OPINION MINING TWITTER CLIMATE CHANGE DISCUSSIONS WITH PRETRAINED BERT-MODELS

Class	`Statistical Natural Language Processing (ELEC-E5550)`
Department	Department of Signal Processing and Acoustics
Date	May 5, 2021
Students	Samuel Piirainen Ville Saarinen Lena Hegemann

Contents

1. Abstract

Social media posts are widely used to express opinions, making them an interesting source for understanding trends in the public opinion on controversial political topics. Due to the large scale of available data on social media platforms automatic methods such as opinion mining are necessary to generate comprehensive insight from social media data. We employ opinion mining to analyze the polarity of tweets in a data set collected from English-speaking Twitter on the topic of climate change. We try different BERT models, fine-tune them so that they are able to infer the sentiments of the tweets in the climate discussion, and compare their performance. We experiment on multiple different optimization strategies such as batch size, dropout, and learning rate. Since twitter data lacks explicit labels for sentiment and opinion, we use the tweet metadata to construct a network representation of social connections between Twitter users and cluster the network based on social proximity. We achieve this through unsupervised machine learning methods, namely community detection which lets us obtain labels for the different opinion groups. The best accuracy (0.922) and highest training efficiency could be reached with the CT-BERT-V2 model, which has already been pre-trained on Twitter data.

2. Introduction

Despite the mounting evidence for anthropogenic origins of climate change and the overarching consensus of the climate scientists, climate change has remained a divisive topic in the public debates during the last decade, with recent studies suggesting patterns of polarization in the public opinion to pro- and anti-mitigation attitudes [2, 3, 11]. This phenomenon has been shown to co-occur with insular echo chambers in which participants increasingly interact only with like-minded people [2]. According to communication and democracy theorists, this poses a danger in two main ways: 1) it threatens to weaken the communication efforts of climate scientists and policy experts [4], 2) it threatens to impair democracy by hindering the public's exposure to diverse opinions and increasing the likelihood of conflict [8].

New technological means by which to understand and alleviate the problems of polarization and echo chamber effects are urgently needed. Complementing the traditional toolbox available to social scientists, late advancements in areas of signal processing, complex networks, and machine learning form a diverse selection of new tools by which to approach and analyze communication data. In this work, we aim to prototype a pipeline that combines tools from both complex networks and statistical natural language processing to study an existing real-world problem of climate communication polarization. We also demonstrate that it is possible to obtain quality labels for supervised learning by using unsupervised methods which alleviates one of the major bottlenecks of applying deep learning pipelines to unclassified real-world data.

The project has practical relevance in demonstrating how state-of-the-art deep learning models can be utilized for a relevant real-life problem in sentiment analysis and opinion mining. Opinion mining of such political topics on social media can help analyze both temporary trends as well as more deep-rooted attitudinal positions in public discussions. Clearly, as this is a course project, a large practical use and motivation for the whole group is to learn about BERT models and their use in natural language processing.

3. Methods

3.1 Data

The data set consists of 1.6 million climate-related tweets of which 0.4 million are original tweets and 1.2 million are retweets. Data was collected during August 2020 and it was obtained for our use with the permission of ECANET research group[1]. Data was prefiltered during Twitter collection by defining a search keyword *climate* in the Twitter streaming API query.

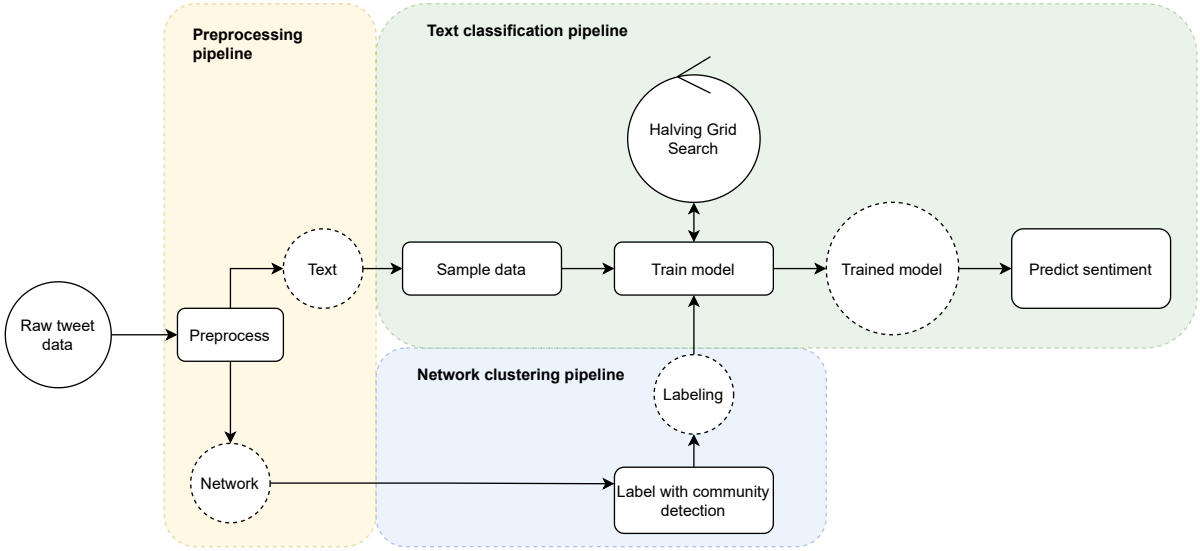


Figure 3.1: Diagram showing the combined sentiment analysis pipeline.

3.2 Preprocessing Pipeline

In their raw format, tweets are represented as JSON-objects (JavaScript Object Notation, an open standard file format consisting of attribute-value pairs), which we parsed into appropriate input formats for both the unsupervised network clustering pipeline as well as the supervised text classification pipeline. Data was thus split into two complementary sets; Interaction data, particularly about retweets, was used for detecting communities in the networks while non-retweets were used for text classification.

[1]ECANET (*Echo Chambers, Experts and Activists: Networks of Mediated Political Communication*) is an interdisciplinary research group consisting of both Aalto University and University of Helsinki researchers. See https://www.aka.fi/en/research-funding/programmes-and-other-funding-schemes/academy-programmes/media-and-society-mediasoc-2019-2022/projects/

Network Preprocessing

We constructed a network representation using the tweet metadata on users and interactions between them. Here we discarded all tweets that had no interaction data available and filtered out all non-retweet type interactions. We built the network representation with users as nodes and retweet interactions as edges. The resulting complete network had 809,157 nodes and 1,089,988 edges. We validated the network by computing summary statistics. We computed the component distribution and observed the existence of the largest connected component in the network. This component contained the majority (77%) of all nodes and (98%) the edges. We disregarded all of the other connected components in the network. For the largest component we measured the degree distribution (Figure 3.2), average degree (3.41), density (2.76e-06), diameter (26), average clustering coefficient (1.4e-04) and assortativity (-0.06).

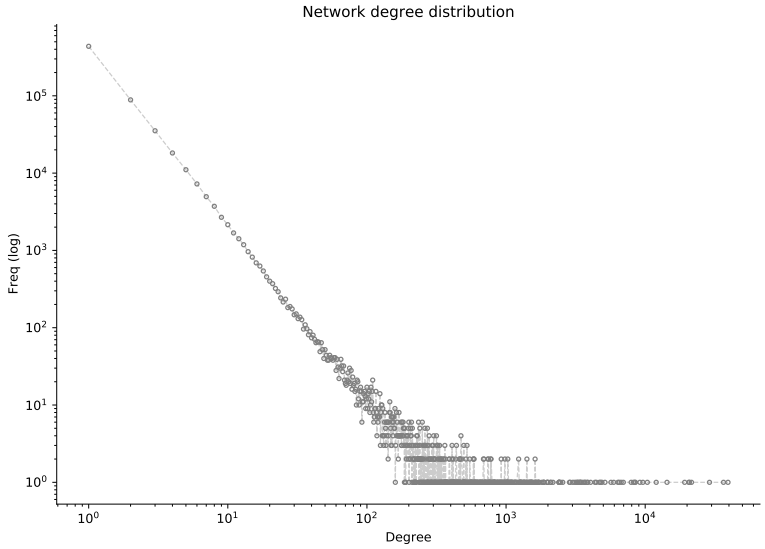


Figure 3.2: Degree distribution for the retweet network in logarithmic scale. Distribution resembles powerlaw distribution which is typical for social networks; frequency of observed user degree is relative to its ranking.

Text Preprocessing

Before the text preprocessing task we filtered out retweets from the data set. By doing this we avoided both model overfitting as well as potentially untrustworthy test accuracy due to repetitive contents in the tweets. For the remaining 0.4 million original tweets we removed hashtags, mentions, hyperlinks and all special string combinations that Twitter uses for its own classification algorithms. We replaced short common abbreviations (such as 'u') with their long-forms ('you') after which we removed unit length characters and extra whitespaces from the text.

Finally, we tokenized and embedded the input tweet texts using model-specific tokenizer for each BERT model. Typically, a BERT or similar tokenizer consists of following subroutines: breaking down of each input sentence into (subword) tokens, adding the [CLS] token at the beginning and [SEP] token at the end of the sentences, padding the sentence using [PAD] tokens to create equal length sentences and mapping each token into corresponding space of natural numbers. The RoBERTa tokenizer has a differing embedding type which has been derived from the GPT-2 tokenizer, and it takes advantage of the byte-level Byte-Pair-Encoding. RoBERTa tokenizer also encodes whitespaces

differently depending on whether whitespace occurs at the beginning of the sentence or if it is the first word in the sentence.

3.3 Network Clustering Pipeline

We obtained labeling for the training and validation of the deep learning models by employing an unsupervised clustering approach specifically designed for efficiently finding dense clusters in large networks.

Towards this we employed the efficient multilevel Leiden [10] community[2] detection algorithm with the CPM (Constant Potts Model) quality function. We first parametrized the model with resolution parameter $\gamma = 1 \cdot 10^{-6}$ which yielded 2248 communities in total. Of these the three largest communities (53.26%, 18.96%, 14.57%) included 86.79% of the nodes in the network with the quality of the partitioning being high (> 0.91). Further fine-tuning led us to select $\gamma = 2 \cdot 10^{-7}$ as the final parameter choice, yielding all in all 182 communities, with two major communities (81.86%, 13.59%) including 95.45% of the nodes in the network. The quality of the partitioning was also very high (> 0.97). We validated the clustering result visually by projecting the network on two dimensional space using ForceAtlas2 [1] force directed network layout algorithm. The final clustering result is shown in the Figure 3.3.

Once we obtained the clustering, we performed a manual validation of the result by associating all available user labels with the tweet contents and sampling tweets from each relevant cluster for inspection. Matching labels and tweets reduced our sample size from 0.4 to 0.15 million tweets. We inspected a small random sample (N=20) for each of the five largest classes to see if the community labels correlated with polarized opinions. Based on this small randomized sample, the opinion in the largest community was generally found to be voicing worry about climate change and promoting mitigation efforts. In the second community we observed the opposing sentiment. We did not find any false positives in the clustering result in these two classes, however the variation between the contents in the messages was high. Third, fourth and fifth community labels seemed to be selected based on other features such as trending discussion topics or geographically bounded discussions mentioning the word *climate*. Some of the tweet contents in these communities were not found to be explicitly related to climate change.

Finally, we assigned the sentiment classes for the communities; sentiment *Positive* to the largest community, *Negative* to the second largest community and rest of the 180 communities with third sentiment label *Unknown/Other*. We based our decision to group the rest of the communities under single class on two observations: i) we didn't identify clear sentiments on the topic of climate change in other communities and ii) the combined size of the rest of the communities in the network was relatively small (only 4.55% of all nodes in the network).

3.4 Text Classification Pipeline

For the opinion mining part, we followed the methods used by Martin Müller et al. when training CT-BERT-V2 [14] and Serena Y Kim et al. when fine-tuning the RoBERTa model for solar energy sentiment classification [13]. We found a set of optimal training hyper-parameters (learning rate, dropout and batch size) by using both Grid Search and Halving Grid Search [7].

We fine-tuned three BERT-models (BERT, RoBERTa, and CT-BERT-V2) using the obtained set of optimal hyper-parameter values. We used the Adam optimizer with decoupled weight decay (AdamW) proposed by Ilya Loshchilov and Frank Hutter [5] as the optimizer and Cross Entropy Loss as the loss function. The structure of the fine-tuning layer is as recommended in the BERT paper

[2] In the complex networks literature the term *community* is used to describe network clusters.

Figure 3.3: Final network communities projected with ForceAtlas2 network layout algorithm. The resolution parameter value used for the community detection was $\gamma = 2 \cdot 10^{-7}$. The two largest communities (pink, green) are clearly separated showing the fingerprint of a polarized discussion network. From third largest community (other colors) onwards, communities are negligible in size.

[6]: the output from the pre-trained model is fed through a dropout layer into a single linear output layer.

4. Experiments

We experimented with fine-tuning models under various conditions. First, we tried different hyperparameters to find an optimal set of parameters for the later experiments. Furthermore, we tested three different models under equal conditions, to see which model is most suitable for our use case. This chapter describes the detailed differences in the setup of the model fine-tuning runs we compared.

4.1 Hyperparameter Optimization

We used grid search to fine-tune the BERT-BASE-CASED model on small datasets with varying hyperparameters to find optimal training configuration. It is a smaller version of the BERT-LARGE-CASED we used in final fine-tuning, so it is suitable for efficient hyperparameter optimization. Due to the high time complexity of performing a grid search on a BERT-like model we assumed that these hyperparameters are also optimal, or at least effective, with the other tested models RoBERTa and CT-BERT-V2, because they are closely related to BERT. Performing searches on the other models is out of scope for this project.

Grid search is a type of search algorithm that performs some process with each combination of the parameters defined and selects the best combination based on some scoring function. Our process is the language model training loop, and our scoring function is the Jaccard index $J(A, B)$:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

where A is the set of predicted labels and B is the set of ground truth labels.

Because a grid search on a BERT-model is computationally very expensive, we used a variant of grid search called the halving grid search proposed by Manoj Kumar et al. [7] The intuition behind it is that it starts with a lower number of samples for each training iteration at the beginning, then eliminates the worst parameter and runs the training again with a larger sample size for the remaining parameters. The samples are increased after each iteration with a chosen factor c. In practice this decreases the training time drastically while not sacrificing search robustness too much, because the worst parameters would likely perform worse even with smaller sample sizes where variance plays a larger factor.

The hyperparameters we initially searched are the learning rate, dropout and batch size. We set the baseline hyperparameters for the grid search to be the ones recommended in the original BERT paper: learning rate $2\mathrm{e}{-5}$, dropout 0.3 and batch size 32 [6]. We searched in the range dropout $[0.1, 0.2, 0.3, 0.4, 0.5]$, learning rate $[2\mathrm{e}{-5}, 1\mathrm{e}{-5}]$ and batch size $[32, 64]$. We used 3-fold cross validation for each parameter combination to prevent overfitting. We set the halving grid search to start at 500 samples with a multiplying factor of $c = 2$ after each elimination. Maximum training sample size was 2000 samples for each three clusters totalling 6000 samples and we ran each iteration for 4 epochs.

4.2 Model Comparison

The combined third class *Unkwown/Other* included only 346 non-retweet tweets and was dropped out from the final fine-tuning data set. We balanced the data set to prevent a bias towards the majority class. Given that the first class (*Positive*) was the majority class we drew random samples equal to the size of the second class (*Negative*) from it. This yielded a final sample size of 49 598 with 24 799 observations for each class. Using a train-validation-test split with a 80%-10%-10% ratio we derived the final fine-tuning data set with 39 678 observations in the training set, 4 960 in the validation set

and 4 960 in the test set. The training was stopped after 1 epoch for an intermediate evaluation and training was continued for second epoch to ensure that model convergence was reached.

With this data, we fine tuned three pre-trained models from the BERT family (BERT large, RoBERTa large and CT-BERT-V2) under equal conditions. During training we monitored the training and validation accuracy of the model 20 times for each epoch. After training, we evaluated the models by comparing their prediction accuracy over a test set against the unsupervised clustering labels. Furthermore, we evaluated the tweets against their labels with manual review of tweet content.

5. Results

5.1 Training Performance

The recommended hyperparameters in the original BERT paper proved to be effective: the halving grid search found the optimal training configuration to be dropout 0.3 (as recommended), learning rate $2e{-}5$ (as recommended) and batch size 64 (batch size 32 was used in the BERT paper). We trained the models with this configuration for three clusters.

Shortly after we found that the model should instead be trained for binary classification because the third cluster consists mostly of retweets. We then trained the models with 2 clusters. Later on, to confirm our prior configuration was still optimal, we also re-ran the grid search with the same parameter options as before but now for two clusters and with a larger sample size. Initial training sample size was 1000, we used a factor of 2 and the maximum sample size was 8000. The parameters discovered with this configuration were, again, dropout 0.3 and learning rate $2e{-}5$. However the suggested batch size was now 32 instead of 64. We expected the batch size to have only marginal effect on the results and thus refrained from retraining the models with batch size 32.

We plotted the training and validation accuracy during fine-tuning for all three binary classification models over two epochs in figure 5.1. As can be seen from the validation accuracy, all models converged relatively quickly. CT-BERT-V2 shows the steepest initial learning. In the second half of the training there was only marginal improvements on the validation accuracy in all models. Hence, we stopped training after this epoch.

5.2 Classification Results

After running the training for each model, we found that all of the models proved to be effective at predicting sentiment against the unsupervised clustering in the task with accuracies above 0.87 on the test set (see table 5.1). CT-BERT-V2 was the best performing model at 0.922 test accuracy, RoBERTa had the second best test accuracy of 0.901 and BERT performed least well with an accuracy of 0.875 on the test set. Table 5.1 shows the final accuracy for each model after 2 epochs of fine-tuning.

Table 5.1: Prediction accuracy for the tested models after 2 epochs.

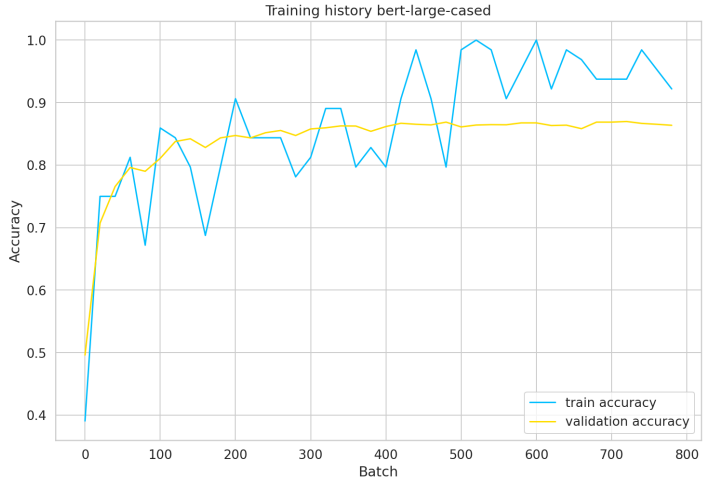
Without retweets		
Model	Train (2nd epoch mean)	Test
BERT-LARGE-CASED	0.934	0.875
RoBERTa-LARGE	0.926	0.901
CT-BERT-V2	**0.944**	**0.922**

In table 5.2 we present several tweets and their predicted cluster labels for the CT-BERT-V2 model. The predictions in the table were randomly sampled from the full preprocessed test data set. [1]

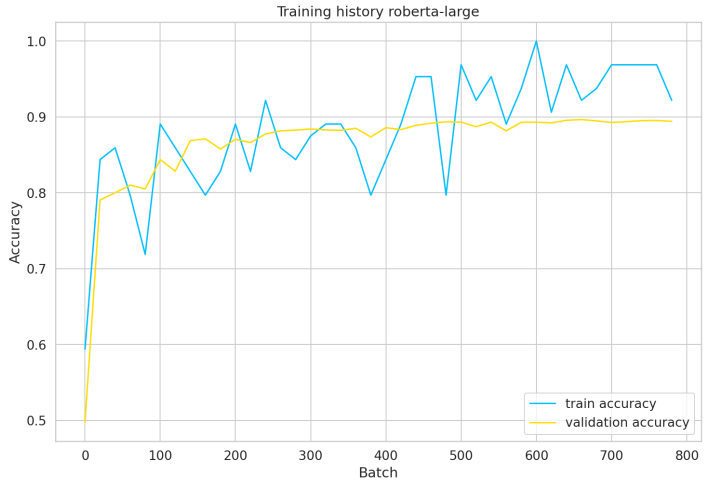
It is quite clear from the tweets that both the labeling as well as the model predictions usually align with the true sentiment well. The positive tweets generally support climate change preventing measures. Tweet #3 is interestingly labeled as positive by both the supervised methods and the predicting model, even though it is not clear what the sentiment of the tweeter is.

Particularly impressive are the last two negative tweets (#5 and #6), as in their case it is not immediately obvious to the reader as to what the writer believes in. They use irony to present the message as negative. This is a hard attribute for a sentiment analysis model to learn, as demonstrated in section 5.3.

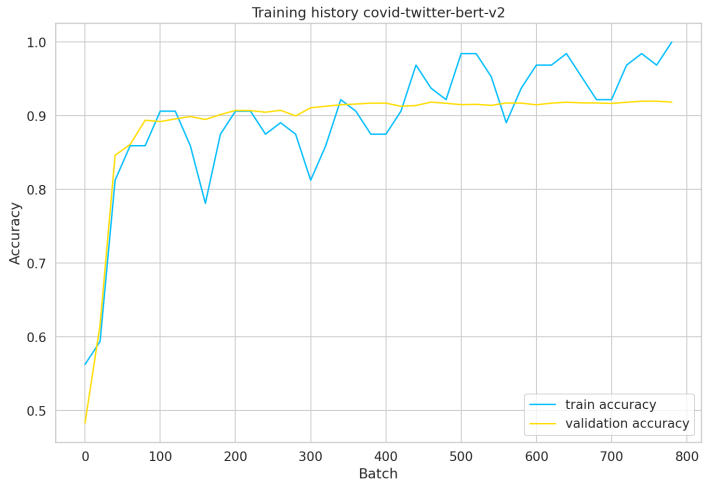
[1] In all of the shown content samples, tweets have been recreated with different wordings. We did this with aim of ensuring anonymity for the original authors of the tweets while trying to stay as truthful to the original meaning as possible.



(a) BERT-LARGE-CASED



(b) RoBERTa-LARGE

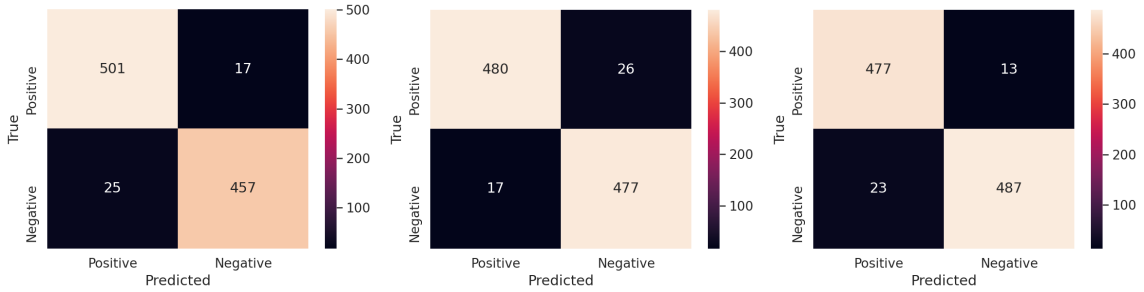


(c) CT-BERT-V2

Figure 5.1: Accuracy on training and validation data during 2 epochs of fine tuning every 20 batches. Models converge relatively quickly. Note, the training accuracy was calculated on the current batch while validation accuracy was calculated on the whole validation set, explaining the difference in variation.

Table 5.2: Tweets and their predicted labels with the fine-tuned CT-BERT-V2 model, the best performing model out of the three models tested. Data without retweets was used.

Correctly predicted tweets with CT-BERT-V2			
#	Tweet	Prediction	Label
1	Big Tech and the Pro-Business coalition are partnering to meet the target set in The Climate Promise; adapt supply chains, support sustainable measurable action; and come up with ways for companies to integrate more holistic climate strategies.	Positive	Positive
2	I keep explaining that organic meat is often worse for the environment, both climate and bd. It's also still red meat and Class 2 carcinogen plus other associated risk factors.	Positive	Positive
3	Trump does away with methane regulations for fossil fuel industry	Positive	Positive
4	COVID19 HAS BEEN FALSELY POSED AS SERIOUS THREAT!!! IT IS ALL ABOUT THE GLOBAL ELITE AGENDA AND RELATES TO CLIMATE CHANGE HOAX!!!	Negative	Negative
5	Indeed. This message is using climate change as the cover for their communist agenda.	Negative	Negative
6	Climate is getting better while we are.... we only have this planet blah blah...	Negative	Negative



(a) BERT large (b) RoBERTa (c) CT-BERT-V2

Figure 5.2: Confusion matrices of each model on test data.

5.3 Misclassifications

None of the models showed a particular preference for one of the classes. Incorrect classifications occurred in all models for either of the classes as can be seen in the confusion matrices in figure 5.2.

We inspected 20 tweets out of which 10 can be seen table 5.3. These tweets show the characteristic tweet types that we could identify among the misclassified tweets. Firstly, there were tweets that were not related to the debate on climate change (#1, #2) e.g. using the word climate in a different sense such as 'political climate'. Secondly, some tweets might have been mislabeled by the unsupervised network clustering (#3, #4). Some tweets might be ironic or use figurative speech. Even though the model was able to classify some of them correctly, it might have taken some others more literally and classified them incorrectly (e.g. #5). Some tweets might also be challenging due to being neutral or missing context, e.g. #9 could be an explanation from either side as an answer to a previous tweet and #10 is missing a word that was likely relevant. Lastly, some tweets are hard to understand due to containing seemingly contradictory arguments (e.g. #7, #8).

5.4 Modified Tweets

We further experimented with the set of misclassified tweets that we had identified in section 5.3 in order to see if we can change their predictions by inducing small modifications. The four modified tweets are presented in table 5.4. Removed parts are struck through and possible additions are

Table 5.3: Tweets and their predicted labels with the fine-tuned CT-BERT-V2 model, the best performing model out of the three models tested.

Misclassifications of CT-BERT-V2			
#	Tweet	Label	Prediction
1	He probably thought it should be understood in the context of the current political climate?	Negative	Positive
2	We used to live in Seattle, which is honestly a great place. After spending time on east coast, my spouse and I want to revisit because of the climate, landscape, drinks...	Positive	Negative
3	I don't really get it why people don't accept the existence of climate change. We are causing this, and it is our duty to mitigate the worst impacts. There is a reason why fighting climate change is high in the list of priorities in our military strategy. If that doesn't persuade you, what will?	Negative	Positive
4	It has been recognized within the Democrat party lines. CC will, like all adverse things in our country, strike hardest the poor people among us.	Negative	Positive
5	This is completely unrelated to record temperature changes and climate change.	Positive	Negative
6	And that will be all put in order once the delusions of the Dems will be carried out as true policies. Yeeha - we can surely save the world from wildfires and climate change while our own people destroy it for profit.	Positive	Negative
7	Joe Biden has a long documented history of deceiving public, including his climate change platform in 2019. His actions speak against him; his promises are empty. Don't be naive	Positive	Negative
8	More on what it means for oil industry in . As senator Harris i) Secured $500m deal with world second largest oil supplier. ii) Helped negotiate the $14b car industry settlement iii) Queried if multiple companies violated the regulations by deliberatively misleading public about climate change	Negative	Positive
9	This demonstration is about responding to the call and really start acting to face the climate crisis	Negative	Positive
10	It's humbling to see us now raising up against the and not just turning blind eye on it without doing anything, like we often have. The Asians will not settle for a climate deal that slows their GDP growth.	Negative	Positive

italized. By simple modifications we were able to change the prediction of two out of four of the modified tweets.

Tweet #1 was long and included religious references. Based on our original manual evaluation we would have labeled it as *Negative* which also agreed with the result of the unsupervised clustering. However, we thought the rhetoric question in the middle "Did we do enough?", might be confusing the model and we wondered if removing it would also help the model predict it correctly, which was the case. We evaluated tweet #2 as *Positive* which again agreed with the unsupervised label. We noted that the tweet used logically rather difficult double negation structure "Nobody is denying that changes in climate aren't happening" which we modified to "Everybody is agreeing that changes in climate are happening". However this didn't affect the prediction. Tweet #3 mentioned *Joe Biden* and was making vivid accusations about him as a character without a strong clear reference to climate change. We wondered if the model had learned to attribute accusations towards Biden to the opposing opinion in the climate debate. We wanted to see if simply changing the name of the well-known US politician would affect the prediction. We changed the name to a more general one, but here prediction didn't change. Tweet #4 shows irony which only the last two words *blah blah* give away, changing the meaning of the whole sentence. By removing the last two words, prediction changed to *Positive* which shows that the model was able to also learn subtle cues such as irony.

5.5 Manually Crafted Tweets

In order to better understand how the clustering method affected the labeling and to detect potential biases learned by the model, we conducted brief test on hand crafted messages. Largest English-speaking geographically bounded area being United States, we tested if the model had learn to recog-

Table 5.4: Misclassified tweets and their modified versions with the prediction of fine-tuned CT-BERT-V2 model before and after modifications. Removed words are struck through with possible replacements italized. Using simple modifications we were able to change the classification result for two out of four of the modified tweets.

Predictions on Modified Messages (CT-BERT-V2)			
#	Tweet	Before	After
1	What do you think will happen at the eve of Redemption? Do you think God will be pleased? ~~**Did we do enough?**~~ Having green lifestyle and all of that is not how Bible guides us to live our lives. Please don't push your ideas about climate change on me	Positive	**Negative**
2	~~**Nobody**~~ *Everybody* is ~~denying~~ *agreeing* that changes in climate ~~**aren't**~~ *are* happening, both humans as well as natural cycles play part in climate change	Negative	Negative
3	~~**Joe Biden**~~ *Roger Dunham* has a long documented history of deceiving public, including his climate change platform in 2019. His actions speak against him; his promises are empty. Don't be naive	Negative	Negative
4	Climate is getting better while we are.... we only have this planet ~~**blah blah**~~...)	Negative	**Positive**

nize US specific political claims. We constructed eighteen variations in total of the following sentence type:

Subject is **wrong/right** about **all/none/some** of the things anyway.

Here subject was referring to either one of the main US presidential candidates in 2020 or one of the two main US political parties. We intentionally left out all references to climate change from the messages. Table 5.5 shows the results.

Interestingly the model was able to differentiate rather well different levels of nuances which we controlled by changing wordings (see table 5.5). Prediction aligned with expected Republican and Democrat positions also when word **all** was substituted with **none** and word **wrong** with **right**, suggesting that i) the model had learned to associate climate opinions with political positions and ii) the fine-tuned model itself has robust language modeling capabilities.

Finally when we changed from strong claims that used words **all** or **none** to using more nuanced word **some** we saw that class *Negative* became more closely associated with claims about political parties whereas *Positive* changed to predict claims about presidential candidates.

Table 5.5: Hand crafted sentences with variable political loadings (left) and their predictions (right). Words **all**, **none** and **some** were injected into the sentences to positions marked with (*). Prediction for class *Positive* is marked by symbol + and class *Negative* with symbol -.

Predictions on Modified Messages (CT-BERT-V2)				
#	Crafted sentence	all	none	some
1	Republicans are wrong about **(*)** of the things anyway.	+	-	-
2	Democrats are wrong about **(*)** of the things anyway.	-	+	-
3	Donald Trump is wrong about **(*)** of the things anyway.	+	-	+
4	Joe Biden is wrong about **(*)** of the things anyway.	-	+	+
5	Donald Trump is right about **(*)** of the things anyway.	-	+	+
6	Joe Biden is right about **(*)** of the things anyway.	+	-	+

6. Discussion and Conclusion

Overall, the accuracy of our models compares well to other state-of-the-art models predicting sentiment on Twitter data about political topics. The RoBERTa model fine-tuned by Kim et.al. [13] reached an accuracy of 0.907 on a binary classification task of sentiment on solar energy in the US, a performance only slightly higher than our RoBERTa model and worse than our CT-BERT-V2 model. In the original CT-BERT-V2 paper [14] accuracy ranged between 0.654 and 0.949 on five different datasets. Our best model achieves performance that is comparable to the best of the original paper.

The differences in training and test performances of our models are as expected given the releases and specialization of the models. RoBERTa was published more recently than BERT as an improvement performing slightly better in many tasks [9]. In our case, this finding was replicated. CT-BERT-V2 showed a steeper increase of accuracy on the validation set during training and had the highest accuracy at the end of training. Both can be explained by the specialization of CT-BERT-V2 on Twitter data. While BERT and RoBERTa had to learn the specific use of language on Twitter, CT-BERT-V2 only needed to fine tune to the polarity of opinions on climate change.

When training the models we opted for using the same hyper-parameters for all models for better comparability and due to the high computational cost associated with performing the grid search. However, the optimal hyperparameters for training CT-BERT-V2 and RoBERTa might be different from BERT. Therefore, a slight increase in accuracy might be possible to reach by searching optimal hyperparameters for each model. This could be done by performing a grid-search for each model. We decided to leave this for future work, given that both models already outperformed BERT. We also omitted to select the lower batch size suggested by the latter grid search run. Our assumption relied on the expectation that this would mostly affect speed of convergence, but we must acknowledge that together with learning rate it could have also affected the dynamics of gradient descent optimization. We still expect the potential effect to be marginal.

For handling the imbalance of the data set, we decided to sample from the majority class to obtain a balanced dataset. Alternatively, the imbalance could be kept, which would allow for using more data and, hence, possibly increase the performance of the models. In case the imbalance would affect the prediction negatively, the loss function could be changed to pay more attention to the minority class. Furthermore, accuracy would need to be complemented with measures that are less sensitive to imbalance such as the F score, precision or recall. However, we also noticed that, with the available balanced data, the models were already converged or close to convergence after only one epoch, so that the improvement through more data might only be minor.

Manually analysing misclassified tweets showed us that removing hashtags during pre-processing might have led to a loss of relevant information. Hashtags are often used in the middle of the sentence and, therefore, carry semantical information about the sentence context. Removing only the hash symbol from the tags, might have kept this information intact and increased the likelihood of predicting the tweets correctly.

Most opinion mining on Twitter data that we are aware of use manually labeled tweets. While human-made labels have the advantage to clearly represent the intended classes *positive*, *negative*, and possibly *neutral*, they incur a much greater cost in the form of human labor cost. Our approach of using unsupervised clustering is potentially advantageous in that it can efficiently label large data sets, but at the same time it relies on the assumption that labels are represented by clearly separable clusters in the network. In our case both visualizing the resulting clusters using network layout algorithms and inspecting Tweets manually helped to confirm that the assumption holds.

Related to this, we also made the following observations: i) based on the density and clustering measurements, the observed retweet network is on average extremely sparse ii) the discussion topic is already known to be polarizing[2] and iii) in our data set climate attitudes generally align with the current bipartisan political divide in US. When put together these factors allow the unsupervised network clustering to uncover the major clusters efficiently. However, finding two large clusters with opposing attitudes may not be possible in other cases, especially cases where network topology is fragmented into multiple smaller communities or is more homogeneous in structure.

We also want to stress that some noise in the clustering result is unavoidable as users may voice opinions that are divergent from the views of the group they usually interact with or inconsistent with their other posts. An example such a post that we noticed was a neutral to positive post by a news source that belonged to the negative sentiment cluster about research on melting arctic ice. There exist noise handling techniques that can be applied in cases of automatically generated noisy labels [12] that could be explored in future work to further improve the results on this dataset.

Predictions for hand crafted tweets seemed to confirm the hypothesis that the bipartisan divide in the US politics is strongly associated with climate sentiments in the data. On the other hand, while political parties did show to influence the prediction when representing the sole subject of the sentence (as in the hand-crafted examples), the topical content seemed to have the bigger impact in other examples. For instance, for the examples sentence #3 in table 5.4 where Joe Biden was replaced by a neutral name, the prediction stayed the same. Also example #8 in table 5.3 criticising Kamala Harris was predicted to be *positive* even though criticism of Harris would be associated with the *negative* cluster.

Finally, we saw that class *Negative* became more closely associated with claims about political parties whereas *Positive* changed to predict claims about individual political candidates. Reasons for this are unclear. We think that one possible interpretation would be that users with differing climate opinions are using language differently, namely people whose messages align with class *Positive* and seem to be generally more aligned with Democrat positions, could be more likely to make claims about individuals than groups, whereas people who use class *Negative* and are generally aligned with Republican positions, could be more likely to make claims about groups instead of individuals. However the data was collected from rather short period and this difference could also be caused by temporal changes in topics within different parties. This trend was seen the most in hand crafted posts with the least strong claim towards one political party where the model might have had a lower certainty (i.e. “Donald Trump is wrong about some of the things anyway” could be uttered by one of his supporters while “Donald Trump is wrong about all of the things anyway” could not). This might also indicate that the effect of referring to groups or individuals has a weaker influence on the prediction than the political position of the mentioned entity.

We positively noted that our model was able to detect irony at least in some cases. The original example #4 in table 5.4 is such an instance and changing the sentence so that it was not ironic also changed the prediction. It also might have interpreted example #6 in table 5.3 as ironic (which misled it in this case because the irony was not on climate change). However, the model also missed the irony in some less obvious cases such as example #5 in table 5.3. Possibly, the model relies on key words such as “blah blah” or “yeeha” to detect irony. Further study of ironic examples would be needed to verify this hypothesis.

All in all, we automatically generated quality labels through unsupervised clustering on a large set of Twitter data and analysed the underlying opinions of the major clusters. We searched optimal hyperparameters for training a BERT model on this data. We fine-tuned three transformer-based language models from the BERT family to reach a high accuracy in predicting the labels based on the text of Tweets. Finally, we analysed the strengths and weaknesses of our best model (CT-BERT-V2) based on original and modified posts.

7. Division of Labor

Ville contributed by setting up the initial data preprocessing pipeline and the clustering pipeline. He also did most of the preprocessing of the data and ran early versions of the models. Lena and Samuel set up the template for the training. Samuel set up the grid search, ran it and ran most of the fine-tuning runs. Everyone participated in tuning the models, analysing intermediate and final results as well as writing. Everyone also contributed to the pipeline code.

8. Acknowledgements

We would like to thank the interdisciplinary ECANET group (Echo Chambers, Experts and Activists: Networks of Mediated Political Communication) at Aalto University and University of Helsinki for providing us with over 1.6 million twitter posts on the topic of climate change.

Bibliography

[1] Mathieu Jacomy et al. “ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software”. In: *PloS one* 9.6 (2014), e98679.

[2] Hywel TP Williams et al. “Network analysis reveals open forums and echo chambers in social media discussions of climate change”. In: *Global environmental change* 32 (2015), pp. 126–138.

[3] Riley E Dunlap, Aaron M McCright, and Jerrod H Yarosh. “The political divide on climate change: Partisan polarization widens in the US”. In: *Environment: Science and Policy for Sustainable Development* 58.5 (2016), pp. 4–23.

[4] Jack Zhou. “Boomerangs versus javelins: how polarization constrains communication on climate change”. In: *Environmental Politics* 25.5 (2016), pp. 788–811.

[5] Ilya Loshchilov and Frank Hutter. “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (2017).

[6] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[7] Manoj Kumar et al. “Parallel architecture and hyperparameter search via successive halving and classification”. In: *arXiv preprint arXiv:1805.10255* (2018).

[8] Jennifer McCoy, Tahmina Rahman, and Murat Somer. “Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities”. In: *American Behavioral Scientist* 62.1 (2018), pp. 16–42.

[9] Yinhan Liu et al. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019).

[10] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. “From Louvain to Leiden: guaranteeing well-connected communities”. In: *Scientific reports* 9.1 (2019), pp. 1–12.

[11] Ted Hsuan Yun Chen et al. “Polarization of Climate Politics Results from Partisan Sorting: Evidence from Finnish Twittersphere”. In: *arXiv preprint arXiv:2007.02706* (2020).

[12] Michael A Hedderich et al. “A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios”. In: *arXiv preprint arXiv:2010.12309* (2020).

[13] Serena Y Kim et al. “Public Sentiment Toward Solar Energy: Opinion Mining of Twitter Using a Transformer-Based Language Model”. In: *arXiv preprint arXiv:2007.13306* (2020).

[14] Martin Müller, Marcel Salathé, and Per E Kummervold. “Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter”. In: *arXiv preprint arXiv:2005.07503* (2020).

A. Source code

Notebooks describing different phases of the project with intermediate results are publicly available via github: https://github.com/Decitizen/SNLP-Project-2021