A Content-based Approach for the Analysis and Classification of Vaccine-related Stances on Twitter: the Italian Scenario

Marco Di Giovanni,1,2 Lorenzo Corti,1 Silvio Pavanetto,1 Francesco Pierri,1 Andrea Tocchetti,1 Marco Brambilla.1

1 Politecnico di Milano, Via Giuseppe Ponzio 34, 20133 Milano MI, Italy
2 Università di Bologna, Via Zamboni 33, 40126 Bologna BO, Italy
name.surname@polimi.it

Abstract

One year after the outbreak of the SARS-CoV-2, several vaccines have been successfully developed to prevent its spreading, and vaccine roll-out campaigns are taking place worldwide. However, an increasing number of individuals is still hesitant towards getting vaccinated, and this poses a serious threat to reaching herd immunity. We collect and analyze Italian online conversations about COVID-19 vaccines on Twitter. We define a hashtag-based semi-automatic approach to label large volumes of tweets as supporters or skeptical about the vaccine. We investigate the geographical, temporal and lexical distribution of data, and we train an accurate binary classifier that predicts the stance of tweets towards vaccines, i.e., it applies a “Pro-vax” or “No-vax” label. This classification approach can be used, in parallel with other affirmed techniques, to promptly detect and prevent the spread of negative and misleading messages about vaccines, ensuring higher rates of vaccine uptake.

Introduction and Related Work

A year after the outbreak in China, the SARS-CoV-2 has radically changed our lives, and despite the countermeasures adopted by countries across the world to prevent its spreading (Bonaccorsi et al. 2020; Spelta et al. 2020), the pandemic has infected more than 123M individuals and caused more than 2.7M deaths worldwide1. Nevertheless, we have seen the rapid development of several vaccines with over 90% effectiveness, the foremost being the one developed by Pfizer-BioNTech, announced in November 2020.2 As of March 22nd, 2021, more than 439M vaccine doses have been administered worldwide, which translates to almost 5.7 doses every 100 individuals3. Italy, in particular, has started its vaccination program on December 27th 2020, with 8M doses given to citizens4 as of March 22nd, 2021.

Although vaccination is considered one of the greatest achievements of public health, it is still perceived as unsafe and unnecessary by a growing number of individuals and the causes of this phenomenon involve emotional, cultural, social, spiritual, political and cognitive factors (Dubé et al. 2013). In particular, after the decline in measles coverage in 12 European countries in 2018, vaccine hesitancy has been included in the top-10 threats to global health in 2019 by the World Health Organization5.

Over the last decades, social media experienced a quick growth in their user-base and daily usage. Echo chamber effects, i.e. reinforcement of users’ beliefs via the interaction with a closed set of similar users, have been observed during debates about political and socially relevant topics (Colleoni, Rozza, and Arvidsson 2014; Del Vicario et al. 2016). Cossard et al. (2020) observed a similar phenomenon regarding Italian Twitter conversations about vaccines in 2019, focusing on the worrying asymmetry of the chambers’ topology.

The alarming growth of skepticism, powered by social media, caused an increase of scientific contributions inspecting the phenomena from different points of view. Pierri et al. (2021a) studied online misinformation about vaccines in US, Kang et al. (2017) constructed semantic networks of vaccine information from highly shared websites of Twitter users in the United States, D’Andrea et al. (2019) trained an SVM classifier to detect the stance of tweets about vaccines, Gargiulo et al. (2020) discovered an asymmetric behaviour of defenders and critics of vaccines in the French-speaking Twitter, Broniatowski et al. (2018) focused on the effect of bots and trolls in the debate, Guarino et al. (2021) investigated the information disorders on social media. Specific to the Italian context, many contributions have been published after the Law on Mandatory Vaccinations in 2017 (Donzelli et al. 2018; Lovari, Martino, and Righetti 2021; Righetti 2020).

In this work we inspect the SARS-CoV-2 vaccination debate on Italian-speaking Twitter from a textual content point of view. Our goal is to train an accurate stance classifier that detects patterns in tweets shared by supporters and skeptics of the vaccine. We design a semi-automated, human-

1https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6
4https://www.governo.it/it/cscovid19/report-vaccini/
5https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019
in-the-loop, hashtag-based approach to label a large set of Italian tweets. We inspect the obtained labeled dataset by focusing on the location and date of tweets, and lexical patterns, looking at possible correlations and induced biases. Finally, we successfully train a BERT (Devlin et al. 2019) model to classify the stance of tweets (“No-Vax” vs “Pro-Vax”), observing high values of AUROC and F1 score also on a dataset of manually labeled tweets that cannot be classified by the semi-automated approach previously defined. Our model can be used to monitor on real time the vaccination debate, independently on both the shared trending hashtags and the underneath social graph.

Data Collection

We collected data from Twitter, a micro-blogging platform widely used in Italy to discuss trending topics, whose official API allows for a fast implementation and a comprehensive collection. We query Twitter’s Streaming API searching for Italian tweets containing at least one of the keywords reported in Table 1. The collection is running continuously since December 20th, 2020, and by March 12th, 2021 we obtained about 30,000 tweets shared by 250k different users (Pierri et al. 2021b).

Table 1: List of keywords used to filter tweets. They refer to vaccine, vaccinate, vaccination.

<table>
<thead>
<tr>
<th>vaccine</th>
<th>vaccinare</th>
<th>vaccinerai</th>
</tr>
</thead>
<tbody>
<tr>
<td>vaccino</td>
<td>vaccinare</td>
<td>vaccineremo</td>
</tr>
<tr>
<td>vaccinazioni</td>
<td>vaccinamocici</td>
<td>vaccinerete</td>
</tr>
<tr>
<td>ionomivaccino</td>
<td>vaccinam Carnival</td>
<td>ionomivaccinero</td>
</tr>
</tbody>
</table>
| vaccino2021 | vaccinocovid                     | iononsononcavac |}

Data Labeling

The manual labeling of big datasets is an expensive and non-scalable approach. Graph-based approaches have obtained impressive results when applied to detect stance in controversial debates (Garimella et al. 2018; Cossard et al. 2020; de Zarate et al. 2020). However, these approaches are mainly used to categorize users, scoring their membership with respect to one side of the debate, but not to label single tweets.

We design a content-based, human-in-the-loop approach to semi-automatically label large sets of tweets as “Pro-Vax” or “No-Vax”. This approach is based on hashtags, often used to express the stance of users about a topic (Mohammad et al. 2016). Trending hashtags attract audience and get the attention of other users in the social network.

We define Gold hashtags those that clearly indicate either a positive or negative view in the vaccine debate. We collect two sets of Gold hashtags, one for each side of the discussion and we label tweets according to the hashtags they contain. We set the stance of a tweet based on the stance of the Gold hashtag, whereas tweets containing at least one of the 50 most-frequent ones pass the validation procedure. We iteratively add new hashtags by searching from the most frequent co-occurring ones, manually selecting the most pertinent ones and choosing them based on their meaning. An example of discarded hashtag is #vanityfair (name of a fashion magazine), highly co-occurring with #iomivaccino, since we cannot safely assume that it is used only by supporters. We also discard hashtags that equally co-occur with hashtags from both sides in similar percentages. An example is #novax, that co-occurs with both #iomivaccino and #iononmivaccino about 50 times in original tweets (tweets that are not retweets). By manually inspecting tweets which shared this hashtag, we notice that it is used by skeptical users to state their side, but also by supporters to refer to their opponents (e.g., Table 2 A-B).

Finally, three annotators manually validated the selected hashtags, as previously described for the initial Gold hashtags. A hashtag is not validated (and thus discarded) if any annotator classified more than 10% of the associated tweets as belonging to the opposite class. In this way we reliably discard hashtags that are meant to be used by a specific side of the debate, but are also often used by the other side in a criticizing or ironic manner. An example is #NoVaccineToNoVax (“No Vaccine To No-Vax”), that is used by “Pro-Vax” partisans in an attempt to prevent people, currently against vaccines, to change their minds and get vaccinated in the future. However, it is also largely used by “No-Vax” users to remark that they do not want to get vaccinated (e.g., Table 2 C-D). After three iterations we obtain a final set of three “Pro-Vax” Gold hashtags and three “No-Vax” Gold hashtags, shared in almost 500 original tweets, by manually labeling only a few hundreds tweets (statistics of the Gold hashtags are reported in Table 3). Since no other hashtag among the 50 most-frequent ones passes the validation procedure, we end the labeling process.

Gold hashtag from both sets are discarded. To obtain the final set of Gold hashtags, we start from two Gold hashtags, one for each stance: #iomivaccino (“I will get vaccinated”) and #iononmivaccino (“I will not get vaccinated”). Three annotators manually validate this selection by inspecting 50 tweets for each hashtag, finding only 2 tweets that clearly belong to the opposite stance.

Table 2: Translated examples of tweets containing both a starting Gold hashtag (#iomivaccino or #iononmivaccino) and #novax (A and B) or #novaccinoainovax (C and D). Note that the two hashtags cannot be selected as Gold hashtags since they are used with different purposes by users from both sides of the debate.

| A | in a world of conspiracies and #novax, me and my brother-in-law bet on who between the two of us would get vaccinated first. If there won’t be #vaccines this afternoon I will win the net #Iwillgetvaccinated #Vaccine |
| B | Please get vaccinated, do in the movie 2022…. the survivors You won’t be there #Iwillnotgetvaccinated #novax |
| C | #Iwillgetvaccinated even 17 times, to save the world from the terrible pandemics #COVID-19 #NoVaccineToNoVax, they don’t deserve the help of science, rather #DonateVaccineToAMigrant let’s help them and be inclusive |
| D | Still relevant, I share to wake up some sleeper #Iwillnotgetvaccinated #HamNotAGiuneaPig #NoVaccineToNoVax |

#NoVax

#IoMiVaccino

#NoVaccinoAinovax
Table 3: Statistics related to Gold hashtags: \( N \) is the total number of collected tweets, \( p_{orig} \) is the percentage of original tweets (tweets that are not retweets), \( N_u \) is the number of unique users that shared the hashtag, \( p_u \) is the percentage of unique users that shared the hashtag in an original tweet. Translations from top to bottom: IWillGetVaccinated, VaccinateH24, LetsNetwork; IWontGetVaccinated, IamNotAGuineaPig, HealthDictatorship.

Data Description

In this section we investigate the geographical, temporal, and lexical distribution of our labeled tweets, looking for relationships and correlations with the computed stance.

Geographical Analysis

What is the geographical distribution of users who tweet about pro- or anti-vax views?

Twitter offers to its users the possibility to geographically tag shared tweets, but many users do not usually enable this functionality. For example, in our dataset only 881 tweets are geolocalized (0.03% of the total data). To investigate the geographical provenance of our data, we devised an approach to obtain the Italian region in which a tweet was posted, by looking at the location of users as specified in their profiles. We use a basic string matching algorithm to match it with the names of the 20 Italian regions, the 107 provinces and the 7903 municipalities\(^7\), also including the most common English translations (e.g., Milan, Tuscany). We obtained the locations of tweets (including retweets), 19k of which also contain one gold hashtag. Figure 1 shows the geographical distribution of the ratio between the number of tweets using a “No-Vax” gold hashtags and the number of tweets using a “Pro-Vax” gold hashtags. We note that Umbria is the region with the highest “No-Vax to Pro-Vax” ratio, with only about twice as much “Pro-Vax” tweets compared to “No-Vax” ones.

Temporal Analysis

What is the temporal dynamics of the two factions?

The data analysed in this study spans the months from 20/12/2020 until 12/03/2021. Figure 2 shows the daily ratios of tweets labelled as “No-Vax” versus the ones labelled as “Pro-Vax”, using as reference the gold hashtags from Table 3. We notice a steep valley at the beginning of January, since #vaccinareh24 was trending, and a spike at the beginning of February, most likely due to a debate about vaccines between Dr. Amici and Dr. Bassetti, broadcasted live during an episode of Non è l’arena on La7 (an Italian mainstream television channel). The conflict between doctors fueled the controversy and resulted in an influx tweets with a skeptical inclination about the vaccine.

Lexicon Analysis

What is the lexicon overlap between tweets shared from the two factions?

Our goal is to train an accurate stance classifier of tweets (see next section). A big lexicon overlap between training texts belonging to opposite classes forces a classifier to learn the meaning of sentences. On the other hand, if the lexicon overlap is small, a classifier could rely on the presence of specific, often unrelated, words to make the right prediction. We quantify the lexicon overlap of the two classes by computing the Pointwise Mutual Information (PMI) between words and classes (Gururangan et al. 2018): a word has a high PMI score with respect to a class when that word occurs mainly in tweets from a single class (e.g., a word used only by “No-Vax” users). For this analysis, we discard Italian stop words and apply text tokenization using the NLTK library (Bird, Klein, and Loper 2009). We report in Table 4 the first five tokens for each class, ranked by PMI score, and the proportion of texts in each class containing each token.

\(^7\)https://it.wikipedia.org/wiki/Comuni_d\/%27Italia
found among the ones with high PMI score is “facciamoinformazione” (“we spread information”), that is found only in 2.1% of the texts labeled as “Pro-Vax”. Therefore, a classifier cannot safely rely on the presence of specific words since the most indicative ones are not very frequent.

**Stance classification**

**Data Cleaning** Before training the classifier, we cleaned the text of tweets through the following procedure. Texts are lowercased, URLs are removed and spaces are standardized. **We remove Gold hashtags** (Table 3) since they were used to automatically label tweets, thus maintaining them will introduce a strong bias in the trained models. Tweets containing at least half of the characters as hashtags are also removed, since they are too noisy. To prevent overfitting we remove duplicate texts, including retweets. We also remove texts shorter than 20 characters, that usually refer to URLs or other tweets, being difficult to understand and contextualize. The cleaning procedure reduces the number of tweets to about 10k, of which 1.8k labeled as “No-Vax”.

**Methodology** Given the large set of labeled tweets using Gold hashtags, we train six text classifiers to predict the stance of a tweet. We select the following models:

- Majority classifier (Baseline);
- Logistic regression and SVM, both fed with TF-IDF of Bag of Words vectors (Joachims 1998; Fan et al. 2008);
- FastText (Joulin et al. 2016), a fast baseline approach widely used for text classification. Its architecture is similar to the CBOW model in Word2Vec (Mikolov et al. 2013). It is known to reach performances on par with some deep learning methods, while being much faster;
- BERT (Devlin et al. 2019), a Transformer-based model (Vaswani et al. 2017) that reaches state-of-the-art performances on many heterogeneous benchmark tasks. The model is pre-trained on large corpora of unsupervised text using two self-supervised techniques: Masked Language Models (MLM) task and Next Sentence Prediction (NSP) task. Pre-trained weights are available on the Hugging face models repository (Wolf et al. 2020). We select a model pre-trained on a concatenation of Italian Wikipedia texts, OPUS corpora (Tiedemann 2012) and OSCAR corpus (Ortiz Suárez, Sagot, and Romary 2019), performed by MDZ Digital Library. We fine-tune the model on our data\(^8\). This pre-trained model has knowledge of the Italian language, lexicon and grammar, but it has few information about our topic (SARS-CoV2 vaccine). We apply Adaptive fine-tuning (AF) (Ruder 2019) to tackle this issue. The pre-trained Italian BERT is unsupervisedly fine-tuned on a MLM task using our full dataset (removing retweets to prevent overfitting). We obtain a specialized model about COVID-19 vaccine, that is fine-tuned on supervised data (like the original Italian model). We refer to this configuration as BERT+AF.

**Results** In Table 5 (left) we report Area Under ROC, weighted F1 score and F1 score on the “No-Vax” class. The values are average of 5-fold cross validation on the training set obtained with Gold hashtags. As expected, the BERT+AF model obtains the best results. To test the generalization capabilities of our classifiers, we feed them with a Test set of 1000 general tweets: tweets that does not contain any Gold hashtags. The tweets are manually labeled by three annotators in four classes: “Pro-Vax”, “No-Vax”, “Neutral” and “Out of Context”. We removed “Neutral” and “Out of Context” tweets obtaining 412 tweets, of which 132 labeled as “No-Vax”. In Table 5 (right) the metrics confirm that BERT+AF handles general tweets better than the baseline models, thus can be applied to detect and prevent the spread of negative and harmful messages.

**Conclusions**

To fight the spread of misinformation, the first step is the detection of negative and harmful messages. In this work, we collected and analyzed tweets about the Italian vaccination campaign. We designed an approach to semi-automatically label a large dataset using Gold hashtags, we analyzed it in terms of geographical, temporal, and lexical distribution, and finally, we used it to train an accurate BERT-based binary-classifier. Our approach suffers from some limitations. First, the selection and usage of Gold hashtags has a strong relationship with date they were trending. Results on the test set suggest small overfitting, but further investigations are required to confirm how relevant it is. Second, by construction of the training dataset, our classifier does not detect neutral tweets or tweets whose stance is undefined, even if widely shared. Future work will focus on the implementation of a 3-class classifier, including the Neutral class, and the real time application of the obtained model to promptly detect the daily trend of “No-Vax” tweets.

---

\(^8\)https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased

<table>
<thead>
<tr>
<th>Model</th>
<th>Validation AUROC</th>
<th>F1\text{_}\text{__}novax</th>
<th>F1\text{____}novax</th>
<th>Test AUROC</th>
<th>F1\text{____}novax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.80</td>
<td>0.74</td>
<td>0.47</td>
<td>0.7c</td>
<td>0.55</td>
</tr>
<tr>
<td>LR</td>
<td>0.83</td>
<td>0.83</td>
<td>0.39</td>
<td>0.71</td>
<td>0.67</td>
</tr>
<tr>
<td>SVM</td>
<td>0.83</td>
<td>0.84</td>
<td>0.45</td>
<td>0.73</td>
<td>0.71</td>
</tr>
<tr>
<td>FastText</td>
<td>0.75</td>
<td>0.81</td>
<td>0.32</td>
<td>0.71</td>
<td>0.60</td>
</tr>
<tr>
<td>BERT</td>
<td>0.89</td>
<td>0.87</td>
<td>0.60</td>
<td>0.76</td>
<td>0.73</td>
</tr>
<tr>
<td>BERT+AF</td>
<td>0.93</td>
<td>0.89</td>
<td>0.68</td>
<td>0.80</td>
<td>0.75</td>
</tr>
</tbody>
</table>
Acknowledgments
This work has been supported by the EU H2020 research and innovation programme, COVID-19 call, under grant agreement No. 101016233 “PERISCOPE” (https://periscopeproject.eu/).

References


