Transfer Learning for Stance Analysis in COVID-19 Tweets

Erik Tjong Kim Sang[*] E.TJONGKIMSANG@ESCIENCECENTER.NL
Marijn Schraagen[**] M.P.SCHRAAGEN@UU.NL
Shihan Wang[**] S.WANG2@UU.NL
Mehdi Dastani[**] M.M.DASTANI@UU.NL

[*] *Netherlands eScience Center, Amsterdam, The Netherlands*

[**] *Utrecht University, Utrecht, The Netherlands*

Abstract

Governments invoked a large collection of measures to fight the effects of the COVID-19 pandemic. One way of measuring the public response to these measures is to study social media messages related to the pandemic, for different policies/topics such as face mask use, social distancing, or vaccine willingness. However, computational analysis requires large efforts of data annotation for each new topic. In this abstract we explore the technique of *transfer learning*, predicting the response to one pandemic policy measure with a machine learning model trained on annotated data related to another measure.

1. Introduction

The negative effects of the COVID-19 pandemic brought governments to impose a wide range of measures to protect their citizens. Measures such as social distancing, mandatory face mask wearing and lockdowns had a strong impact on the lives of people worldwide. Organizations such as the Corona Behaviour Unit of the Dutch National Institute for Public Health and the Environment (RIVM) were set up to monitor the public's response to the different measures[1]. These assessments involve surveys in which thousands of people answer a long list of questions related to the pandemic approximately once per month. These data require a lot of effort to collect.

As an alternative, we are interested in extracting the public's response to anti-pandemic measures from an analysis of social media, such as Twitter. Techniques for sentiment analysis and machine learning can be used to mine the opinions of social media users about pandemic topics. However, in prior work we found that machine learning analysis of each measure required manual annotation of measure-specific data (Wang et al. 2020b). This time investment prevents us from analysing all the interesting pandemic topics (Tjong Kim Sang et al. 2021).

In this abstract we will explore the technique of *transfer learning* (Weiss et al. 2016), to use labelled data related to one pandemic measure for predicting stances on several other measures. The transfer between topics may require adaptation of the data, which is investigated in the current study.

2. Data and Methods

We analyze Dutch tweets as collected by the website twiqs.nl (Tjong Kim Sang and van den Bosch 2013). We focus on the period from March 12th, 2020 (the date of the first Dutch national pandemic press conference) to March 31st, 2021. This data set contains about 20 million tweets per month of which about 13% are pandemic tweets, defined by a hand-crafted query containing 67 tokens like *corona, vaccine, pandemic, curfew* and *patient* (Tjong Kim Sang et al. 2021). In this study we focus on four pandemic topics: wearing *face masks*, *social distancing*, *testing* and *vaccination*.

1. `https://www.rivm.nl/en/novel-coronavirus-covid-19/research/behaviour`

Topic	Total	From	Until	Supports	Rejects	Irrelevant
face masks	1,011 tweets	2020–03–01	2021–03–31	21.2%	23.6%	55.2%
social distancing	5,977 tweets	2020–02–01	2020–07–03	56.2%	20.0%	23.8%
testing	1,181 tweets	2020–03–01	2020–12–31	28.8%	17.8%	53.4%
vaccination	1,007 tweets	2020–01–01	2021–01–31	9.6%	24.3%	66.1%

Table 1: Overview of the annotated data for four pandemic topics: wearing *face masks*, *social distancing*, *testing* and *vaccination*

Small random selections of topic tweets were annotated by a single annotator at different time points. Three different labels could be assigned to a tweet: Support, if the tweet author supports the related pandemic measure, Reject, if the tweet author rejects the measure, and Irrelevant, if a tweet is not relevant for the measure or if the opinion of the sender could not be determined. A subset of the social distancing tweets (898) was annotated by a second annotator. The inter-annotator agreement was $\kappa = 0.55$ (Cohen 1960) which indicates that labeling was a hard task.

Next, we used the word embedding-based machine learner fastText (Joulin et al. 2017) for building models for labelling other relevant tweets. The skipgram models were built starting from word vectors derived from all (230 million) Dutch tweets of 2020 (Bojanowski et al. 2017). The word vectors have a length of 300 and model training employed 200 epochs and a learning rate of 0.2.

We have only a limited amount of annotated data available on a small number of topics. More data is needed for building better machine learning models and analyzing more topics, but resources for data annotation are limited. Therefore we explore the technique of transfer learning (Weiss et al. 2016), applying machine learning models trained on a source data set to a target data set. This technique is also know as domain adaptation (Plank 2011) and is related to multi-task learning (Zhang and Yang 2021). The source data set in the experiments consists of the *social distancing* tweets, the other three topics are used as target. We explore several baseline techniques for supervised learning listed in (Daumé III 2007): training only on the in-domain target data (TgtOnly) training only on the out-of-domain source data (SrcOnly), and training on both (All). Additionally, we evaluate two variants of All, one where only the relevant part of the out-of-domain data (i.e., excluding the label Irrelevant) is used during training (RlvOnly) and a feature augmentation technique (FeatAug) following the method described in (Daumé III 2007). In this method each tweet is offered to the model twice during training and testing, with features labeled as source, target, or general, for example $\{I_t\ always_t\ wear_t\ face\_masks_t\ I_g\ always_g\ wear_g\ face\_masks_g\}$ for a target tweet and $\{I_s\ always_s\ keep_s\ distance_s\ I_g\ always_g\ keep_g\ distance_g\}$ for a source tweet.

Three different evaluation measures will be used. The first is label accuracy, which in the full-data-set experiments is always equal to precision, to recall and to F_1. Since we are interested in obtaining accurate moving averages over time, we also incorporate two graph-based evaluation measures: Pearson correlation coefficient (r) and absolute difference. Both measures compare a predicted time graph to a graph representing the available gold standard data.

3. Results

The trends in support for the measures related to wearing *face masks*, *social distancing*, *testing* and *vaccination* are plotted in Figure 1. The Irrelevant class has been omitted in the plots for clarity of presentation. The patterns for *social distancing* and *testing* show large support at the start of the pandemic and decreasing support towards Summer 2020, which corresponds to Dutch population survey studies[2].

2. `https://www.rivm.nl/gedragsonderzoek/maatregelen-welbevinden/draagvlak`

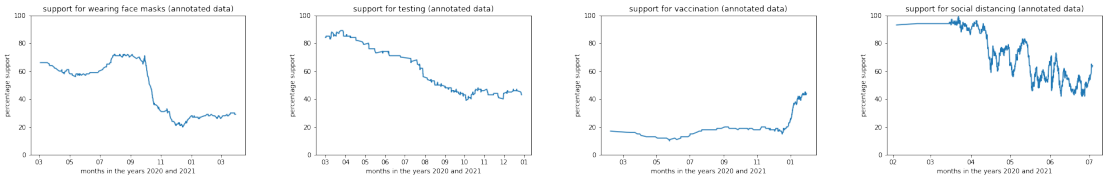

Figure 1: Graphs representing the support on Dutch-speaking Twitter (using a 100 tweet moving window) for the pandemic measures related to wearing face masks, testing, vaccination and social distancing over time, as derived from the manually annotated data.

Accuracy	TgtOnly	SrcOnly	All	RlvOnly	FeatAug
face masks	**0.592**	0.414	0.543	0.569	0.583
testing	0.561	0.452	0.541	**0.572**	**0.572**
vaccination	0.608	0.488	0.572	0.585	**0.616**
social distancing	0.656				

Table 2: Accuracy of transfer learning experiments.

For the topic *face masks* we see a support drop from 60% to 20%, starting around 30 September 2020 when the Dutch government advised face mask wearing in shops. However, support for this measure has been consistently around 80% according to the Dutch survey studies. Support for *vaccination* on Dutch Twitter has been low (20%) throughout 2020 and only started to increase slightly in January 2021. But recent Dutch survey analysis[3] showed that the support for COVID-19 vaccination is between 60% and 85% while previous research found 79% of relevant Dutch tweets of 2012-2017 supported general vaccination (Kunneman et al. 2020). So the support measured on Dutch Twitters does not match the support of the Dutch population. We believe this is caused by the large number of troll tweets in the two topics. These tweets often strongly express opinions against the measures and their large numbers in the data cause the measured support to be low.

The annotated data only provide insights in the stance with respect to the four topics for limited time frames. We need machine learning models to classify the tweets outside of those time frames. The performance of these models is evaluated by processing the annotated data sets with ten-fold cross validation: building ten models on 90% of the data and evaluating them on the remaining 10%. The results of these evaluations can be found in Table 2, in the column TgtOnly

Because of the differences in size of the data sets, the *social distancing* model performs better than the other three models. The transfer learning experiments aim to employ this advantage for improving the performances for the other three data sets. A comparison of the performances shown in the columns TgtOnly, SrcOnly and All reveals that model developed for *social distancing* does not perform very well on other topics. For the three topics, the models learned from only topic data outperforms the out-of-domain models. This result can be explained in part by the setup with three classes Support, Reject and Irrelevant, for which the class Irrelevant is not transferable between topics by definition. Therefore we also performed two groups of stacked experiments where the relevant tweets were first selected with the in-domain model (TgtOnly), after which stance labelling was performed by an out-of-domain model without relevance processing (columns RlvOnly and FeatAug). In the second group, extra features spaces for the two domains were added to enable the learner to distinguish between the two tasks.

Restricting the out-of-domain data to classifying relevant data improved the results. For two topics, *testing* and *vaccination*, improvement went beyond the marks set by TgtOnly, with RlvOnly

3. `https://www.rivm.nl/gedragsonderzoek/maatregelen-welbevinden/vaccinatiebereidheid`

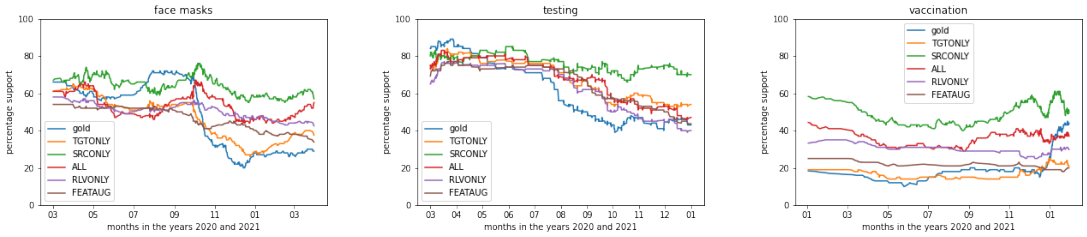

Figure 2: Time graphs applied for assessing the transfer learning results.

(for *testing*) and with FeatAug (for *testing* and *vaccination*). However, the improvements were small, at most 0.011 (1.1%) and they did not lead to better time graphs: for all three cases the Pearson r of the TgtOnly runs remained better (see Figure 2). For *testing*, the absolute difference of the graph of RevOnly was a bit better (smaller) than that of TgtOnly. Therefore we conclude that the applied transfer learning methods did not lead to useful improvements for our data.

4. Discussion

The goal of this study was to use extra data from other topics to improve single topic social media text classification (column TgtOnly in Table 2). In hindsight, it was not surprising that SrcOnly and All did not perform well since these involved determining relevance. This task is quite topic-specific and it is unlikely that the task is easily transferable from one topic to another.

The two approaches using relevant messages only did much better. As could have been expected, FeatAug, which includes three different feature spaces (source, target and combination) did well. Still the improvements over TgtOnly were small. We believe that an important reason for this is the small performance difference between the source data set (*social distancing*, about 0.65) and the target data sets (about 0.60). A better performance on the source date set will most likely enable more gains from transfer learning but as shown by (Wang et al. 2020a) this will require a much larger data set to be manually annotated. Another issue might be the size difference between the source set (6k tweets) and the target sets (1k tweets), causing the source set to dominate the classifier. A two-step approach with fine-tuning a pre-trained model may be able to address this issue.

The effect of troll activity on the measured support for *face masks* and *vaccination* (see Figure 1) limits the applicability of Twitter data for this type of topics. It is not an easy task to identify and remove these tweets from the data, one must be careful not remove genuine opinions. A second-order effect could play a role here as well: when a discussion about a topic is poisoned by a small group of users with strong opinions, users with other opinions might avoid it.

The *vaccination* graph in Figure 2) shows that none of the methods was able to predict the January 2021 surge of support. Most likely, the vocabulary used in the topic tweets changed in that month. During the year 2020, COVID-19 vaccination was a theoretical topic, while in January 2021 it became a realistic topic, influencing people's daily lives. Machine learning approaches will likely have difficulty identifying such shifts without enough examples. Manual annotation of more data will continuously be required when monitoring large ongoing events like this pandemic.

5. Concluding remarks

Four different methods for transfer learning have been applied to Dutch pandemic tweets, targeting on three pandemic topics, wearing *face masks*, *testing* and *vaccination*. The goal was to improve analysis of pandemic measure stances over time. Of the four applied methods feature augmentation (Daumé III 2007) performed best. However, the measured gains of incorporating out-of-domain data were small. Larger annotated data sets are necessary to increase the benefits of this approach.

References

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017), Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics* **5**, pp. 135–146.

Cohen, Jacob (1960), A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20** (1), pp. 37–46.

Daumé III, Hal (2007), Frustratingly easy domain adaptation, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Association for Computational Linguistics, Prague, Czech Republic, pp. 256–263. https://www.aclweb.org/anthology/P07-1033.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov (2017), Bag of tricks for efficient text classification, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Association for Computational Linguistics, pp. 427–431.

Kunneman, Florian, Mattijs Lambooij, Albert Wong, Antal Van Den Bosch, and Liesbeth Mollema (2020), Monitoring stance towards vaccination in Twitter messages, *BMC Medical Informatics and Decision Making* **20** (1), pp. 1–14, BioMed Central.

Plank, Barbara (2011), *Domain adaptation for parsing*, PhD thesis, University of Groningen.

Tjong Kim Sang, Erik and Antal van den Bosch (2013), Dealing with big data: The case of Twitter, *Computational Linguistics in the Netherlands Journal* **3**, pp. 121–134.

Tjong Kim Sang, Erik, Marijn Schraagen, Mehdi Dastani, and Shihan Wang (2021), Discovering pandemic topics on Twitter, *DHBenelux*.

Wang, Shihan, Marijn Schraagen, Erik Tjong Kim Sang, and Mehdi Dastani (2020a), Dutch general public reaction on governmental COVID-19 measures and announcements in Twitter data. ArXiv preprint, 2006.07283.

Wang, Shihan, Marijn Schraagen, Erik Tjong Kim Sang, and Mehdi Dastani (2020b), Public sentiment on governmental COVID-19 measures in Dutch social media, *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Association for Computational Linguistics, Online. https://www.aclweb.org/anthology/2020.nlpcovid19-2.17.

Weiss, Karl., Taghi M. Khoshgoftaar, and DingDing Wang (2016), A survey of transfer learning, *Journal of Big Data*, Springer.

Zhang, Y. and Q. Yang (2021), A survey on multi-task learning, *IEEE Transactions on Knowledge and Data Engineering*, IEEE. https://arxiv.org/pdf/1707.08114.

The texts in these appendices are not part of the submitted abstract

Appendix A. Topic queries

Topic	Definition
face mask wearing	`mondkapje`
social distancing	`1[.,]5[ -]*m|afstand.*hou|hou.*afstand|anderhalve[ -]*meter`
testing	`\btest|getest|sneltest|pcr`
vaccination	`vaccin|ingeënt|ingeent|inent|prik|spuit|bijwerking|-->|💉|pfizer|moderna|astrazeneca|astra|zeneca|novavax|biontech`

Table 3: Regular expressions used for searching for relevant tweets in the four pandemic topics studied in this paper, where `|` stands for the logical OR operation, `\b` represents a word boundary, `.` (period) is an arbitrary character and `*` indicates an arbitrary number of repetitions of the previous character.

Appendix B. Graph-based evaluation

Pearson r	TgtOnly	SrcOnly	All	RlvOnly	FeatAug
face masks	**0.877**	0.533	0.435	0.743	0.806
testing	**0.949**	0.819	0.882	0.856	0.848
vaccination	0.361	0.586	**0.592**	0.528	-0.824
social distancing					

Table 4: Results of fastText experiments. The TgtOnly column contains the results of in-domain runs while all other columns display out-of-domain run results where the social distancing data was used for training and other data sets for testing.

Absolute difference	TgtOnly	SrcOnly	All	RlvOnly	FeatAug
face masks	**0.080**	0.177	0.143	0.133	0.125
testing	0.093	0.169	0.097	**0.087**	0.102
vaccination	**0.078**	0.260	0.149	0.107	0.102
social distancing					

Table 5: Results of fastText experiments. The TgtOnly column contains the results of in-domain runs while all other columns display out-of-domain run results where the social distancing data was used for training and other data sets for testing.