*. 2020;5:
684 e004206. doi:10.1136/bmjgh-2020-004206

685 53. Pierri F, Perry BL, DeVerna MR, Yang K-C, Flammini A, Menczer F, et al. Online
686 misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Sci Rep*.
687 2022;12: 5966. doi:10.1038/s41598-022-10070-w

688

689 **Supporting information captions**

690 **S1 Table:** Detailed example of implementation of SVM algorithm on 9 extra tweets from the
691 external dataset

692

Evaluated algorithms:

- Support Vector Classifier
- Multilayer Perceptron
- Random Forest Classifier
- Ensemble model (Support Vector Classifier + Multilayer Perceptron)
- Recurrent Convolutional Neural Networks
- BERT

Most important validation metrics:

- Matthews Correlation Coefficient
- ROC AUC Score (One-Versus-One)
- ROC AUC Score (One-Versus-Rest)

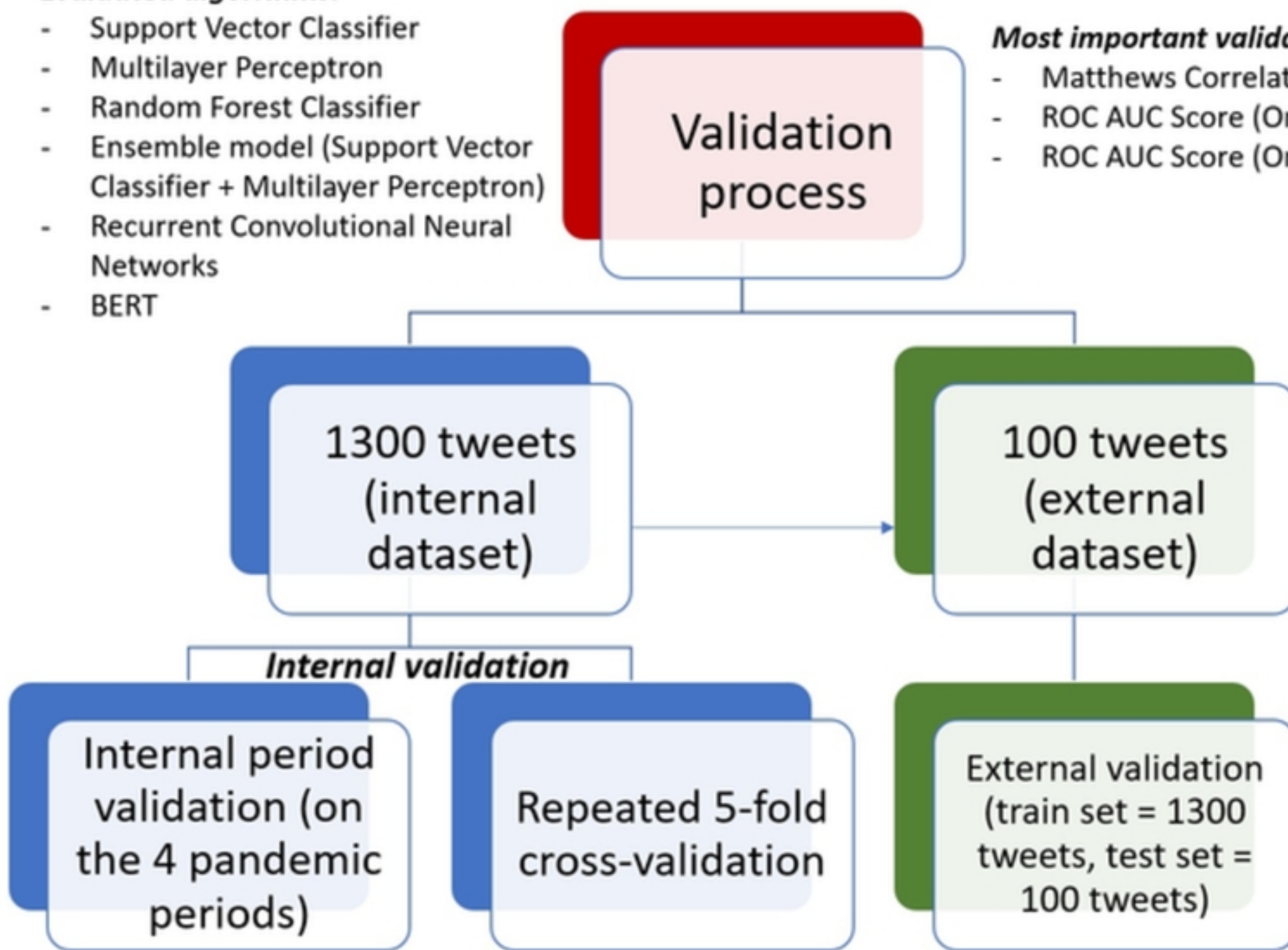


Figure 1

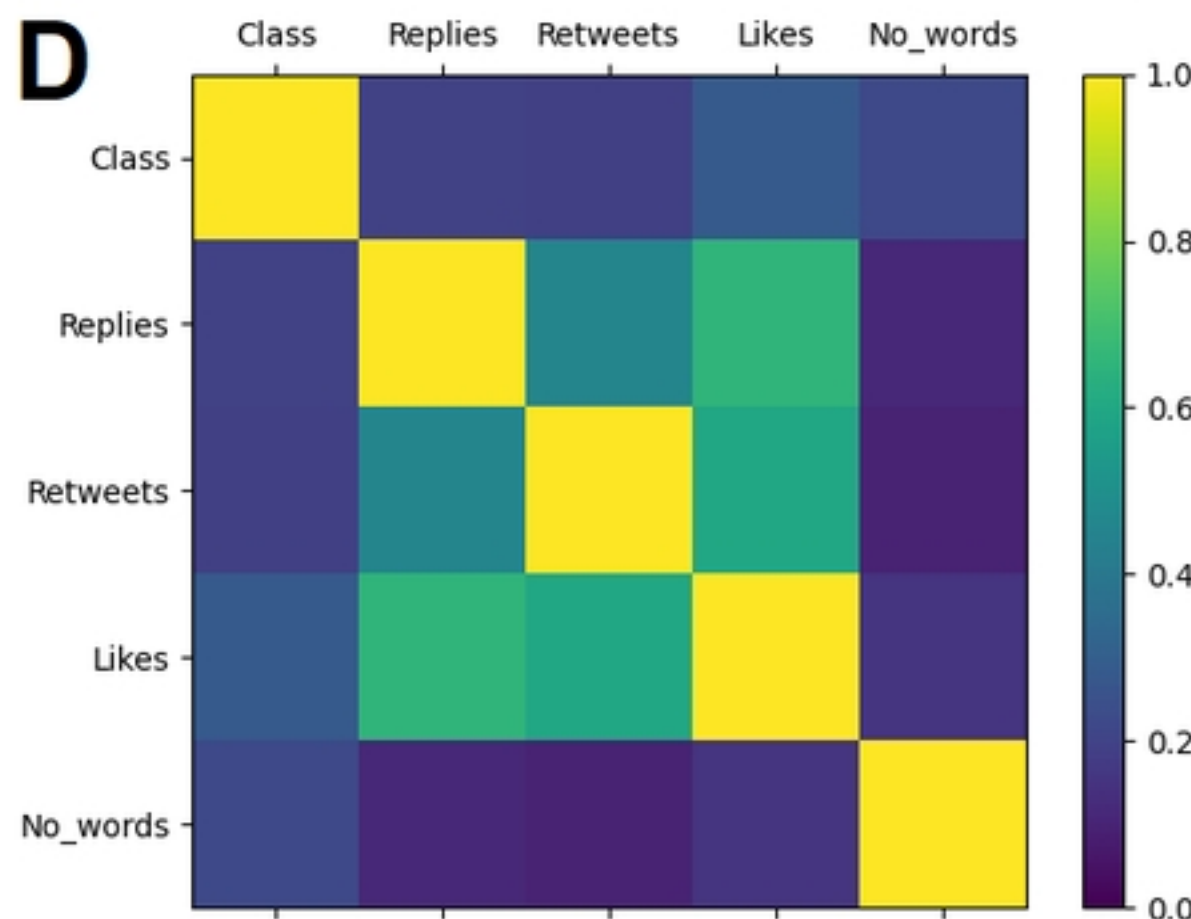
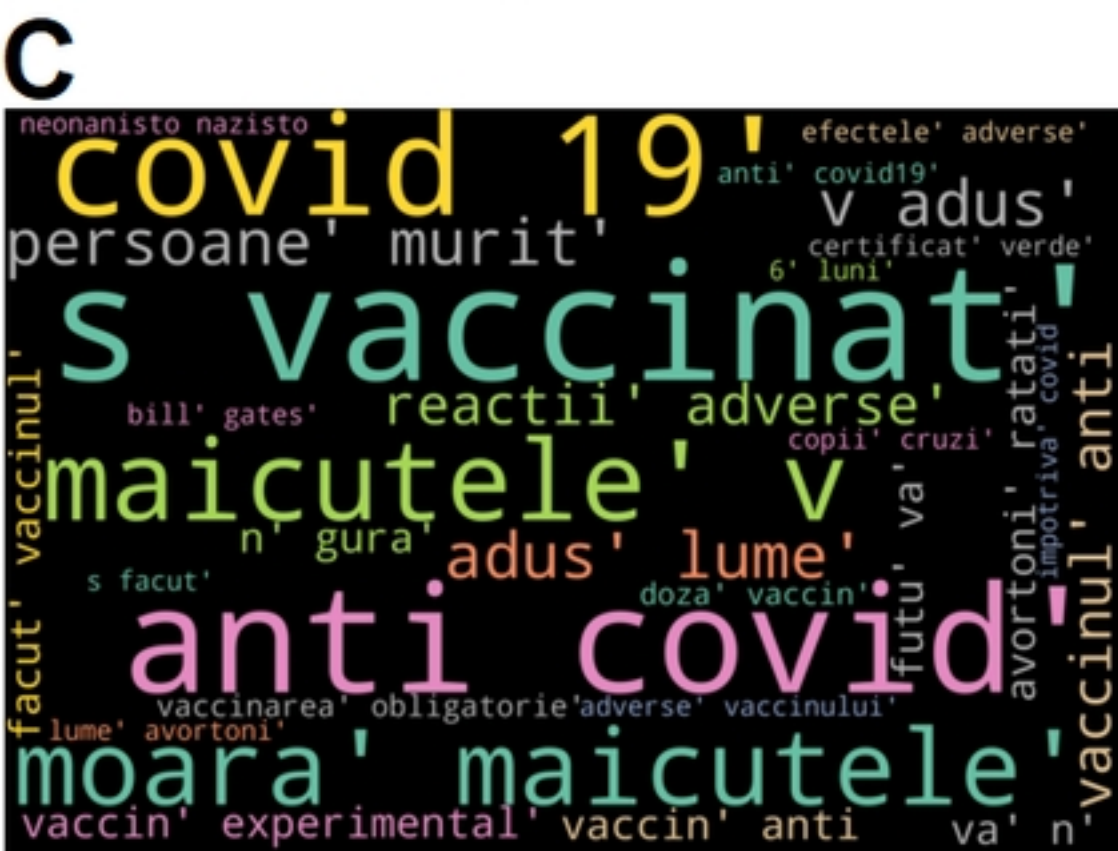
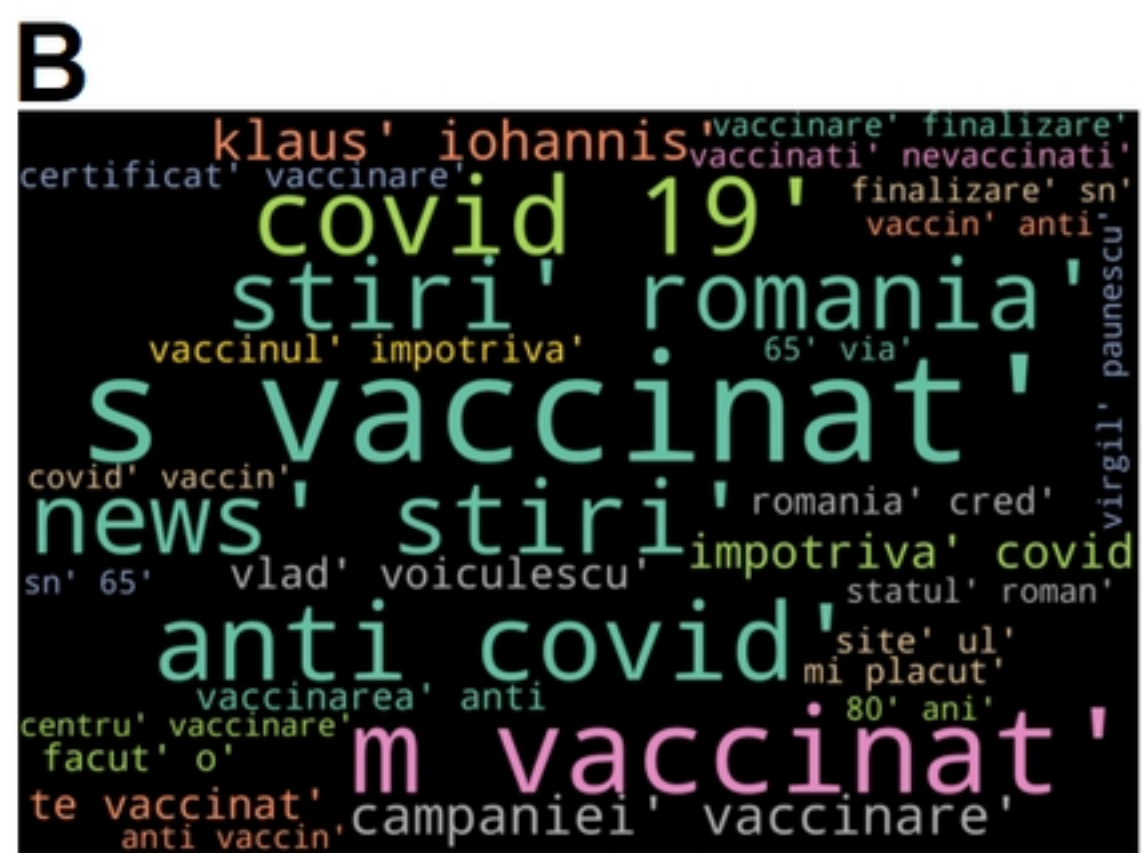
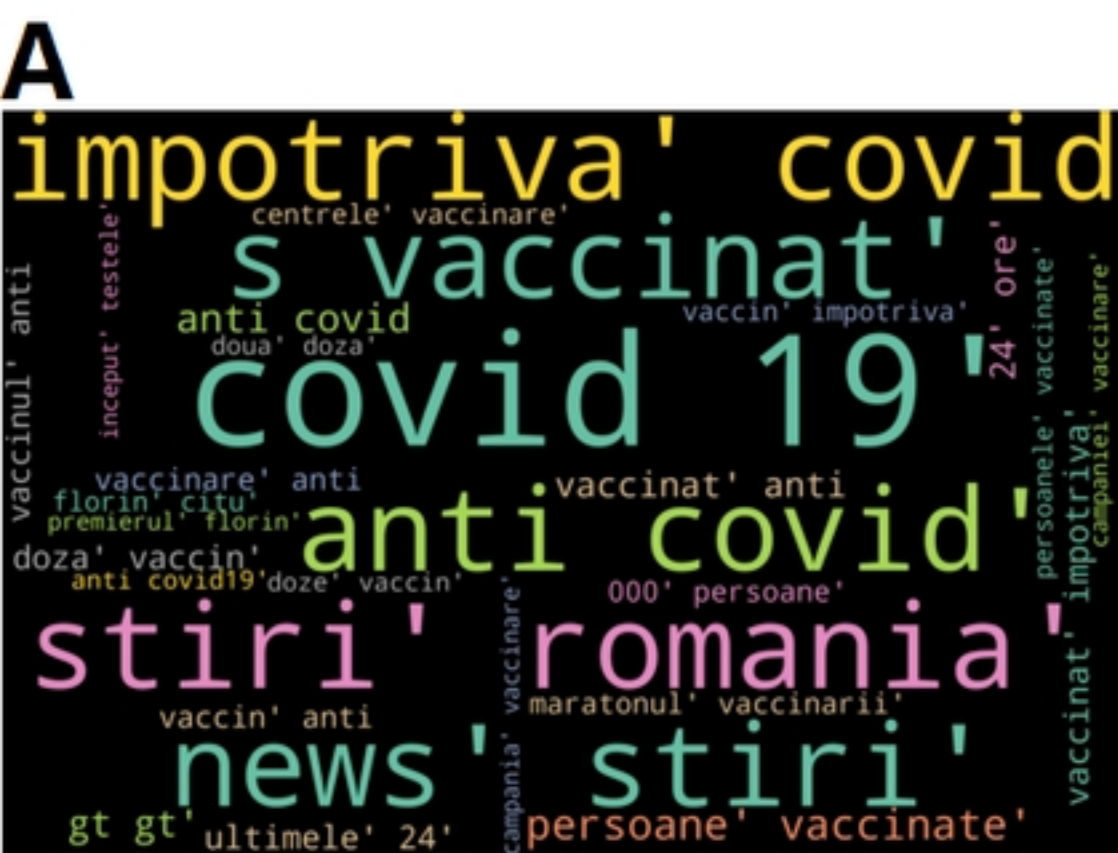


Figure 2

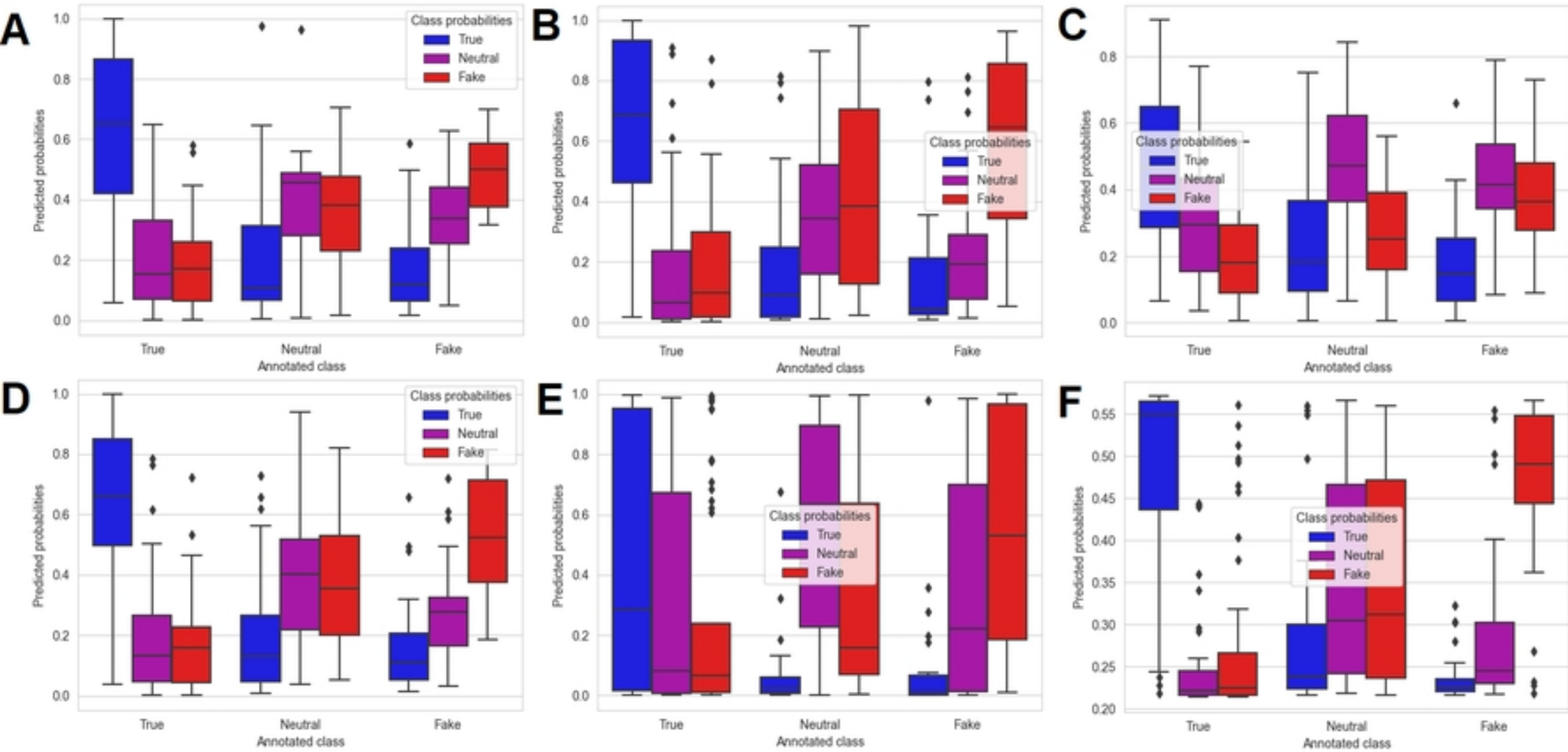


Figure 3