Towards a Multilingual System for Vaccine Hesitancy using a Data Mixture Approach

Oscar Araque[1], María Felipa Ledesma-Corniel[1] and Kyriaki Kalimeri[2]

[1]Universidad Politécnica de Madrid, ETSI Telecomunicación, Intelligent Systems Group, Madrid, Spain
[2]ISI Foundation, Turin, Italy

Abstract

Understanding public narratives on contentious topics like vaccination adherence is vital for promoting cooperative behaviors. During the COVID-19 pandemic, significant polarization arose from concerns about vaccines, with misinformation and conspiracy beliefs proliferating on social media. While many studies have analyzed these narratives, the focus has largely been on English-language content. This linguistic bias limits comprehensive global insights. Our study introduces a novel multilingual approach that addresses this gap. By integrating Italian examples into a primarily English dataset, we detect vaccine-hesitant language and demonstrate the model's adaptability to diverse linguistic data. Our findings highlight the importance of incorporating varied linguistic datasets for a more holistic understanding of global narratives on vaccine hesitancy.

Keywords

vaccine hesitancy, natural language processing, machine learning, transformer models

1. Introduction

Automatically understanding peoples' narratives on controversial social issues is fundamantal to efficiently address the real concerns as they occur fostering collaborative, prosocial behaviours. Vaccination adherence is an exemplar case where society witnessed a notable polarisation concerning possible adverse reactions [1, 2]. Especially during the COVID-19 pandemic and despite vaccines being the most efficient and cost-effective intervention, the spread of misinformation [3], the scepticism around the scientific development of COVID-19 vaccines and the dissemination of conspiracy beliefs [4, 5], proliferated on social media platforms.

Numerous studies analysed user generated text [6, 7, 8, 9, 10], almost exclusively focusing on the English language due to the availablity of models and tools. Even if often English is universally spoken limits the analysis in specific sociodemographic groups. Lenti et al. [11] in a purely network based approach showed the existence of a global misinformation network, calling for a multilingual analysis to further understand the drivers of vaccine hesitancy in the various languages.

Here, in light of these issues, we propose a novel approach for multilingual language understanding able to deal with language unbalance. More specifically, here we progressively include Italian instances in a predominantly English dataset for the task of vaccine hesitant language detection and demonstrate the ability of the model to generalise on previously unseen data. Often, researchers and practitioners have access to large English datasets but data in other languages, such as Italian, are lacking. We show that including small datasets in different languages can improve overall performance when analyzing texts in several languages.

2. Data and Methods

2.1. Data Collection

Although several Twitter datasets were constructed to monitor COVID-19 pandemic and are openly available to researchers, they differ in the number, timing, and language of tweets collected, as well as the search keywords used for collection [6, 12]. Here, we opted of a large multilingual dataset (MultilingTw [11]), an Italian dataset [13], while we also performed a new data collection based on a time invariant hashtag list, manually annotated as per their vaccination stance, which we share with the community.

A. Twitter-AntiVax This dataset was collected for this specific study and has been generated by capturing English Twitter messages ranging from December 2020 to March 2023. It aims to capture opinions and narratives expressed by anti-vaccination users, balancing between pro and anti stances. We collected the data using a variety of phrases and hashtags related to vaccination (e.g., "kill jab", "covid jab", "#vaccineskill", "VaccinesAreNotTheAnswer", "vaccineswork", "vaccinessavelives"), manually

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
o.araque@upm.es (O. Araque); mf.ledesma@alumnos.upm.es (M. F. Ledesma-Corniel); kyriaki.kalimeri@isi.it (K. Kalimeri)
https://gsi.upm.es/oaraquem (O. Araque)
0000-0003-3224-0001 (O. Araque); 0000-0001-8068-5916 (K. Kalimeri)

Dataset	Pro-vaccine	Anti-vaccine	Total	Total no. tokens	Avg. no. tokens
A. Twitter-AntiVax (ours)	14,374	13,560	27,934	170,299	30.5
B. TwitterVax	7,210	7,150	14,360	61,312	21.3
C. MultilingTw (EN)	416	52	468	16,543	35.3
C. MultilingTw (IT)	141	45	186	6,309	33.9

Table 1
Distribution of instances for each class, number of tokens in total, and average number of tokens per document.

inspecting the relevance of a sample of the obtained messages. From these hashtags, we have identified a set of users (480 users in total) expressing pro and anti-vaccine stances. Finally, we extracted the tweets generated by these users in the considered period. The dataset with the respective annotations is freely accesible at https://github.com/gsi-upm/multilingual-vaccine-hesitancy.

B. TwitterVax dataset [13]. Originally, this dataset contains 9,068,389 Italian tweets on vaccines tweeted from 1st January 2019 to 1st June 2022. The authors annotate each captured user as anti-vaccine and pro-vaccine through network analysis. For this work, to reduce the computational load and to work with similar dataset sizes, we have selected a sub-sample of approximately 14,000 tweets. We have categorized each tweet as anti-vaccine and pro-vaccine by means of the user's annotation.

C. Multilingual Twitter dataset (MultilingTw) [11]. This dataset is composed of Twitter messages in 18 languages from October 2019 to March 2021. While the original size is around 316 million messages, we select a subsample of 1,246 tweets in English and 449 in Italian, manually labelled for their vaccine stance. To comply with the other datasets, we selected the messages labelled as pro and anti-vaccine. These sets are used as test sets.

Both the Twitter-AntiVax and TwitterVax dataset have been split into train and test sets, randomly sampling 20% of instances as test set. Some statistics for the used datasets are detailed in Table 1.

2.2. Methods.

This work is based on a multilingual approach to vaccine hesitancy analysis. To this regard, we use a DistilBERT [14] model (`distil-base-multilingual-cased`[1]). This transformer model was trained in the most common languages in Wikipedia and thus is capable of generating internal representations for a variety of languages, including English and Italian. Nevertheless, it has been shown that this kind of models do not compute language-agnostic representations but rather generates partitioned representations for each language [15]. In practice, this implies that the instances used for training in English are not directly useful for predicting vaccine hesitancy in other languages since the internal representations vary with the language.

Here, our goal is to model the effect of including small sets of data in a multilingual approach. To do so, we use the Twitter-AntiVax train set as English training data and the TwitterVax train set and Italian training data. As test sets, we use the test sets of Twitter-AntiVax and TwitterVax, as well as the Multilingual Twitter dataset in both English and Italian. To modulate the number of Italian instances included in the training set, we define the α parameter that can take values in the range $[0, 1]$. Thus, the instances in the training set are composed with the following expression:

$$\text{Train instances} = \alpha * \text{IT} + (1 - \alpha) * \text{EN}$$

where IT and EN represent the Italian and English datasets, respectively. In this way, a training set composed with $\alpha = 0$ is composed entirely of English instances, while the opposite is correct when $\alpha = 1$. Of course, with $\alpha = 0.5$, the training set would have the same number of instances for English and Italian.

Since the English and Italian datasets contain a different number of instances, this could have the undesired effect of a varying number of training instances that may affect the results. We control this to produce the same number of train instances for all possible values of α.

Evaluation. Finally, all the models are evaluated with the macro-averaged F-score of each model. This allows us to consider the effect of unbalanced data. We opted for an evaluation without label propagation via retweet networks as proposed in other studies [16, 11] since these are likely to introduce uncertainty in the groundtruth. Our evaluation is strictly based on manually annotated data regarding the vaccination stance.

3. Results

As described, the proposed experiment aims to study the effect of including Italian instances in an English dataset and train a multilingual learning model with the generated dataset. Figure 3 shows the macro-averaged f-scores obtained for an increasing number of α. The horizontal axis shows the variation of the α parameter (see Sect. 2), and the vertical axis the performance of

[1] https://huggingface.co/distilbert-base-multilingual-cased

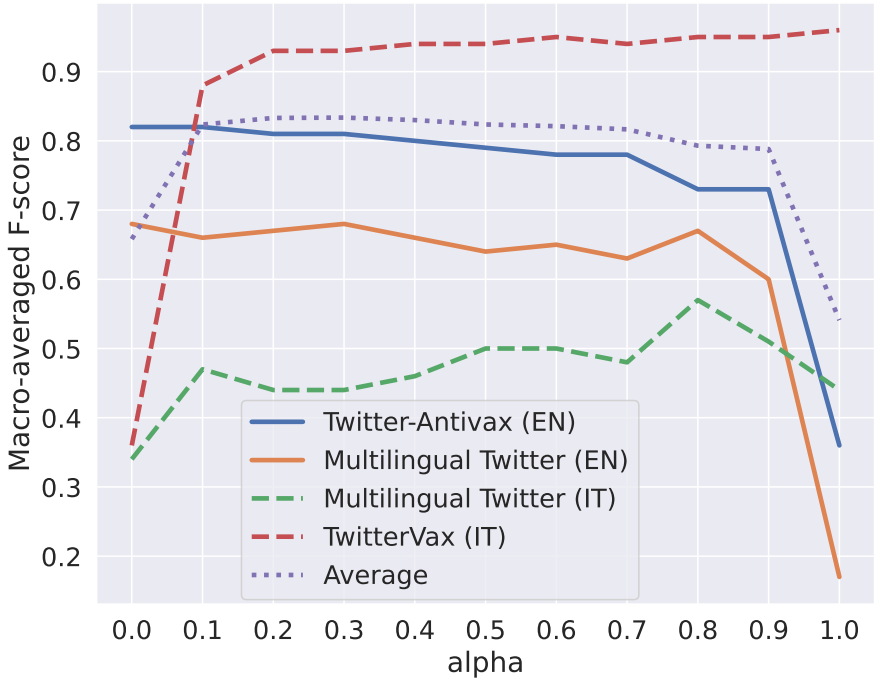

Figure 1: Macro-averaged f-scores in all test sets, both English (EN) and Italian (IT). The averaged curve is weighted with the number of instances of each test dataset.

the model in all test sets, both English and Italian. The average curve is weighted with the number of instances of each test set so that the number of correctly classified instances is better reflected.

It is worth noticing a prevalent behaviour in the obtained learning curves: the elbow evolutions with α in three of the four curves. Attending to the evolution of the performance in the Twitter-Antivax test set, we see that the best performance is obtained when $\alpha = 0$, that is, when all training instances are in English. As α increases, the performance decreases slowly. Nevertheless, when α changes from 0.9 to 1, a faster reduction is observed. The lower performance corresponds to the case where there are no English instances in the training set, negatively affecting the performance in the English language. A similar behaviour is shown by the performance on the Multilingual Twitter data in English.

In contrast, the performance in the Italian data progresses differently. We can see a large improvement in performance at the change between $\alpha = 0$ (no Italian training data) and $\alpha = 0.1$ (10% of training instances in Italian) for the TwitterVax dataset. As more Italian data is included in the composition of the training set, the performance on this dataset increases slowly. As for the Multilingual Twitter dataset in Italian, the performance tends to increase with α.

Attending to the obtained results, we can derive a general trend: the higher the percentage of training instances in a language, the higher the performance in that language. This is to be expected, as follows common experimental observations when training learning models. Besides, it is interesting to see that there is a large portion of cases where performances in both English and Italian are kept high. Practically, this situation is observed when $\alpha \in [0.1, 0.9]$ and can be better understood by attending to the averaged curve.

This interesting behavior may indicate the robustness of the proposed method to the proportion of language mixture. That is, it seems that the model successfully generalizes to a different language even when its training set is composed in a small proportion (e.g., 10%) by instances of that language. The previous observation indicates that the multilingual model may be learning to classify Italian documents while being trained with English instances, and that adding a small proportion of Italian instances facilitates such performance.

While this work is an initial attempt at describing a multilingual system trained with a mixture of data, further work should explore whether the observed behavior is maintained with more languages. How the internal representations of the evaluated model can be used for multilingual applications has yet to be thoroughly stud-

ied.

4. Conclusions

Here we design and evaluate a method that achieves multilingual vaccine hesitancy detection. The experimental design considers training a multilingual classification model on a mixture of English and Italian text excerpts. Progressively varying the combination of languages in the training data, we obtain a better understanding of the of the classification problem in the two languages. Additionally, we undertook a novel data collection effort on Twitter, manually annotating content based on vaccination stance. This curated dataset is now freely accessible to the scientific community, providing a valuable resource for further research.

By adjusting the language composition in our training data, we gained deeper insights into the classification intricacies across both languages. Notably, our findings suggest that the model can effectively generalize to a different language even when its training set contains a minimal proportion (e.g., 10%) of instances from that language. This indicates the model's robustness and adaptability in handling linguistic variations with limited data.

Importantly, this approach is an important tool for researchers and practitioners who often have access to large datasets in English, but limited resources in other widely spoken languages such as Italian or Spanish. The evaluation shows that composing a mixture dataset can be effective in generating a model that classifies instances in two languages. In fact, the experimentation shows that this mixture is flexible, maintaining consistent performances across different ratios of language presence. This consistency suggests that the mixture approach is promising.

Given its language-neutral nature, our technique holds promise for broader applications across multiple languages and diverse domains. As a next step, we aim to explore various multilingual models and languages to further ascertain the scalability and adaptability of our approach.

Acknowledgments

This work has been funded by the Spanish Ministry of Science and Innovation through the COGNOS project (PID2019-105484RB-I00) and by the European Union with NextGeneration EU funds. KK gratefully acknowledges the support from the Lagrange Project of the Institute for Scientific Interchange Foundation (ISI Foundation) funded by Fondazione Cassa di Risparmio di Torino (Fondazione CRT).

The authors would like to acknowledge the support of Yelena Mejova, from ISI Foundation in Italy, for sharing the Multilingual Twitter dataset.

References

[1] A. A. Dror, N. Eisenbach, S. Taiber, N. G. Morozov, M. Mizrachi, A. Zigron, S. Srouji, E. Sela, Vaccine hesitancy: the next challenge in the fight against covid-19, European journal of epidemiology 35 (2020) 775–779.

[2] L. Betti, G. De Francisci Morales, L. Gauvin, K. Kalimeri, Y. Mejova, D. Paolotti, M. Starnini, Detecting adherence to the recommended childhood vaccination schedule from user-generated content in a us parenting forum, PLoS computational biology 17 (2021) e1008919.

[3] Y. Mejova, K. Kalimeri, Covid-19 on facebook ads: competing agendas around a public health crisis, in: Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies, 2020, pp. 22–31.

[4] K. Kalimeri, M. G. Beiró, A. Urbinati, A. Bonanomi, A. Rosina, C. Cattuto, Human values and attitudes towards vaccination in social media, in: Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 248–254.

[5] M. G. Beiró, J. D'Ignazi, V. Perez Bustos, M. F. Prado, K. Kalimeri, Moral narratives around the vaccination debate on facebook, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 4134–4141.

[6] K. Hayawi, S. Shahriar, M. A. Serhani, I. Taleb, S. S. Mathew, Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection, Public health 203 (2022) 23–30.

[7] S. Nyawa, D. Tchuente, S. Fosso-Wamba, Covid-19 vaccine hesitancy: a social media analysis using deep learning, Annals of Operations Research (2022) 1–39.

[8] A. Fasce, P. Schmid, D. L. Holford, L. Bates, I. Gurevych, S. Lewandowsky, A taxonomy of anti-vaccination arguments from a systematic literature review and text modelling, Nature Human Behaviour (2023) 1–19.

[9] Y. Mejova, G. Crupi, J. Lenti, M. Tizzani, K. Kalimeri, D. Paolotti, A. Panisson, Echo chambers of vaccination hesitancy discussion on social media during covid-19 pandemic, in: XX ISA World Congress of Sociology (June 25-July 1, 2023), ISA, 2023.

[10] N. E. MacDonald, Vaccine hesitancy: Definition, scope and determinants, Vaccine 33 (2015) 4161–4164. URL: https://www.sciencedirect.com/science/article/pii/S0264410X15005009. doi:https://doi.org/10.1016/j.vaccine.2015.04.036, wHO Recommendations Regarding Vaccine Hesitancy.

[11] J. Lenti, Y. Mejova, K. Kalimeri, A. Panisson, D. Paolotti, M. Tizzani, M. Starnini, Global misinformation spillovers in the vaccination debate before and during the covid-19 pandemic: Multilingual twitter study, JMIR Infodemiology 3 (2023) e44714. doi:10.2196/44714.

[12] C. E. Lopez, C. Gallemore, An augmented multilingual twitter dataset for studying the covid-19 infodemic, Social Network Analysis and Mining 11 (2021) 102.

[13] V. Lachi, G. M. Dimitri, A. Di Stefano, P. Liò, M. Bianchini, C. Mocenni, Impact of the covid 19 outbreaks on the italian twitter vaccination debat: a network based analysis, arXiv preprint arXiv:2306.02838 (2023).

[14] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[15] J. Singh, B. McCann, R. Socher, C. Xiong, BERT is not an interlingua and the bias of tokenization, in: Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 47–55. URL: https://aclanthology.org/D19-6106. doi:10.18653/v1/D19-6106.

[16] F. Gargiulo, F. Cafiero, P. Guille-Escuret, V. Seror, J. K. Ward, Asymmetric participation of defenders and critics of vaccines to debates on french-speaking twitter, Scientific reports 10 (2020) 6599.