Detecting harmful information against humanitarian organizations

Selim Fekih, Prof. Karl Aberer, Rebekah Overdorf, Rémi Lebret

Semester project - 8ECTS
Distributed Information Systems Laboratory, EPFL, Switzerland

***Abstract*—The aim of this project is to identify tweets which are negative and directed at the Red Cross. We first begin by identifying harmful tweets citing the Red Cross. Then, we group them into different clusters and get the general topic of each one. Using this work, humanitarian analysts can save time by focusing on the clusters directed against the Red Cross.**

I. Introduction

Twitter is a popular social media and while it is helping improve the communication between individuals, it is also a platform that contains a lot of attacks, directed against individuals, countries or intergovernmental organisations, such as humanitarian organisations. Finding and preventing these attacks is crucial as they can cause a lot of harm the reputation of any organisation. This project aims at detecting harmful tweets directed against humanitarian organisations is general, and more specifically the International Committee of the Red Cross (ICRC) and other national Red Cross organisations.
Our dataset consists if tweets citing the ICRC, the Red Cross and the Red Crescent. The tweets have been published between 01/01/2022 and 24/03/2022. We want to detect four types of attacks:

- Disinformation: Propagating false information about the Red Cross.
- Misleading: Giving a wrong impression about the Red Cross.
- Troll Attacks: General hateful tweets directed against the Red Cross.
- True, but of interest: Tweets citing true scandals the Red Cross was involved into.

Much work has been done on sentiment analysis related tasks on twitter datasets. However, no work has been done to prevent twitter based attacks against humanitarian organisations in general and the ICRC more specifically.
To resolve this problem, we process in two steps. Firstly, we begin by filtering out possibly harmful tweets, whether or not they are directed against the Red Cross. We use pre-trained models to analyze sentiments of each tweets and then keep relevant ones. Secondly, we apply clustering and topic modelling techniques to return predefined clusters and their topics.

II. Methods

In our work and to be able to analyze and understand results, we focus on tweets of three languages: English, French and Arabic.
We begin by reporting the proportion of data present in the dataset for each language.

	English	French	Arab
Number of tweets	194.024	18.625	21.098
Proportion of the dataset	83%	8%	9%

TABLE I: Table showing the proportion of overall tweets for each language for each language

Due to the data imbalanceness and the fact that the tweet content is different across countries and languages, we choose to make independent pipelines for each language.
We illustrate below the general workflow of the system implemented.

Algorithm 1 General work flow

```
for each language do
    Compute negative score for each tweet.
    Select the 5% tweets with highest negative scores.
    Group Selected tweets into clusters.
    for each cluster do
        Get topics
    end for
end for
```

A. Tweets preprocessing

We apply a different preprcessing to tweets for each part of the work

1) Harmful tweets filtering: section II.B: For this task, we do the following preprocessing steps, following the preprocessing directives given by models documentation:

- remove 'http' hyperlinks.
- changing each tagged account by '@user'.

2) Clustering and topic modelling: section II.C: The two parts share many preprocessing steps but contain one difference, which is the way to preprocess the tagged users. WE inspire ourselves from the Coding Club blog [5]

a) Clustering: section II.C.1:

- Remove all sorts of hyperlinks ('http', 'bitly', '[link]').
- Remove tags.
- Remove stop words and lower tweets.
- When the language is English, stem and lemmatize tweets.

b) Topic modelling: section II.C.2:

- Remove all sorts of hyperlinks (‘http‘, ‘bitly‘, ‘[link]‘).
- Remove tagged users but without the ‘@‘.
- Remove stop words and lower tweets.
- When the language is English, stem and lemmatize tweets.

B. Harmful tweets filtering

After that, we proceed in two steps. We first compute the overall negative score for tweets then do the filtering with respect to the score.

1) Negative scores computation: To filter out potentially harmful tweets, we use finetuned sentiment models on tweets datasets. For this task, we use models from ‘HuggingFace‘.

a) English and French Languages: We get the sentiment scores from two models that yield an anger score[6] and an offensiveness score[8]. Both models consist of a roBERTa-base model which is then finetuned on the specified tasks[1].
We report a correlation score between the two models of 0.62 for English and of 0.53 for French. We also get the following values for the final scores.

	English	French
anger Model	0.3	0.20
Offensiveness Model	0.14	0.11

TABLE II: Mean scores for each sentiment for each language

In order to give the same weight to both models, we begin by normalizing scores:

$$final\_score = \frac{score - mean\_score}{std\_score}$$

Then, we get define the final negative score as the mean of normalized scores.

b) Arabic language: We proceed differently for the Arabic language. We use an Arabic specific
The Fig. 1 below shows the general workflow for the tweets sentiments modelling.

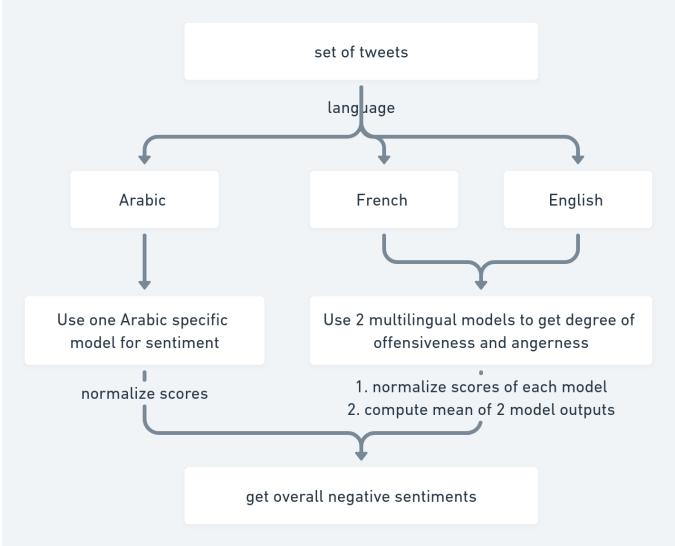

Fig. 1: final workflow for sentiments modelling

2) tweets filtering: After getting the overall negative sentiments, we filter out for each language the tweets containing the 5% highest numbers. Therefore, we have the following number of filtered tweets for each language:

	English	French	Arab
Number of tweets	194.024	18.625	21.098
Number of filtered tweets	9.701	931	1054

TABLE III: Table showing the number of kept tweets for each language

C. Clustering and topic modelling

1) Clustering: One important specification of the clustering is that we cannot impose a predefined number of cluster. We tested two clustering algorithms, the Louvain clustering algorithm and the HDBscan algorithm.

a) Louvain clustering: We show below the work flow to get clusters using the Louvain algorithm[2].

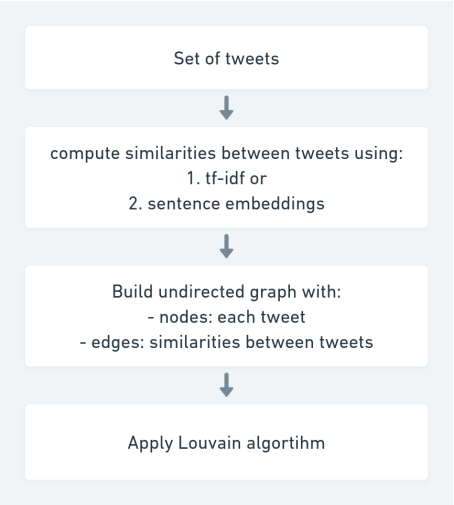

Fig. 2: workflow for louvain clustering

After manual inspection of results, we noticed that the clustering did not help filter out harmful tweets. Moreover, the complexity of the model was too high and could not be modularized for future work (in cases where the number of tweets we want to keep is higher). We therefore implemented the HDBscan algorithm.

b) HDBscan clustering: We show below the work flow to get clusters using the HDBscan algorithm. The workflow is inspired from the Toward Data Science blog [4].

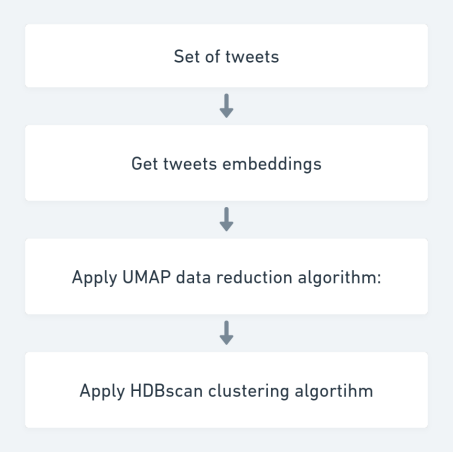

Fig. 3: workflow for HDBscan clustering

We manually finetune the hyperparameters and used metrics for the UMAP and the HDBscan algorithms.
We show below the kept hyperparameters for the UMAP algorithm.

hyperparameter	n_neighbors	n_components	metric
value	10	8	cosine

TABLE IV: hyperparameters chosen for the UMAP algorithm

Moreover, we show below the kept hyperparameters for the HDBscan algorithm.

hyperparameter	min_cluster_size	cluster_selection_method
value	5	eom (Excess of Mass)

TABLE V: hyperparameters chosen for the HDBscan algorithm

2) Topic modelling: For this task, we use the LDA algorithm after vectorizing the corpus using tf-idf. The workflow is inspired by the Towards Data Science blog[3].

III. Results

A. General results of modellings

	English	French	Arabic
Number of Clusters	275	29	32
Proportion of unclustered tweets	45%	18%	30%
Mean cluster size	19	26	23
Median cluster size	13	13	17

TABLE VI: Cluster sizes with HDBscan clustering

First of all, we see that the HDBscan algorithm does not cluster all entries. This is due to the fact that the dataset contains harmful tweets but which are out of context. Many of these tweets consist of:

- General Troll Attacks: Example: 'Fck @RedCross they are so corrupt http' (tweet_id=1500115567772065792)
- Negative Tweets out of context: Example: 'Dorian is really hot until he opens his stupid mouth and starts telling you that the red cross is run by vampires http' (tweet_id=1501733862505955328)
- Troll Attacks not directed against the Red Cross: Example: 'Just donated. Fuck Putin and his cronies. http ' (tweet_id=1497229342807625728).

The proportion of unclustered tweets differs with the language, most of the unclustered tweets are for the English language (45%). This can be explained by the high number of English tweets. There is more variability in tweets and is it therefore more difficult to get fixed clusters. Moreover, the number of clusters returned by the algorithm are different. We see a high difference between the number of clusters returned for the English language (275) compared to other languages (29 for French and 32 for Arabic). This can be explained by the fact that the total number of English tweets is much higher than other languages. This is verified by the fact that the median and mean number of cluster is similar for all languages.
After that, we visualize the cluster lengths distributions for the three languages:

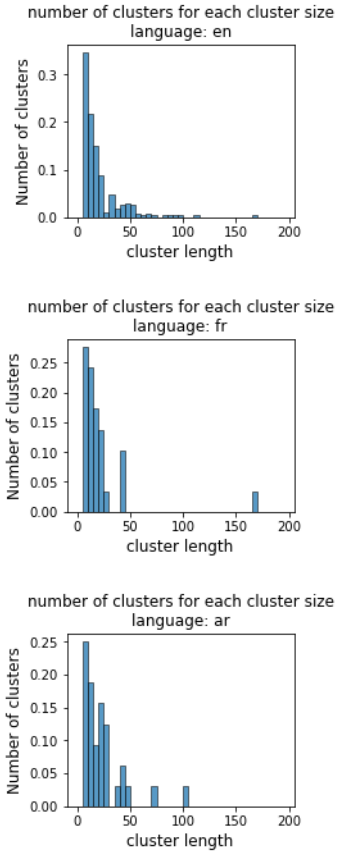

Fig. 4: Clusters lengths (using HDBscan) visualization

Comparing the three histograms plot above, we see that the clusters' distribution is the same for all languages. More precisely, the majority of clusters have a low length, and the more the cluster length augments, the less number of clusters. After that, we get a general overview of attacks, some of them being directed against the Red Cross and some not.

B. Attacks against the Red Cross in general

We manually go through clusters and filter out tweets not directed against the ICRC. Examples of these tweets are:

- Against politicians and their relatives
  - Putin for the war he initiated (English and French clusters).
  - Melania Trump because of the good relationship between Donald Trump and Putin (English clusters).
- Against world organizations: NATO were criticised for not getting involved enough in the war in Ukraine (English and French tweets).
- Against countries: Kuwait and their relationship with terrorist groups (Arabic clusters).

After that, we report the top 3 tweet clusters directed against the Red Cross organization for each language.

	English	French	Arab
# 1 topic	Against American Red Cross: Gay blood donations scandal	True, but of interest against ICRC: Cyberattack the ICRC was victim of.	Troll Attacks against the Red Cross
# 2 topic	Against ICRC: Accused of racism and favoritism for Ukraine over other countries	Disinformation against Red Cross: Not acceding plasma donations of vaccinated people.	Against Red Cross: Disinformation and Misleading attacks in Ukraine war.
# 3 topic	Against ICRC: Accused of racism and favoritism for Ukraine over other countries	Troll & Hate attacks aginst ICRC: accused of hypocrisy.	Troll + Hateful Attacks against Red Cross: Accused of not doing enough in Yemen.

TABLE VII: Top 3 extracted topics for each language (translated to English for French and Arab languages) of attacks against the Red Cross

We see that some topics are repeated across languages (Red Cross being accused of hypocrisy and favoritism as well as Troll attacks). However, some topics are more language specific. For example, there is no mention of the Yemen situation for French and English tweets, on the contrary to Arabic ones.

C. Specific case of the ICRC

We finally treat specifically tweets directed against the ICRC. For this, we only keep tweets that mention the ICRC (whether in the text or as a tag). In this case, many topics are hard to classify as they mention debatable subjects mainly linked to the ICRC and its geopolitical implication in the world. We therefore chose to add one more type of attack: geopolitics.

Language	English	French	Arabic
Total number of tweets mentioning ICRC	39.942	3.889	1.410
Number of clusters containing harmful information against ICRC	15	6	5
Number of ICRC tweets after applying harmful tweets fitering	2.919	519	177
Number of Disinformation based attacks	98	0	106
Number of Misleading based attacks	0	0	0
Number of Troll Attacks	210	327	70
Number of True, but of interest attacks	11	52	0
Number of Geopolitical attacks	152	17	24
Total number of tweet-based attacks	416	396	130

TABLE VIII: Number of extracted attack based tweets against ICRC

First of all, note that the types of attacks for each language don't always sum to the total number of extracted attack-based tweets, as some clusters contain more than one type of attacks. We see that the number of attacks extracted does not reflect on the number of total tweets. For instance, we have a similar number of tweet based attacks against the ICRC for the English and French languages (respectively 416 and 396). On the other hand, we have extracted approximately 3 times less attack based tweets for the Arabic language (130).
In order to have a better insight on the type of attacks, we normalize each type to the total number of extracted tweets for each language and visualize the heatmap of attacks:

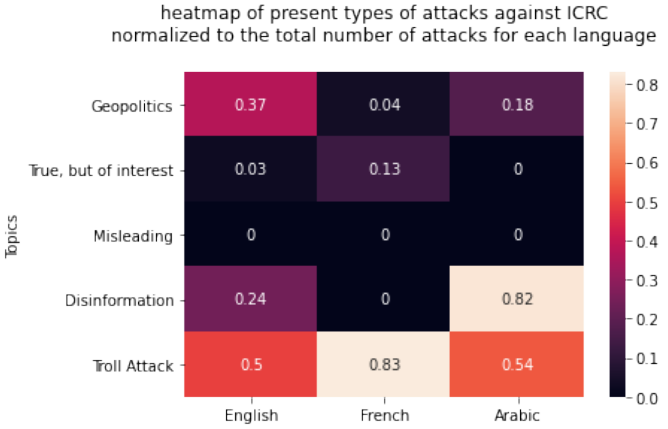

Fig. 5: Heatmap of types of attacks for each language against the ICRC

We first notice that no 'Misleading' based attacks were detected. This is mainly due to the fact that Misleading based tweets are hard to detect as they are based on fact checking and nuances of meaning, not on sentiment analysis of tweets. Moreover, very few 'True but of interest' attacks were extracted.

On the other side, Troll Attacks are the most present for all languages, followed by Disinformation and Geopolitics. The distribution also differs depending on the language. For instance, the ‘Geopolitics‘ topic is more present in the English tweets (37% of the total number of tweet based attacks), while the ‘Disinformation‘ based attacks are more present for the Arabic language (82% of the total number of tweet based attacks).

IV. Discussion

A. Negative results

1) Labeled dataset creation: One option we had was to finetune pre-trained models on another dataset. Therefore, we searched for labeled datasets which are focused on humanitarian organisations.
From around 40 sources, clean disinformation and misinformation datasets are most focused on politics, celebrities news and fact checking from newspapers.

2) Harmful tweets filtering (section II.B):

a) Zero shot classification models test: One option we tested but that was not precise enough was using zero shot classification models[..]. We used two labels: [’relevant’ and ’irrelevant’]. The output of such models depended on the tweets structure more than the content. For instance, tweets which did not contain clear sentences were always classified as irrelevant.

b) Experiments before having the current pipeline: Some experiments did not yield good results. Firstly, we tried to find out Misleading tweet attacks using pretrained irony detection models[7]. However, manual inspection of results showed that the models did work well. Moreover, we worked with an Arabic specific model because the offensiveness and anger detection model were not effective enough for the Arabic language. We can see below the correlation matrices between the anger, offensiveness and irony models for English, French and Arabic.

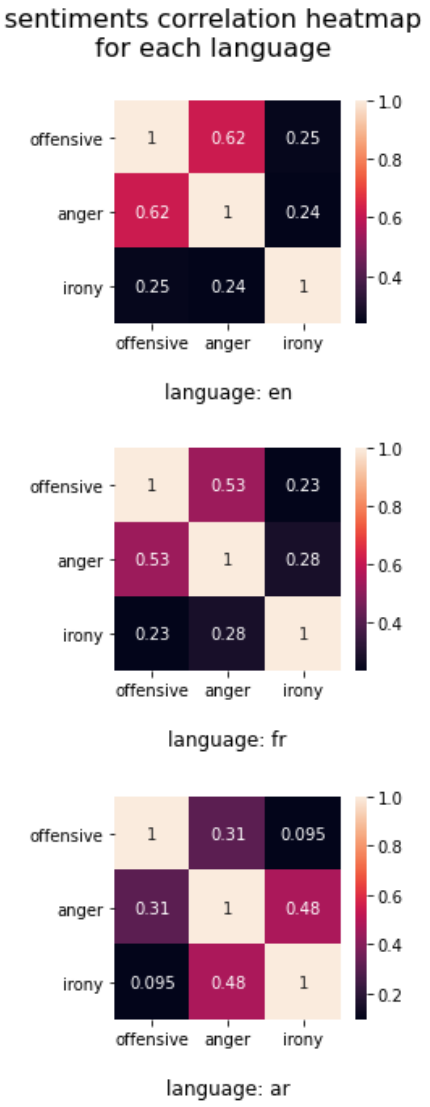

Fig. 6: Correlation heatmaps of sentiment models for different languages

The two assertions cited above are here verified. First, we see that the correlation matrix for the Arabic language is different from the English and French ones. This shows that the output of models is generally not the same. Moreover, the irony sentiments are weakly correlated to the anger and offensiveness sentiments.

B. Challenges

Doing this project, we encountered many challenges. We classified them in four different topics. They are about the lack of annotated data, the problems encountered in the modelling part and the limitations of the task and the project.

- On the lack of annotated data. Because of this:
  - We could not fine-tune bigger models on our specific task and had to use already fine-tuned models.

- We could not create a valid test set. We cannot have any numerical performance of models (other than an overview using samples).

- All the experiments done were validated through manual inspection of outputs. Therefore, the dataset was restricted to languages fluently spoken by the authors: English, French and Arabic.

- Many tweets contain more than one type of attacks. For example, many tweets contain tweets which contain true and of interest information but are coupled to troll attacks or disinformation. Many of these tweets consist of Disinformation coupled to Troll Attacks.

- On the first part of modelling: Harmful tweets filtering

  - It is hard to put a threshold on the number of tweets to filter out and potentially present to analysts. In this project, we chose to keep the 5% tweets with highest negative scores. However, this proportion mainly depends on the human capacity that can be deployed on the project.

  - The current models focus most on hateful, offensive and angry tweets. Therefore, they hardly identify misleading tweets directed against the Red Cross. On this topic, it can be relevant to implement fact-checking programs directed at news articles.

- On the second part of modelling: Clustering and topic modelling

  - The Louvain algorithm has a high complexity and it therefore limits the maximum number of tweet samples we can cluster.

  - The HDBscan algorithm we worked with has some drawbacks. First, it is too precise for our task. In fact, it yields a high number of clusters (as shown in TABLE V), with many clusters having overlap in their topics. For instance, we filtered out many clusters that talk about the Melania Trump and her donation to Ukraine, with each cluster describing a different sentiment (shaming the Red Cross, Anger against the Red Cross and Trump etc.). Secondly, a non negligible proportion of the data was not clustered, because some tweets were too general. This is an important issue to address as it implies a high time loss for analysts. One option to explore on this is to create custom embeddings, where pronouns have more weights. This can help improve clusters creation.

  - The topic modelling systems does not always give a clear understanding of the topic. For instance, tweets which have a high number of swear words are clustered together. However, not all of these tweets are directed against the Red Cross.

- On results generation

  - The number of information based attack tweets given in the section III.A are only a rough estimation. We did not go go through tweets one by one. We chose clusters which have relevant topics (tweets directed against Red Cross) and counted the total number of tweets in them. Some more in depth analysis needs to be made to have more accurate results.

  - It is hard to understand Arabic based tweets as they contain many different dialects.

  - For the section III.C (Specific case of the ICRC), it is hard to differentiate tweets that talk about the ICRC from those who mention national ICRC organisations (example: Americam Red Cross). The keywords we used for filtering are very restrictive. Some work on the keyword filtering, or topic based filtering, can help improve results.

- On the problem of generalisability

  - It is hard to make any further work to accelerate the processing time as there is a high risk of overfitting on the current tweets set. For example, making automatic filtering on the topics of clusters can work on the current dataset (given the current context and war) but would not work well on other future contexts.

  - A direct implication of the point before is that the filtering of tweets cannot be fully automatic and therefore requires more software development. One example to help analysts further gain time is to develop a software that helps them filter out clusters according to specific keywords.

Finally, while we restricted this project to the ICRC, we can easily generalize it to not only all humanitarian organizations but also any organization, public personality.

V. Conclusion

Through this project, we generated a pipeline for filtering out potential harmful tweets and clustering them depending on their topics. We treated tweets of each language differently, from the preprocessing to the negative score generation models. Our methodology helped us filter out a high number of tweet based attacks against the ICRC and other specific Red Cross organisations across the world. We filtered out mostly Troll Attack tweets, and in a smaller proportion, Geopolitics and Disinformation based attacks. However, noo misleading based attack tweets were extracted.

References

[1] Francesco Barbieri et al. "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1644–1650. DOI: 10.18653/v1/2020.findings-emnlp.148. URL: https://aclanthology.org/2020.findings-emnlp.148.

[2] *Louvain's Algorithm for Community Detection in Python, howpublished = https : / / towardsdatascience . com / louvains-algorithm-for-community-detection-in-python-95ff7f675306*.

[3] *Topic Modeling in Python: Latent Dirichlet Allocation (LDA), howpublished = https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0*.

[4] *Topic Modeling with BERT, howpublished = https : // towardsdatascience . com / topic - modeling - with - bert - 779f7db187e6*.

[5] *Topic Modelling in Python, howpublished = https://ourcodingclub.github.io/tutorials/topic-modelling-python/.*

[6] *Twitter-roBERTa-base for Emotion Recognition , howpublished = https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion.*

[7] *Twitter-roBERTa-base for Irony Detection, howpublished = https://huggingface.co/cardiffnlp/twitter-roberta-base-irony.*

[8] *Twitter-roBERTa-base for Offensive Language Identification , howpublished = https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive.*