Detecting Disinformation on Twitter targeting Non-profit organizations

Philipp Guldimann
ETH Zürich
guphilipp@student.ethz.ch
June 27, 2022

Supervisor:
Rebekah Overdorf (rebekah.overdorf@epfl.ch)
& Ryan Cotterell (ryan.cotterell@inf.ethz.ch)

Abstract

In recent years, social media has repeatedly been used to spread wrong information. The Russian incident during the presidential election of Donald Trump has been a significant event leading to more focus on how social media platforms permit the spreading of wrongfully created content with malicious intent. Nonprofit organizations rely on social media to operate in conflict areas and support minority groups. Disinformation may lead to the inability to help efficiently. With this bachelor thesis, various methods used in conjunction are reviewed and evaluated to detect disinformation against the International Committee of the Red Cross (ICRC). Concretely, methods like entity recognition, sentiment analysis, and topic analysis are combined with the powerful machine learning paradigm of transfer learning, relying heavily on the Transformers architecture. Furthermore, we will evaluate each part of the constructed pipeline individually to ensure a high level of robustness. Multiple tweet datasets collected on the ICRC will be used and hand-annotated to base the evaluation on domain-specific knowledge. The project is about measuring possible threats against nonprofit organizations. After perceiving such a thread, the evidence collected should be presentable to a nonprofit organization so that it can act accordingly. The goal of this thesis is to provide nonprofit organizations with the automated process to visualize and identify disinformation campaigns effectively.

1 Introduction

The concept of disinformation campaigns has been around for a long time. With widespread access to the internet and various social media platforms, it has become a possibility to be spread by everyone with access to the internet. The possible results of such an attack can be seen in the especially fatal cases of the US 2016 election, and the manipulation attempts of the elections carried out by the Internet Research Agency, which is based in St. Petersburg, Russia. (Bastos and Farkas, 2019)

At the École Polytechnique Fédérale de Lausanne (EPFL), Distributed Information Systems Laboratory, a study is performed to help the ICRC build an early detection system of targeted disinformation. In 2018, the box incident showed that Misinformation against the ICRC does happen and can be misleading and dangerous. In this incident, videos of trunks with money having a red cross emblem appeared. (ICRC) On social media, it was mainly interpreted as the ICRC smuggling money for an illegal cause. The carrier of the truck turned out not to be an entity of the ICRC but rather an approach to target the legitimacy of the ICRC. Efforts must be implemented to safeguard the ICRC and non-profit organizations against such future attacks.

Especially right now, in the context of the Ukrainian-Russian crisis, disinformation seems to affect the ICRC. False information against the ICRC is spread on various social media sites according to which the "ICRC is participating in forced evacuations of Ukrainian refugees to Russia and negotiating the opening of a refugee camp in southern Russia." (Swissinfo)

Those incidents raise the question if there is a way to detect such targeted attacks against the ICRC. Coming up with a suitable and maximal efficient detection mechanism is exactly what this paper is trying to solve. Before delving further into the technical aspects, it is worth understanding the distinction between disinformation and misinformation. Both are sometimes used interchangeably by the news media. Misinformation is describing information which is wrong. However, there is not necessarily any malicious intent behind it. Disinformation, on the other hand, carries intrinsic intentionality, as further described by (Tucker

et al., 2018). When an unknowing end-user forwards disinformation, it can take the character of misinformation since the intention is most likely only associated with the user of the original post. Taking this perspective, detection of not only disinformation but more general misinformation against the ICRC is necessary.

Moreover, hate speech is another categorization. Hate speech does not necessarily need to arise from a planned disinformation campaign. Nonetheless, it is crucial to capture since it may refer to information propagated by disinformation campaigns.

To avoid ethical problems, we want to be able to measure the likelihood of tweets being disinformation. Special care must be taken to avoid accidentally promoting a particular identity group or introducing unfair biases. Introducing such a bias would pose an abuse of power. Transparency is vital, and every part of the project needs to be clearly understood to enable the third parties to verify the legitimacy of our method.

We do not want to carry out the final decision if some tweets are harmful or not. These tasks should be left to the human supervisor working at the non-profit organization. Also, as social media posts are highly convoluted, we aim to increase the average blocking of hate speech and disinformation campaigns rather than being an absolute protection against them.

One important thing to note is that we focus much more on recall than precision. Lower precision is not as bad since we require a human supervisor to review the results. Recall, however, is critical since all potentially harmful content must be detected early on.

In the next section 2, an overview of the concepts and methods used in this paper from to fields of Natural Language Processing but also Machine Learning will be given. Additionally, related work will be summarized and put in context. The 3 section will deal will with the main part of constructing a pipeline that solves the problem of detecting disinformation campaigns. The results will be criticized in the final 4 section, and ideas for further improvement will be discussed.

2 Related Work

This paper uses recent advances in Natural Language Processing, including sub-fields of Named Entity Recognition, Zero-Shot learning using Transformers, and Sentiment Analysis. A quick introduction to these sub-fields will be given in the following subsections.

2.1 Topic Analysis

Topic Analysis is the discipline of extracting common topics from a collection of documents. There are different ways of modelling each document. Viewing each document as a bag of words is a standard way. This representation was also consistently used in this project. In the case of this paper, each document consists of precisely one tweet.

One prominent method is Latent Semantic Analysis (LSA). Using LSA, the collection of tweets can be represented as a matrix. Each document is associated with a column and each term of the vocabulary with a row. Moreover, each entry can be weighted according to the inverse term frequency. Weighting each element of the matrix inversely proportional to the occurrence of its words allows common but irrelevant words like stop words to be filtered out.

Standard linear algebra methods can be applied in formulating the collection of documents as a matrix, including Singular Value Decomposition and Dimensionality Reduction. In Latent Semantic Analysis, precisely those methods are used. When performing the SVD decomposition on the above-described matrix, we end up with a diagonal matrix with singular values in decreasing order. The left-adjoint orthonormal vectors are the topics of the LSA method. The topic vectors corresponding to the highest singular values equal the most meaningful topics meaning that they capture maximal variance of the document corpus. Each left-adjoint vector describes a linear combination of terms, indicating how strongly each term correlates with a topic.

Another approach is the so-called Non-negative matrix factorization (NMF). Another technique is used instead of factorizing the document-term matrix using the SVD decomposition. The matrix is approximatively factorized into two non-negative matrices U, V, such that $UV = X$. This factorization is possible since matrix X contains per definition only positive elements. This problem has been shown to be equivalent to minimizing the K-means clustering objective whenever V is an orthogonal matrix (Ding et al.). The factorization exhibits the benefit that U and V often have lower dimensionality than X. Therefore it has the effects of dimensionality reduction coupled with clustering

	precision	recall	f1
Roberta	0.29	0.25	0.27
S-NER Segm	0.62	0.35	0.44
S-NER Class	0.30	0.27	0.29

Table 1: Baseline of entity recognition on tweets

effects.

One final, very efficient approach is the statistical method called Latent Dirichlet Analysis (Blei et al., 2003). It is based on a generative approach, modelling the topics using latent variables. Not only the topics but also the documents and the terms are modelled using latent variables. As a result, LDA has a high degree of expressiveness since the assumptions are as minimal as possible. The estimation procedure is a variation of the Expectation-Likelihood maximization method.

2.2 Entity Recognition

The field of Entity Recognition has been around for a long time. Its goal is the retrieval of entities inside a document. What one considers an entity needs to be specified. Prominent entity classifiers distinguish between Person, Organization, and Location classes. Usually, pronouns like she and him are excluded as those can be found more easily by Part-Of-Speech tagging. An appropriate annotation scheme must be chosen once suitable classes have been selected. Typical annotation schemes include IOB (Inside-Outside-Beginning) and IOB2. In this project, IOB2 is used.

The whole document is split into tokens, whereas each entity is a continuous span of those tokens. Each document token is annotated with an *O* when it is outside an entity, *I* when it is part of an entity, and *B* when it is the start of an entity. After the entity boundaries have been annotated using these three tags, the class can be appended to the end of each tag. For example, *B-Org* would specify the start of an Organization entity. State-of-the-art transformer architectures like Roberta (Liu et al., 2019) can achieve high f1, recall, and precision scores.

The Roberta model is trained on large text corpora like Wikipedia. Being trained on highly structured corpora, those models perform worse on noisy texts like social media posts. Social media posts contain a lot of non-standard language and have standard abbreviations which are not generally known. Moreover, Twitter-specific constructs like hashtags and mentions make it even more difficult. Performing Entity Recognition on Tweets experiences a significant drop of the f1, recall, and precision scores.

Further, fine-tuning the Roberta model for the entity recognition task directly on the tweets yields low results, as seen from the results in 1. The Stanford NER method (Finkel et al., 2005) yields also low results, but performs better then the Roberta model used directly on Tweets. Those methods will be described further in the next section.

There must be a distinction between the Recognition of entities segmentation and entity detection. The former specifies the task of segmenting entities correctly without further predicting the entity class they belong to. Then Entity Detection is the prediction of the corresponding type of the entities. In this project we are interested mostly on the first one, since we are interested on the negative sentiment towards entities and don't distinguish between the classes.

2.2.1 Evaluation of Entity Recogntion

Before moving on, I want to clarify the metrics f1-score, precision, and recall as used in (Pedregosa et al., 2011). The four fundamental metrics are the false positive rate fp, the true positive rate tp and corresponding for the negative results, false-negative rate fn and false positive rate fp From these metrics, we can construct the precision score $precision = \frac{tp}{tp+fp}$ describing how many of the classification attempts were correct. On the other side, the recall rate is expressed as $recall = \frac{tp}{tp+fn}$ and denotes how many positives were identified out of all the positives. A combined measure of those two metrics is given by their geometric mean expressed as the f1-score $f1 = \frac{2*precision*recall}{precision+recall}$.

In the event of binary classification, those metrics are explanatory enough. In the case of multiple classes, an average of the scores for the different classes needs to be computed. There are three major approaches to this problem: the micro average, macro average, and weighted average. The micro average doesn't distinguish between classes and considers the positives and negatives of each class as the same. This score doesn't work for imbalanced data as well as a result.

The macro average, on the other hand, takes the average of the precision, recall, and f1 scores of each class with uniform weights. By further adjusting the weights according to the sample size

	precision	recall	f1
T-NER Segm	0.62	0.35	0.44
T-NER Class	0.73	0.49	0.59

Table 2: Baseline results for tweet entity recognition from (Ritter et al., 2011)

of each class, we arrive at the weighted average.

If we have multiple tweets, there is also the question of calculating the total precision, recall, and f1 scores. There are two ways to do that. The first method is calculating false/true positives and negatives for each tweet and then summing them up.

On the other hand, we can also look at the whole collection of tweets and identify the set of entities detected for each class. In a second step, false/true positive and negative rates are calculated. The true positive would then, for example, correspond to the set of entities detected of a class versus all the annotated entities of the same class. This method has the benefit that the set of entities dismisses duplicates and treats them as one. Therefore, this method is not biased towards entities that occur more often. In case of the first method of summarizing the results, the name Entity Mention Detection (EMD) will be used. The name Entity Detection (ED) is used in the second method. Those names are chosen to be in accordance with the codebase and paper by (Bhowmick et al., 2022) that are used as a starting point for this project.

2.2.2 First attempt

One of the first attempts to create Entity Recognition specifically geared toward tweets was carried out by (Ritter et al., 2011). Their approach is based on feature generation. In the method they came up with, they used both part-of-speech tagging and chunking to generate features that are used to train a final classifier.

They explain that the many Out-Of-Vocabulary-Words in tweets have a negative impact on the results. Further misclassification stems from the various capitalization patterns found in tweets according to them. Also, they argue that the grammatical structure of tweets is often different, having a verb at the start of a sentence instead of a noun. According to their work, hashtags and mentions are captured using regular expressions and thus can be found with a certainty of 100%. Both of them are not viewed as entities in this project for this reason, since they can be easily extracted by regular expressions.

Next to entity segmentation, entity classification is also tackled. Their paper states that categorizing entities on tweets is complicated since entities often consist of many modifications and changes and are not necessarily found in standard dictionaries. The use of numbers and various short-hand versions adds to this problem. Therefore, using an ontological dictionaries only like Freebase is not a suitable method, according to the paper. Instead of a fixed dictionary, a topic model based on the freebase dictionary and their own generated clusters of variation of the dictionary entries are used. This approach can be seen as a distant supervision task. For the clustering method, Brown clusters are used to find clusters representing different versions of the exact words to capture the whole lexical variety of tweets.

Having gathered various features describing the tweets, a Conditional Random Field is used to predict the Entity Segments. Asides from POS-tagging, shallow parsing or chunking is performed to identify the different grammatical compounds of the tweets. Capitalization is the building block for additional features as well, since different capitalization is used extensively in tweets and often indicates entities.

The final prediction is made using a conditional random field using a combination of all those features. The results show a significant improvement of 51 percent compared to the results of the Stanford NER prediction. (Finkel et al., 2005) Their results for both the task segmentation and classification are shown on 2 One limitation of tweets is their restriction to 140 characters which leads to the fact that often necessary content is missing in tweets to decide whether specific tokens are entities or not. What is more, annotated data for training is relatively limited. For languages other than English, available data is even scarcer.

2.2.3 State of the Art

Neural Networks Fast-forward 6 years, a competition called WNUT-17 took place in 2017, which had the task of detecting entities of tweets. They provided a dataset of annotated tweets for developing, testing, and validation. After the submission, the competition organizers evaluated all the solutions on a private dataset. The competition's task was segmenting and classifying entities according to the entity classes Person, Organization, and Location, among others. The winner of this

competition was the paper from (Gustavo Aguilar, 2017).

They use the multi-task learning approach, combining an entity segmentation task with the main study of the classification, to improve the results. They are using neural networks to capture lexical and orthographical information as well. On top of that, they are using gazetteer lists to capture familiar entities. Similar to (Ritter et al., 2011), they are using a Conditional Random Field as the final step to arrive at a sequence of entity classifications. The first part of the proposed pipeline shows the recent trend of relying more on neural networks to generate meaningful features instead of hand-crafting them.

Using Transformers The rise of the Transformer architecture and it's easy accessibility through Huggingface (Wolf et al., 2019) led to a new way to tackle entity recognition on tweets. Next to general language models like Roberta, there exists a model which is trained explicitly on a large Twitter corpus of around 80 gigabytes called BERTweet (Nguyen et al., 2020). This model was the primary building block of this thesis.

A global view So far, the approaches have tackled the problem by looking at each tweet individually. However, tweets following the same topic and following each other timewise usually use similar phrases and entities. Taking a global perspective gives rise to the possibility of analyzing entities of tweets by keeping track of entities encountered so far. Precisely this was carried out by (Bhowmick et al., 2022). They approach the problem by looking at a stream of tweets instead of individual ones. They divide the whole model into two sub-parts, one acting on individual tweets and one operating on a global representation of multiple tweets.

The local part is mainly based on the BERTweet model or the model developed by (Gustavo Aguilar, 2017). It starts by performing Entity Mention Detection on each tweet. The entities from different tweets are then accumulated into a prefix tree data structure.

After having a collection of all entities occurring in the tweets, a second pass over the tweets detects missed entities of the first pass. Additionally, it will classify whether the entities are good, bad, or ambitious candidates. In regular operation, only good and ambitious candidates will be included in the final selection of entities.

2.2.4 Transfer learning and Transformer models

For a long time, recurrent and convolutional neural networks were the main backbone for various language tasks, including entity recognition. One particular issue of recurrent networks is that they cannot be parallelized efficiently since they are built on a linear sequence.

A new architecture gave rise to shorter training times and better efficiency due to the allowed parallelism of the Attention-mechanism, developed by (Vaswani et al., 2017). That architecture had an important impact on the transfer learning paradigm, enabling a model to be reused by further fine-tuning it on a more specific dataset. The Web platform and library Huggingface make it easy to access many different Transformers models, build on them, and finetune them for various tasks.

Based on this new architecture, the language model BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) was introduced, which is pre-trained on a large language corpus, effectively giving the model the ability to understand the language. However, the language is not explicitly encoded but implicitly captured through the encoder-decoder architecture of Transformers.

In 2018, a new model called RoBERTa (Robustly optimized BERT Approach) was proposed, further increasing the performance of the BERT model by rigorously analyzing the effect of various hyperparameters and making a careful parameter selection leading to further performance improvements. (Liu et al., 2019)

2.3 Sentiment analysis

Understanding the sentiment of a sentence or tweet is possible using transformer-language models like Roberta. Finetuning the language model on sufficient hand-annotated data, the model can achieve a satisfactory performance level. Depending on the annotated data, bias may be introduced. Also, the definition of what is considered positive and negative sentiment is highly individual and depends on the annotation scheme in use. The sentiment models are simple classifiers finetuned on a pretrained transformer language model. A typical selection of class labels is *Positive*, *Neutral*, and *Negative*. The output for each class usually corresponds to the softmax over all three outcomes. The outcomes correspond to the confidence of the tweet belong-

ing to a particular sentiment rather than measuring the polarity of how intense the emotion is. The performance of sentiment analysis is analyzed and summarized in (Mohammad, 2020). Another survey is presented by (Hartmann et al.)

3 Methodology

3.1 Validation on other nonprofit organization

Initially, the idea was to test the final pipeline on a different non-profit organization. The first task was to find an appropriate non-profit organization. There exist a vast amount of different types of non-profit organizations. They are not only diverse in what they do, topic-wise but also in how they do it. The interaction with the public is different. Some organizations work passively by researching various human rights violations, while others actively engage with the affected people to have a direct impact. Of course, many organizations carry out both tasks.

After looking at varios non-profit organizations the United Nations of Human Rights Council (UNHCR) was chosen. The UNHCR is interesting since it is not only a non-profit organization but also a multi-national institution. It spans a vast amount of countries and deals with many different issues. Additionally, it doesn't fall under the category of NGO's. Working with them directly would have allowed us to get an accurate wordlist that encapsulates all entities part of the UNHCR composed in different languages. We arranged a meeting with them to explain our project, and they were interested in it. However, we ended up not hearing back. Since finding another non-profit organization would have been too time-consuming and there was little time for this thesis was limited, we decided to focus on the ICRC for validation solely. In the light of recent activities in Ukraine, choosing to stay with the ICRC makes sense and delivers an excellent validation potential.

3.2 Collecting Tweets

To gather a sensible dataset for training, the author tried to get access a Twitter developer account. However, the request got denied, so the retrieval of tweets could not be performed by the author himself. Therefore, the supervisor of this thesis helped with the acquisition of tweets from Twitter. Besides the data directly from Twitter, it was also attempted to extract tweets from the internet archive (`archive.org`). Since this website stores the past version of websites locally, it was possible to collect some of the archived tweets.

3.3 Datasets

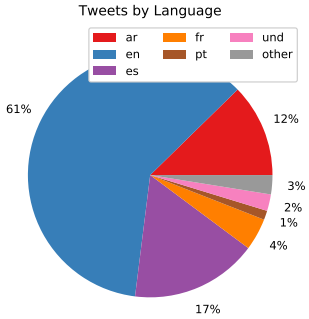
Various publicly available datasets for Entity Recognition have been used throughout the project. In addition to those data sets, we also utilized a private collection of tweets relating to the ICRC. To capture as many tweets as possible relating to the ICRC, we requested a list of keywords containing all the entities belonging to the ICRC written in different languages. This list also includes the Twitter user names for all the entities part of the ICRC. Throughout the project, seven tweet datasets were used, four private and three public ones.

First, a data set composed of tweets with ICRC keywords collected between June and August 2021 was used. This list was mainly used for acquiring an understanding of different Natural Language Processing methods.

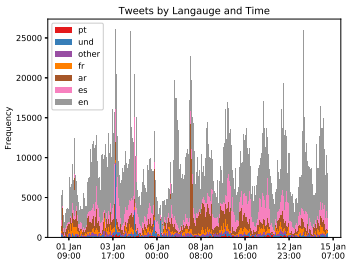
A more comprehensive dataset consisting of a wide variety of tweets relating to the ICRC was gathered as well, spanning the time range of February 2018, the time when to box-incident of the ICRC was happening. The distribution of languages is visualized in 1. The three major languages are English, Arab, and Spanish.

Parallel to this dataset extracted from Twitter, tweet data from the archive was also collected. However, this dataset is noticeably smaller since the internet archive only archives a sample of the Twitter pages each time a scan is done on the Twitter domain. It must be noted that the data collected from this site only contains the start of the Tweets and not the entire content, probably relating to the policy of Twitter regarding which pages are allowed to be scanned by a program. The language distribution is again visualized in 2. English is the most prominent language followed by smaller portions of Portuguese and Spanish. The size of Arab tweets is much smaller than in the collected data directly form Twitter. This result is interesting because it shows that the language distribution on Tweets regarding the ICRC have high variance since the internet archive samples twitter pages in a random fashion.

A further dataset provided by the ICRC is used to evaluate the final pipeline. This dataset consists of tweets relating to the ICRC collected in February 2022. It contains recent development of the ICRC in Ukraine. This dataset also comes with

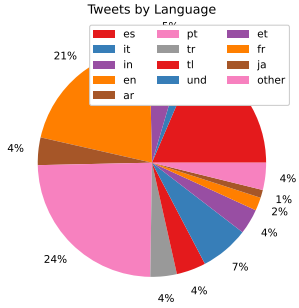


(a) Tweets by language

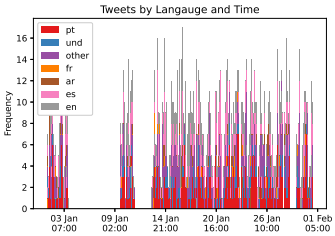


(b) Tweets by langauge and time

Figure 1: January 2018, from Twitter



(a) Tweets by language



(b) Tweets by language and time

Figure 2: January 2018 from `archive.org`

topics assigned to each tweet. The topics partition the whole dataset into a collection of tweets with similar themes.

Next to those private datasets, public datasets are also used for evaluation and training purposes. The dataset provided by the WNUT-17 challenge is used to finetune the entity recognition model on tweets. Apart from that, the data and model are used from (Jain et al., 2019). With the help of their data and model, an alternate model is created to fit the purposes of the directed sentiment detection part.

3.4 Experimental study - A naive version

One has to wonder if it wouldn't be enough to use sentiment analysis to filter out all tweets containing an ICRC keyword. However, there would be a big problem with that. The nature of most of the tweets relating to the ICRC is intrinsically negative. Common sentiment analysis methods quickly tend to tag news tweet about war as very negative. Therefore, this simple approach doesn't work in this scenario. To demonstrate this issue, I compiled a list of commonly occurring words and picked both positive and negative from them. Then I evaluated how many tweets containing those words were attributed to positive sentiment and which fraction to negative sentiment.

Using MNF topic analysis described in 2.1, I automatically created a range of ten topics. Within each topic, I am focusing on the most prominent terms, concatenating all the principal terms of the top 10 topics, and removing duplicates. This procedure leads to a good overview of common words, not just stop words or other frequently used words without any significant meaning. The collection of the ten topics can bee seen on 3. I selected ten words that give a good balance of terms with negative and neutral connotations.

In the next step, I plotted the histogram of tweets with negative meanings for each of the ten selected words. I proceeded with the same for the words with positive sentiment. The graph showing the fraction of tweets with positive versus negative sentiment for each selected word is more interesting. This graph shows negative sentiment is not expressive enough to distinguish whether the tweet contains offensive content. All the graphs can bee seen in 4.

To accomplish the sentiment analysis, the "siebert/sentiment-roberta-large-english" model from Huggingface was used. The model was executed for each tweet from the first dataset.

Before running the model, I preprocessed the input according to the following rules. I removed

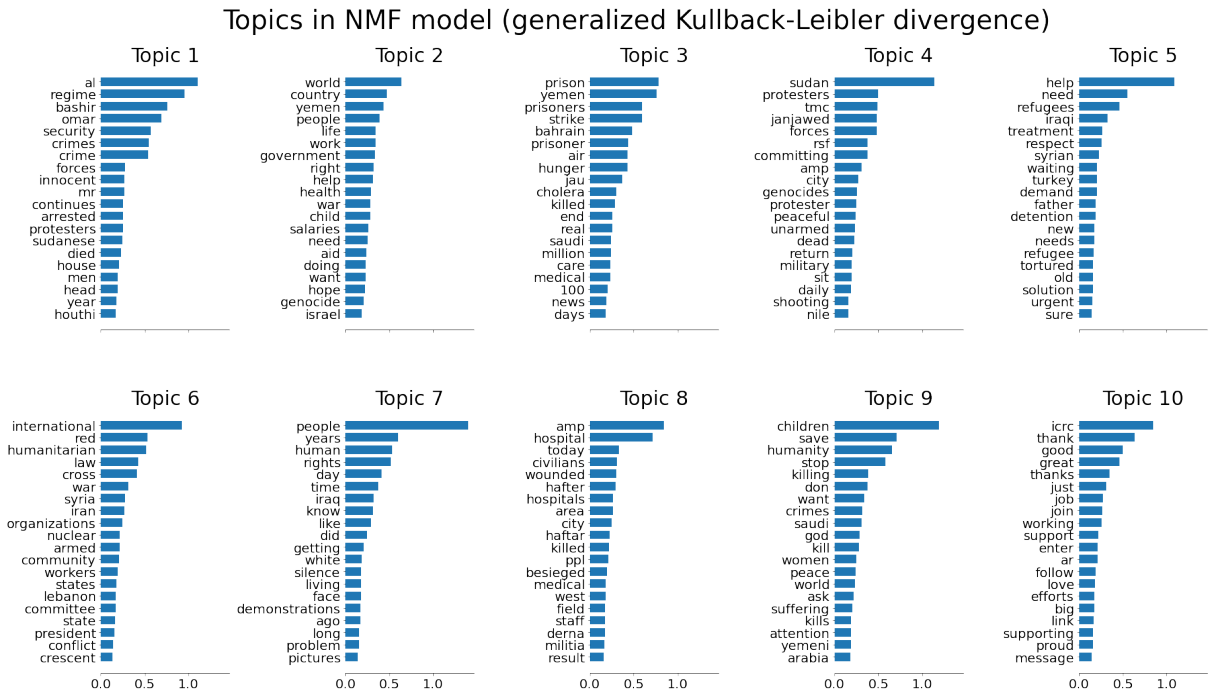


Figure 3: Topics

	precision	recall	f1-score
LOC	0.72	0.39	0.51
ORG	0.56	0.30	0.39
PER	0.59	0.61	0.60
micro avg	0.29	0.46	0.35
macro avg	0.13	0.09	0.10
weighted avg	0.61	0.46	0.51

Table 3: spacy using transformers

	precision	recall	f1-score
LOC	0.48	0.33	0.39
ORG	0.13	0.17	0.15
PER	0.41	0.48	0.44
micro avg	0.18	0.35	0.24
macro avg	0.07	0.07	0.07
weighted avg	0.33	0.35	0.33

Table 4: spacy standard

all the retweeted tweets since they are duplicates of the original tweet. Furthermore, I replaced all URL's with a placeholder to prevent duplicates from tweets with the same content where only the URL alternates by a difference in a numerical identifier within the link. Moreover, I replaced all the hashtags and mentions with a placeholder in the topic analysis phase to exclude typical user handles or common hashtags that don't carry semantic significance.

3.5 Comparison of Common NER methods on tweets

There are various methods to tackle the task of Entity Recogntion. To attain an understanding af various standard methods to tackle this problem, I tried out various tools.

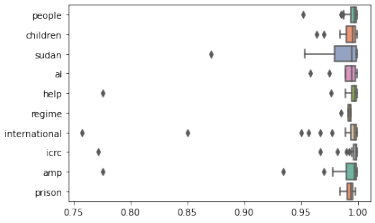
As a first attempt, I tried to use commonly used NER methods directly on the tweet data

	precision	recall	f1-score
LOC	1.00	0.01	0.02
ORG	0.07	0.25	0.11
PER	0.23	0.38	0.28
micro avg	0.13	0.25	0.17
macro avg	0.43	0.21	0.14
weighted avg	0.35	0.25	0.16

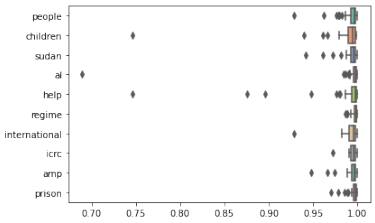
Table 5: NLTK

	precision	recall	f1-score
LOC	0.63	0.71	0.67
ORG	0.34	0.51	0.41
PER	0.61	0.66	0.63
micro avg	0.39	0.62	0.48
macro avg	0.39	0.47	0.43
weighted avg	0.52	0.62	0.57

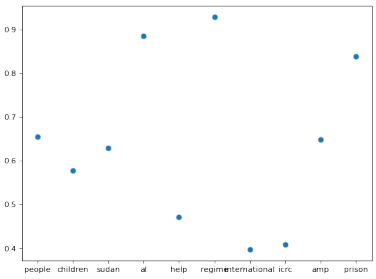
Table 6: Huggingface



(a) Confidence in positive sentiment of tweets containing words on y-axis



(b) Confidence in negative sentiment of tweets containing words on y-axis



(c) Fraction of confidence in positive versus negative sentiment of tweets containing words on x-axis

Figure 4: Analysis of positive and negative sentiment of tweets containing common words

coming from the public conll-2003 dataset (Tjong Kim Sang and De Meulder, 2003).

The most straightforward library for Natural Language Processing is NLTK (Natural Language Toolkit) (Bird et al., 2009). The evaluation results are shown on 5. It is apparent that the performance is quite low with a maximum recall of only 0.48 for the person tag. The low performance is explainable through the fact that NLTK only uses a simple logistic regression to determine named entities. That the method performs best for Person could be due to the reason that persons can be inferred the easiest using simple Part-Of-Speech tagging.

The second table shows the results of using the library Spacy. 4 Spacy is one of the industry-standard library for Natural Language Processing Tasks. It provides convenience methods to set up a whole pipeline, including tokenization, entity recognition, and more. The performance is generally a bit better than the one of NLTK with Person still being the most accurate tag. The better results are explainable through the more advanced used of an BLSTM (Bidirectional Long-/short-term memory) network.

Next, I am using the approach of Transfer learning using the Huggingface library showed on table 6. This approach was explained at the beginning in 2. The results show a clear jump to a higher level of prediction recall and precision rates. Looking at the micro-avg, we can see that it is around 0.62. With the previous methods, the Person tag got predicted the best. This has changed with this method. The Location tag is predicted the best followed by Person, with Organizations being predicted the least accurate. This may be due to the fact that Organizations are described using a variety of different non-standard abbreviations. This explanation was also mentioned in (Ritter et al., 2011).

Finally, I also utilized Spacys new Entity Recognition Backend which uses Transfer learning as well. From the result, this method is inferior to the pure use of Transformer models through Huggingface.

3.6 Named Entity Recognition

3.6.1 Inter-Annotator Agreement

To understand how well two human annotators agree on the entities of tweets, the Inter-Annotator Agreement was calculated using the Cohen-Kappa.

The Annotation itself was performed on a set of 100 English tweets. Two annotators each per-

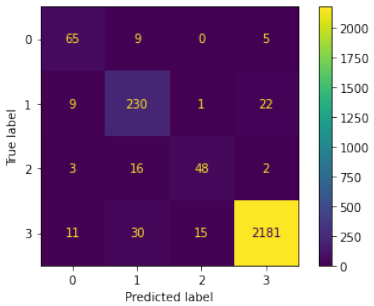


Figure 5: All entities

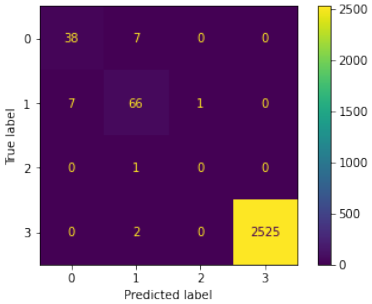


Figure 6: Mention entities

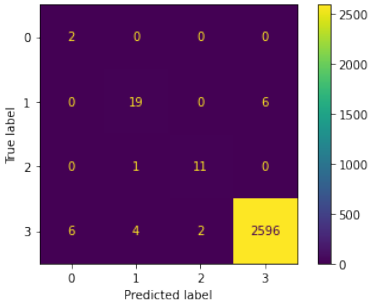


Figure 7: Hashtag entities

	ALL	PER	ORG	LOC
Hashtag	0.77	0.40	0.77	0.88
Mentions	0.92	0.84	0.88	0.00
All	0.84	0.77	0.82	0.71

Table 7: Inter-annotator agreement using Cohen-Kappa

formed the annotation for all the 100 Tweets. The IOB2 scheme was used for this purpose. The class labels ORG, PER, LOC, and NONE for non-entity tokens were used. The open-source software Doccano was used for the annotation process. (Nakayama et al., 2018)

Figure 7 shows the results. The row hashtag shows the scores when only the agreement of annotated hashtags is considered. The same holds for mentions. The last rows All show the agreement scores when the agreement on the whole tweet is taken into account. The columns, on the other hand, indicate whether only Person tags, Organization tags, Location tags, or all tags are used. The inter-annotator agreements for all labels on the whole tweets achieve a score of 0.83. This score is above the threshold 0.8, which is commonly assumed to be the requirement for a significant agreement as argued by Jacob Cohen himself (Cohen, 1960). One value strongly deviating from the other results is the score for only mentions and Location tags. The reason for such a small score is that there are very limited occurrences of such a case in our collection of 100 tweets.

Further, it is observable that the agreement on hashtags for persons is pretty low. This low score could be explained by the missing structure involving hashtags. Usually, multiple hashtags appear at the start and end of a tweet without any specific order. There are also many different hashtags referring to the same entity. The hashtags themselves are usually listed unordered at the end of a tweet to mark specific topics and themes the tweet belongs to.

To highlight the performance of entities, Confusion matrices comparing the true labels against the predicted labels where used. Confusion matrices including all entities, only mention, and only hashtag entities are shown in figure 7. The order of entities is the same for both axis and is the following: *Person*, *Organization*, *Location*, *None*.

All in all, these agreement scores justify the further use of the annotated data for the fine-tuning of the language model.

3.6.2 English only

After seeing the results of commonly used NER methods on tweet data, it is apparent that specialized techniques are necessary to handle Twitter-specific artefacts.

As discussed in 2, two methods stand out in particular. One approach is BERTweet using the

	Precision	recall	f_measure
EMD	0.72	0.62	0.67
ED	0.58	0.53	0.55
Global	0.81	0.73	0.77

Table 8: BERTweet

	precision	recall	f1_measure
EMD	0.62	0.60	0.61
ED	0.52	0.53	0.52
Global	0.80	0.73	0.76

Table 9: BERTweet including hashtags

	precision	recall	f_measure
ED	0.67	0.56	0.61
Global	0.74	0.73	0.73

Table 10: Multitask approach by (Gustavo Aguilar, 2017)

TOTAL ERRORS	MISSED
29	13.00
PARTIAL	**OVERREACH**
12	4.00

Table 11: Classification of errors made by the model

Transformer architecture based on Bert. (Nguyen et al., 2020) Another one is the multi-task learning approach proposed by (Gustavo Aguilar, 2017).

Understanding how hashtags and mentions are handled in the entity recognition procedure is crucial. Since each mention links to a user or organization account and the fact that account names can be very cryptic, all usernames are replaced by the constant @*USER. Each profile handle will be assumed as an entity by default. Since Twitter mentions often appear grouped, one could merge multiple sequentially listed mentions into one single* @USER. This reduction could make it possible to treat all mentions listed at the start as one entity. This pattern of composing tweets is widespread.

Next, we need to deal appropriately with hashtags. Hashtags are keywords making it easier to find the tweets belonging to a particular keyword group. Twitter uses hashtags to cluster tweets together so the Tweets can be searched using hashtags. One can also think of hashtags as tags. Most of the time, multiple hashtags appear either at the tweets' start or end. Therefore, they have a similar pattern to Twitter handles. The groupings of those hashtag groups usually mark the topic the tweet is part of.

Sometimes, the hashtags can also appear in the tweets and are just used as a part of the grammatical structure. The deletion of the # symbol at the start would lead to the correct sentence in this scenario. Heuristic rules could be implemented into the entity recognition preprocessing step to distinguish between those two scenarios described above. Hashtags with no adjacent hashtags would be stripped of the #symbol and, therefore, be part of the text of the tweet itself which can be easier recognized by the model.

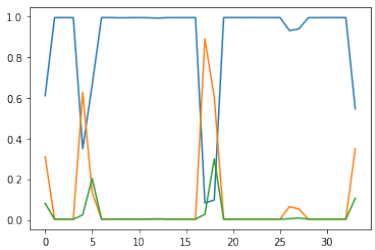
The following results show the performance when Hashtags are considered as candidates as entities (table 9) and when not (table 8). As can be seen, the results differ only slightly, and the specific case of hashtags won't be discussed further in this paper. A potential use case, however, is outlined in the final section. 3.8.

The results of aguilar (table 10) versus those of bertweet are close enough, in order for us to choose the one which fits our task better. Since the multitask approach by Aguilar et. al consists of multiple components and BERTweet just of one, BERTweet is chosen in favour of simplicity.

The results of the global pass exhibit already satisfying recall rates of around 0.73. However, highlighting potentially harmful tweets would be better in order to achieve even a better recall rate. Trading off the precision against the recall slightly to increase the recall rate is permissible since we assume that a human supervisor has to consult the final results manually and go over them. To understand where the model makes the errors, the model's results were analyzed for 50 tweets. The following subsection summarizes the results.

3.6.3 Error analysis

One common source of problems is abbreviations which were also mentioned in (Ritter et al., 2011),



(a) ner-prob

Figure 8: Softmax of model output

are abbreviations. Below are two examples of such. In the first example, *NHS* should have been recognized as an organization. In the second example, *MOC* should have also been recognized as an organization as well. The abbreviations which are correctly identified depend on the pre-training corpus used by BERTweet. However, some minimal typical structure is often shared by those abbreviations. For example, capitalization marks the entities and distinguishes them from the rest of the tweet. Therefore, it may be possible to recognize those abbreviations with the price of less precise results.

> Jeremy Corbyn says Theresa May must come to Parliament to explain how she'll fix the *NHS*' humanitarian crisis' HTTPURL
>
> @USER @USER @USER is my Senator. She is accessible, responsive, sincere in her motives. Her office returns calls. She sent the Red Cross to our 97 yr old great grandmother's house in Puerto Rico to offer assistance and support when other *MOC*'s were taking tours

The softmax probabilities for the second tweets are visualized in 8. Blue represents the Outside tag, orange the Beginning tag, and green the Inside tag.The y-axis corresponds to the probabilities of the respective tag, And the x-axis indicates the index of the Tweet-token. Around the 26-th token, we can see a small spike which indicates that *MOC* could be an entity. Even though it is a small signal around 0.1, it could help us improve our classification results.

Moreover, I classified all the mistakes made by the model into three categories. The result are shown in the table 11. The categories are: *MISSED*, *PARTIAL*, *OVERREACH*. *MISSED* shows entities which were not found by the model at all. *PARTIAL* corresponds to all the cases where only part of the entity was marked. And lastly, *OVERREACH* marks the cases where more than the necessary region was marked. Out of those three categories, the *MISSED* category is the most crucial to improve. In the other ones, at least the entity itself was found even though with alternating segmentation.

3.6.4 Possible enhancements

different metric It is worth noting that the Outside labels using the IOB2 scheme are the most common, with a relative occurrence of over 90%.

That imbalance of class labels causes the model to be very well versed in recognized tokens that don't belong to an entity. However, as we want to increase the recall rate, it may be worth it to move the focus of the model from the Outside tags toward the Inside and Outside labels corresponding to the entity tokens.

One way to achieve that is to pick as the loss function an asymmetric cross-entropy loss function where the Outside label is attached with a smaller weight to limit its influence on the final prediction.

Another way to deal with this issue is to adjust the threshold using a different metric.

Another perspective would be to adjust the model by fine-tuning on a differently labelled tweet data. Choosing another, simpler scheme than IOB2 could achieve that.

The results of each approach are shown on table 12. For consistency, the scores using the Entity Detection (ED) approach are used.

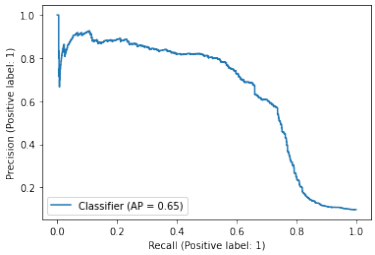
Initially, the scores show only results with an exact boundary match for the entity in question. For example, suppose an entity *international committee of the red cross* is annotated as Organization, and the model only predicts *red cross* as an entity. In that case, this won't be counted as a correct prediction by the evaluation procedure.

This strict evaluation rule makes it difficult to evaluate different improvements since exact matches of the segmentation boundaries are often hard to achieve. To work around this problem, we introduce an alternative metric that measures if a token in question which is of an entity also got predicted as part of the entity. This can be viewed as a binary-classification problem.

	precision	**recall**	**f1measure**
Asym Loss 1	0.42	0.85	0.56
Asym Loss 2	0.53	0.76	0.62
Diff metric	0.54	0.78	0.64
Diff metric -norm	0.57	0.86	0.68
Binary model	0.48	0.85	0.62

Table 12: Various adjustments

different loss The model uses the standard cross-entropy loss function for the training — where the Outside label gets assigned the smallest weight. In the case of the first row, a weight of 0.01 was chosen, while the second row corresponds to the weight of 0.001. In the first case, the other label's



(a) ner-thresh

Figure 9: Various thresholds for ther ner model displayed as precision-recall curve

weights are 0.99 and 0.999 perspectively. Comparing the resulting performance to the results of BERTweet with hashtags and BERTtweet, the recall is higher for both, while the precision is a bit lower for the first and more so for the second row. In conclusion, adjusting the weight for the Cross entropy loss function can benefit the recall.

different threshold Alternatively, the evaluation threshold can be adjusted to trade off between recall and precision. Precision-recall curves are utilized to find a suitable threshold. The effect of different thresholds is visualized on the figure 9. From this figure, a threshold leading to a recall above 0.7 percent while still remaining a precision of above 0.6 is used. The precision-recall curve shown was created using normalization beforehand.

Model adjustment Another approach is to adjust the BERTweet model itself. I used a simpler scheme instead of IOB, which has two labels, ENT and NONE, which correspond to an entity token and non-entity token in the other case. Naturally, the evaluation metric corresponds to the notation of how many tokens are covered and permits partial entities in this regard. The results are shown in the row labelled *Binary Model*.

manual enhancements The preprocessing done by the notebook of (Bhowmick et al., 2022) utilizes some normalization of tweets before running the model on them. To evaluate their effects, I considered a notebook version without this normalization procedure. Apostrophes aren't filtered out, stop words remain, etc. The result shows that the normalization step decreases the performance slightly. The lower score could be because stopwords and punctuation of tweets have indicative power for entities and have a meaningful role in the structure of the tweets. The approach with different evaluation metric with and without normalization is shown on the row*Diff metric -norm*.

	precision	**recall**	**f1**
ED	0.40	0.63	0.49

Table 13: Entity Recognition on french dataset

All the results show the evaluation of the local pass only. Since there is a second pass which can find entities which were missed in the first step, the scores are likely to be higher. But to understand the performance of the entity recognition model itself, it is useful to analyse the results based only on the first pass. From all the results, *Asym Loss 1* and *Diff metric -norm* seem to be the most promising. Therefore, this choice was also selected to be used for the next part of the pipeline.

3.6.5 Multilingual

So far, we have only considered English tweets. As we have seen on the tweet language distribution plots in section 3.3, this assumption doesn't hold considering the ICRC. Russian and Arabic are also very common and don't use the Latin alphabet. The question now arises of how we can adapt our methods we employed so far to different languages.

One option would be to pretrain BERTweet in other languages. Pretraining BERTweet from the start was out of scope since the required resources and training time are too high. The BERTweet model took around four weeks to train on 80 gigabytes of English tweets. (Nguyen et al., 2020)

Another option is to use a multilingual Roberta model directly. However, as we have seen from the results of entity recognition on tweets using standard methods, this method won't even perform in the English language well enough.

To deal with the issue of tweets composed in other languages, we will translate them to English before we perform the existing Entity Recognition method. After performing the Entity Recognition, the entities will be translated back to the original language. We are following the approach of (Jain et al., 2019) to achieve that.

Their approach is divided into two steps. In the first step, potential matches are generated in the target language (French in this case). The source entities are translated using the google translate API. The source entities are also considered valid potential matches in the target language. This is done for the cases where the entity is a name that doesn't change across languages. In addition, dictionary-based translations are used as well. In (Jain et al.,

2019), the best match is chosen out of all the potential matches.

In the second step, the best matches are selected. We omit this state for a high recall rate and treat all potential matches as valid.

After the second step, some source entities may not be matched with an entity span in the target tweet. Some entities are not matched because word-by-word-based translations can be different from correct translations because the context of multi-token spans may not be considered. To remedy this issue, consistency between tweets in the whole collection is exploited, similarly to how multiple tweets are used to generate better results in (Bhowmick et al., 2022).

Testing this approach on annotated French tweets leads to the results presented in 13. A recall rate of 0.63 was achieved. There is quite a bit of a performance gap, and further work must be done to improve that. This task will be left to future work.

3.7 Directed Sentiment Analayis

So far, we have established how entity recognition can be carried out on tweets. The reason for this is that we want to be able to identify all entities which are being addressed in a tweet, possibly with negative sentiment. Sentiment analysis alone is insufficient since it detects if a tweet is negative overall. Especially in the ICRC scenario, most tweets are pessimistic because they address topics like war and human rights. Overall, over 80% of the tweets are negative from the first dataset. This project aims to provide the non-profit organization with a filtered list of possibly harmful tweets. Simply listing all negative tweets makes it impossible to have a look at each entry.

To tackle this problem, we use a similar approach as in (Park et al., 2021). In their work, the issue of directed sentiment analysis is approached using an explicit Question-Answering (QA) model utilizing Transformer models. In the work of (Park et al., 2021), they acquired annotated sentences from the Real News Corpus and articles from various news media. The annotators had to write down entity pairs and specify whether the source entity has a negative, neutral, or positive tone towards the target entity. One caveat is that the dataset consists only of the English language and doesn't contain microblog messages like tweets. The model is fine-tuned using three types of questions. The first question asks if two entities are neutral, whether one entity is negative towards the second, and if one entity is positive towards the other.

The first simplification is to disregard source entities. We don't care about the sentiment-relationship between two entities but rather about what sentiment each entity is addressed by. Or in other words, we are only interested in relationships in one direction rather than both. This decision is motivated by the fact that most tweets do not have an explicit source entity. Often instead, the author poses the source entity. Therefore, most of the time, there are target entities but no source entities. We want to know which entities are being addressed in a negative tone and belong to the ICRC.

The second simplification is to focus only on two possible classifications, being either negative or non-negative. This simplification is also motivated by the fact that positive sentiment is never problematic because it is naturally non-threatening.

Since our problem is a simplified version of the model used by (Park et al., 2021) and our annotated data follows the structure of the simplified problem, we have two possibilities two bridge those differences. One solution is two transfor the annotated data follwoing the simplified problem statement into data matching the orignial model. An artificial entitie is introduced at the start of each tweet where this entity poses the soruce entity. The structure would look as following "[NAME]: 'ACTUAL TWEET'". *NAME* can be replaced by any name. An alternative to that approach is to train a simplified model by transforming the training dataset from the original model by simply disregarding the source entities. Both attempts will be visualized in the results section (3.7.3).

3.7.1 Directed sentiment on English tweets

To tackle this problem, I hope that the effects of transform learning make it possible to use the model trained on the English dataset on English tweets. Further, I am hypothesizing that it will also work for other languages when we fine-tune a multilingual model like Roberta with the English dataset and still get acceptable results. To be able to measure how well the knowledge transfer works both scenarios are evaluated. First, I will create an annotated list of tweets highlighting their entities and negative sentiment. Since annotating is quite time-consuming, I rely on the size of 50 tweets. In the future, I hope to be able to work with a more extensive set of annotated tweets. Moreover, I will annotate 50 tweets in the French language to see

how well the method performs in another language than it was fine-tuned.

Since we want to know which entities are being addressed in a negative tone and belong to the ICRC, it is reasonable to preselect some tweets for annotation which meet certain criteria. One possibility is to use established sentiment analysis to arrive at a selection of tweets that have a negative sentiment. However, this has the issue that most tweets relating to the ICRC are intrinsically negative. Not only are we interested in tweets with negative sentiment, but also in tweets where the expressed attitude is subjective. For example, we are interested in tweets that exhibit signs of anger or sadness. Luckily, there exist models tailored to this objective of detecting emotion rather than just negative and positive sentiment. One of such models is the *cardiffnlp/twitter-roberta-base-emotion* available on Huggingface, which is capable of distinguishing between the emotions joy, optimism, anger, and sadness. Next to this model, *cardiffnlp/twitter-roberta-base-offensive*, *cardiffnlp/twitter-roberta-base-irony*, and *cardiffnlp/twitter-roberta-base-hate* seem also suitable for this problem. The hate speech model finds only five tweets containing hate speech, which is not enough for an evaluation dataset. Therefore, this model was not used in this part of the project.

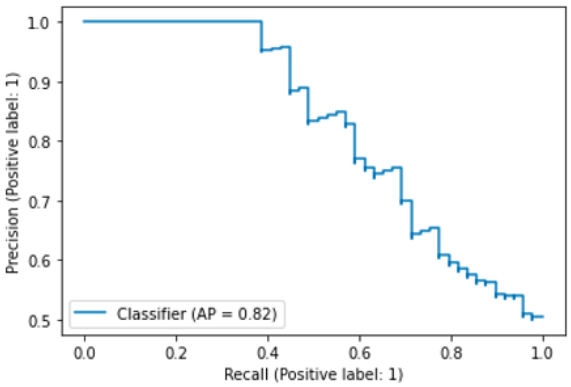
3.7.2 Directed sentiment on French tweets

There are two ways to deal with tweets composed in languages other than French. The first approach is to use a multilingual model instead of an English-only transformer model.

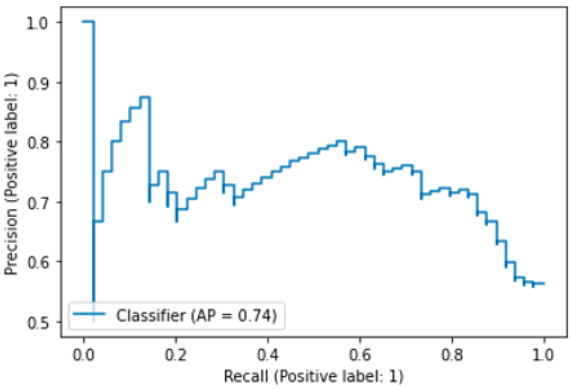
Another idea to tackle directed sentiment in languages other than English would be to translate them to English beforehand. This approach can be seen as an implicit case of transfer learning. Translation has significantly benefited from the recent novelties in machine learning architectures. It is therefore likely that this approach will also lead to practical results. To conclude which of those two approaches is superior, both will be evaluated and discussed.

3.7.3 Results of directed sentiment analysis

The precision recall cures are present in figure 10, 11, and 12. In the first figure, the English model as presented in the paper where our input is transformed in a format suitable and the adjusted model to the restricted input of our problem. It is apparent, that the original model performs best. This is

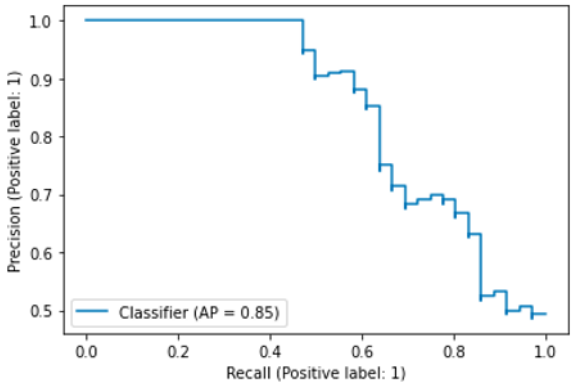


(a) adjusted model

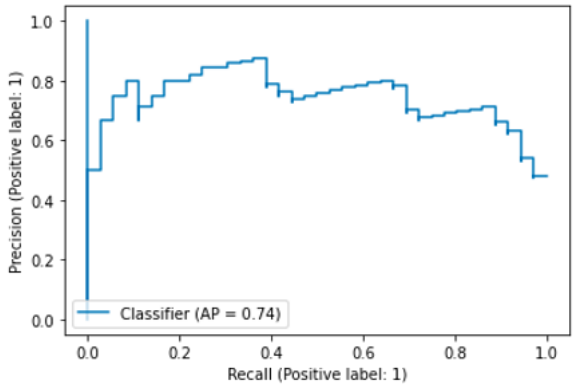


(b) original model

Figure 10: Directed sentiment for English language

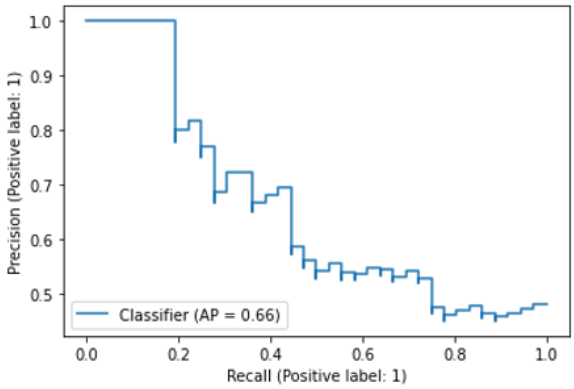


(a) adjusted model

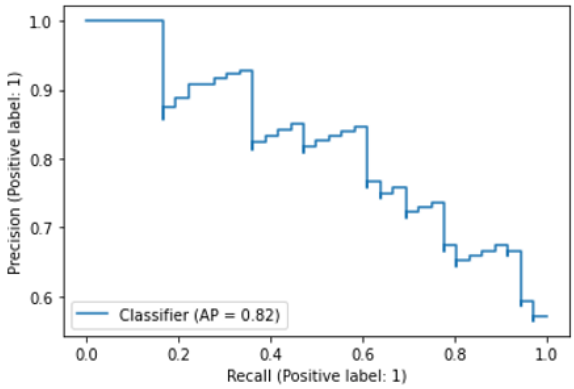


(b) original model

Figure 11: Directed sentiment for French using translated tweets



(a) adjusted model



(b) original model

Figure 12: Directed sentiment for French using multilingual model

possible due to the fact that it is able to incorporate more knowledge since the full data is used and not further simplified.

The second figure presents the results when french tweets are translated beforehand. And the third figure shows the results on french tweets when a multilingual model is used. It is apparent that translating french tweets to English beforehand performs much better than using a multilingual model.

Finding suitable tweets to annotate is an additional challenge. The selection of tweets should contain tweets where entities are being addressed negatively and tweets where no entity receives a negative sentiment to evaluate the precision of the model. The balance of those two cases is important since we want a high recall rate, but the number of tweets an auditor must go through should be as small as possible. The annotated dataset in French therefore contains both, entities which are addressed as negative and entities which don't have any negative sentiment associated with them.

3.8 Final results

In the final step, we combined the results from Directed Sentiment analysis on tweets and Entity Recognition on tweets to create one holistic pipeline.

For the final evaluation, I relied on a dataset from the ICRC, annotated by another student in our lab. The data is based on ICRC-related tweets during the Ukrainian crisis from January until March. All of the tweets were classified as hate speech. Each tweet was associated with a topic after a topic analysis was carried out on the whole collection. I annotated 30 tweets with the respective entity, which is addressed negatively. Usually, this entity is *ICRC*.

The final recall rate has a high result of 0.96 and a precision of almost 1, when the threshold is chosen appropriately. A threshold can be chosen from the precision-recall curves to achieve optimal results fulfilling precision/recall requirements. This number is not necessarily representative since the entity annotated was mostly "Red Cross". Since one correctly 'Red Cross' entity automatically leads to all others being correctly identified by the global pass of the entity recognition method, this result is somewhat biased. Therefore, this final result doesn't sufficiently evaluate the correctness of the entity recognition part. However, we sufficiently surveyed this part's correctness in the section about Entity Recognition 3.6. The final results primarily represent an evaluation of the directed sentiment using a dataset correctly capturing the use-case of our pipeline. Since the annotated data was quite sparse with only around 50 tweets, the final evaluate is not representative and needs further investigation.But from the high precision it is evident, that the directed sentiment part can distinguish between the classes *NEGATIVE* and *NON-NEGATIVE* quite well.

Since the results should be understandable and searchable by a human, visualization is important and useful. There are some figures shown which highlight certain aspects of the data. The first figure 13 shows that topic analysis can be carried out based on entities rather than the full tweets themselves. This maybe helpful to reduce noise in the topic model.

Figure 14 shows the distribution of entities in the whole corpus. Figure 15 shows the distribution of entities for one topic provided by the data itself.

The three figures so far focus on the entities and visualize their distribution. As our task states, we are mostly interested in the negative sentiment towards those entities. Therefore, it is also important to incorporate that aspect into the visualizations. For the 10 most common tweets, their average confidence of them being referred to negatively is shown in the first figure of 17. More interestingly, a heatmap (shown in the second figure of 17) is used to visualize the relation between the entities, topics and their negative attribution. the most prominent entities are shown on the y-axis, whereas the identifiers of the provided topics are shown on the x-axis. The warmer color the higher the negative sentiment towards an entity inside a specific topic. For example, we can see that the entity "the red cross" is appearing to be mentioned extensively negatively in the topics 2, 203, and 174. The heatmap also allows to find topics where multiple entities appear to be referred to negatively. For example, in the topics 104 and 217, both entities "russia" and "the red cross" seem to receive negative sentiment.

Having a closer look at the topic 217, we can come across the following tweet:

> @USER @USER NO FUCKING WAY! RED CROSS is into child sex trafficking out of UKRAINE. Russia is not attacking the citizens but only the military and leader and wiping out the US taxpayer-funded bioweapon labs and testing facili-

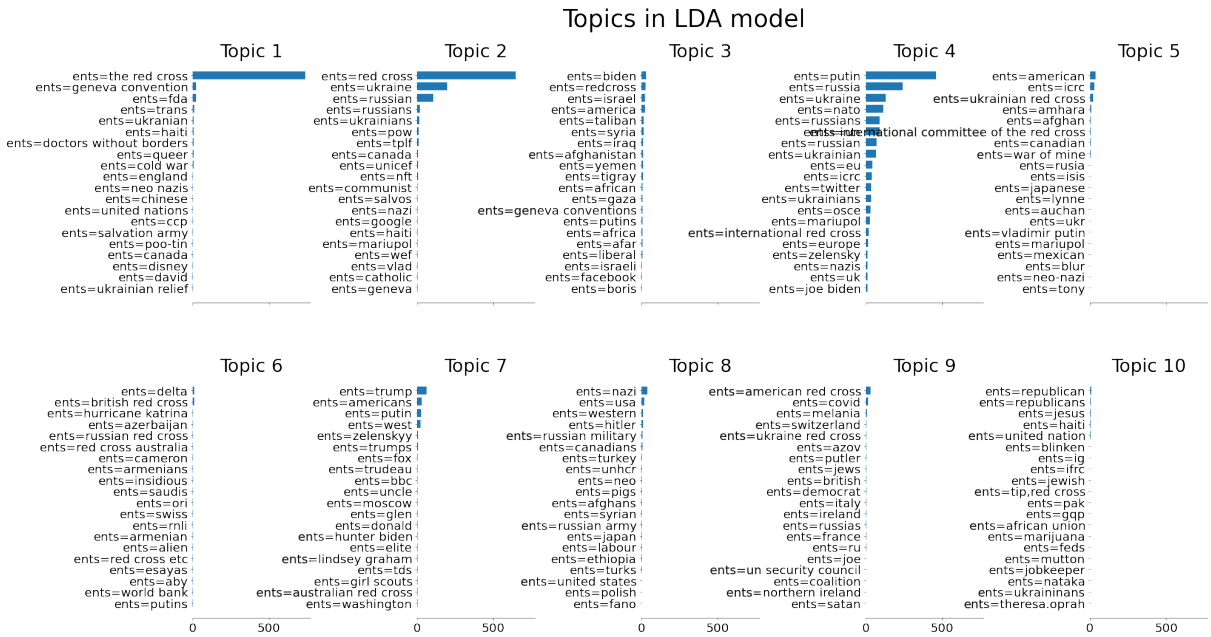


Figure 13: Entity distribution in most relevant topics

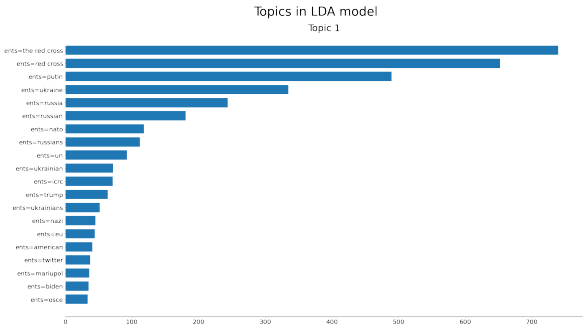


Figure 14: Average confidence in negative sentiment towards most common entities

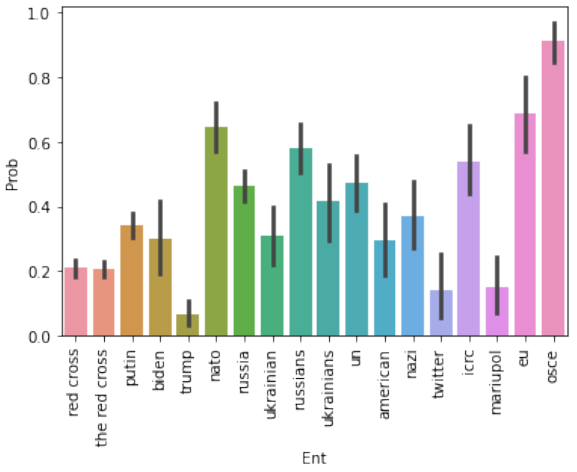


Figure 16: Average probability of negative sentiment for the 10 most prominent entities

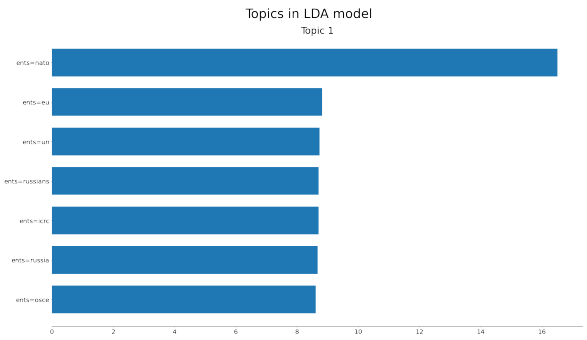


Figure 15: Entity distribution for an individual topic

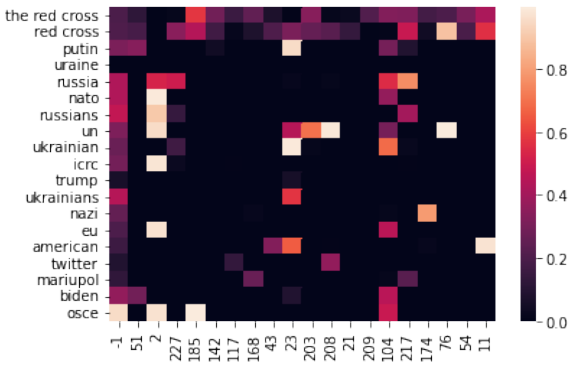


Figure 17: Heatmap

ties. you are a deep state and communist for not believing this.

Such tweets indicate possible disinformation and could be a good starting point for further research where exactly this information comes from and which users are participating in spreading it. Other tweets in this topic generally include tweets about the humanitarian corridors in Ukraine, and how the ICRC and Russia affect it.

4 Conclusion

Critique A translation approach was used to tackle the problem of multilingual Entity Recognition in Tweets. For the translation part, the Google translation API was utilized. One issue with this is the cost of the API usage and the proprietary and closed source nature of the translation service offered by Google. Especially in a case like the ICRC, transparency is a critical requirement to understand and adjust possible biases. For this task, Huggingface translation models could be used. One drawback, of course, are the less powerful translation capabilities.

A significant issue of the method in this paper is that it mainly focuses on languages with the Latin character set, including languages like German, English, and French. Other languages, Arab, Turkish, and Russian, are neglected. The performance of those languages might differ quite a bit since their structure is further away from Latin-based languages.

One of the major shortcomings of this paper is the preliminary evaluations of the methods for other languages. Annotating is a laborious process, and finding people willing to annotate is hard. An approach to tackle this would be using a crowdsourcing service like AmazonTurk to distribute the annotation workload effectively. However, setting up a meaningful procedure to generate valuable annotations is the work for a future project.

So far, BERTweet has been used without additional model components. Since the finetuned BERTweet model outputs softmax probabilities, adding a Conditional Random Field (CRF) model could be beneficial to extract further structure out of the probabilities rather than just using argmax function to determine the entity for each token individually. Conditional Random Fields are also used as the final layer in (Gustavo Aguilar, 2017).

In this work, the focus has been put on increasing the recall. In doing so, the precision suffered slightly. It may be possible to alleviate that by having a second pass over the potential entities and filtering out false positives by various means. For example, words that belong to a grammatical category like verbs could be excluded. Existing databases on the web capturing the grammatical categories of the phrase could be consulted.

Future Ideas Bad grammar is often a problem. Different slangs are often mixed. Highly irregular grammatical structure makes it especially difficult for machine learning. Furthermore, there can also be many temporal and local changes in the grammar and way tweets are composed. This time-drift is also mentioned in (Bhowmick et al., 2022). Moreover, tweets are composed by many non-native English speakers. The many different language backgrounds also influence the average English grammar of tweets.

Another interesting issue that often appears is the phenomenon of code-switching. A lot of Spanish tweets contain partial English fragments for example. (Begum et al., 2016) provides a good analysis of code-switching occurrences in tweets. Implementing the recognition of code-switching could help improve our pipeline further.

A different approach not followed by this paper is adapting the BERTweet model to other languages. Pretraining BERTweet on other languages could be the key to high performance using other languages. Maybe it is not even necessary to have a model for each language but one model for each cluster of highly similar languages. E.g. Italian and French could belong to one such cluster.

Also, an idea for possible improvement is to extend the concept of (Bhowmick et al., 2022) further. Instead of just looking at the repetition of known entities across multiple tweets, one could perform a more sophisticated analysis of entities' occurrence in a time stream and create an even more global approach. Especially considering tweets, I wonder how one can use additional context found in user-profiles and context shared by tweets utilizing similar subsets of hashtags and mentions. Taking further context into account could allow for an even more complete global picture.

Our solution provides a data structure that is query-able by a human worker. As not all users are assumed to be highly technically skilled, suitable visualization needs to be created to provide the end-users with enough visual feedback on what they need to have a closer look at. So far, heatmaps

have given the users a visual sense of what could be of importance. While heatmaps are interesting and already convey a lot of information, a more complex graph-based visualization may be of benefit. In particular, graph-based databases with a query language like GraphQL could be interesting for this task.

Summary In this project, a pipeline was created to detect disinformation campaigns on Twitter. Firstly, an overview of various techniques for natural language presentation was given in the context of Twitter messages. Standard methods of entity recognition on tweets have been evaluated. The most recent advances consisting of the work of (Gustavo Aguilar, 2017) and (Bhowmick et al., 2022) were chosen as a starting point for this project. The problem was decomposed into two subparts. As a first step, a reasonable, effective entity recognition model for tweets was implemented on the basis of the work of (Bhowmick et al., 2022). As a second step, the detection of directed sentiment was realized using the entity recognition mechanism developed in the first part. When the whole pipeline is employed, a stream of tweets can be filtered to arrive at the selection of tweets with negative sentiment towards an arbitrary entity. The final output of the pipeline consists of a data table that is easily understood and can be queried for different information by the end-user. In the last step, a visualization presentation is given using heatmaps. All the pipeline steps have been evaluated and adjusted for higher recall. The results for English tweets are satisfactory, but the results for other languages still have space for improvement. Finally, various ideas were given to improve the results further in the future. I want to thank especially my supervisor Rebekah Overdorf from EPFL, who helped me figuring out the concrete steps of this project and also helped me with any questions I had. I also want to thank my supervisor Ryan Cotterell from ETHZ for taking my project and helping me with any other questions I had.

References

Marco Bastos and Johan Farkas. 2019. "donald trump is my president!": The internet research agency propaganda machine. *Social Media + Society*, 5(3):2056305119865466.

Rafiya Begum, Kalika Bali, Monojit Choudhury, Koustav Rudra, and Niloy Ganguly. 2016. Functions of code-switching in tweets: An annotation framework and some initial experiments. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1644–1650, Portorož, Slovenia. European Language Resources Association (ELRA).

Satadisha Saha Bhowmick, Eduard C. Dragut, and Weiyi Meng. 2022. Boosting entity mention detection for targetted twitter streams with global contextual embeddings.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Chris Ding, Xiaofeng He, and Horst D. Simon. *On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering*, pages 606–610.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan. Association for Computational Linguistics.

A. Pastor López Monroy Thamar Solorio Gustavo Aguilar, Suraj Maharjan. 2017. A multi-task approach for named entity recognition on social media data. Proceedings of 3rd Workshop on Noisy User-generated Text, WNUT 2017. Ranked 1st place in the two evaluation metrics.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. More than a feeling: Accuracy and application of sentiment analysis.

ICRC. False allegations: Icrc condemns video showing cash in trunks. [Last accessed: 2022-04-23].

Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. 2019. Entity projection via machine translation for cross-lingual ner.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Saif M. Mohammad. 2020. Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Kunwoo Park, Zhufeng Pan, and Jungseock Joo. 2021. Who blames or endorses whom? entity-to-entity directed sentiment extraction in news text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4091–4102, Online. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Swissinfo. Explainer: what can the red cross do and not do in ukraine? [Last accessed: 2022-04-23].

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Joshua Tucker, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *SSRN Electronic Journal*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.