They're all biased!

Wrocław University of Science and Technology

Katedra Sztucznej Inteligencji

Mikołaj Libura @254550 — Follow

Katarzyna Broczkowska @255728 — Follow

Dawid Jędrzejewski @276941 — Follow

Project Supervisor

Tomasz Kajdanowicz @tomasz.kajdanowicz — Follow

Michał Madej @276948 — Follow

01 Problem Description

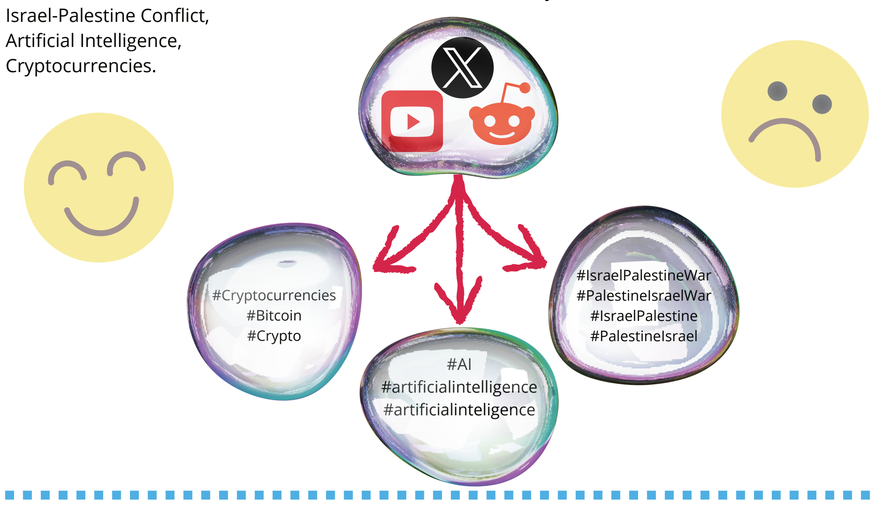
With the importance of opinion mining tasks and the raise in popularity of Artificial Intelligence, there is now an abundance of open source models for "Sentiment Analysis" tasks. The one we end up choosing is often the one that happened to pop up at the relatively high position in our search engine. But how are these models different to one another? Does fine-tune for a specific social media make a model perform worse on other social media? How careful should we be while selecting a model? And how the same slice of reality can be portrayed using different tools?

02 Technologies

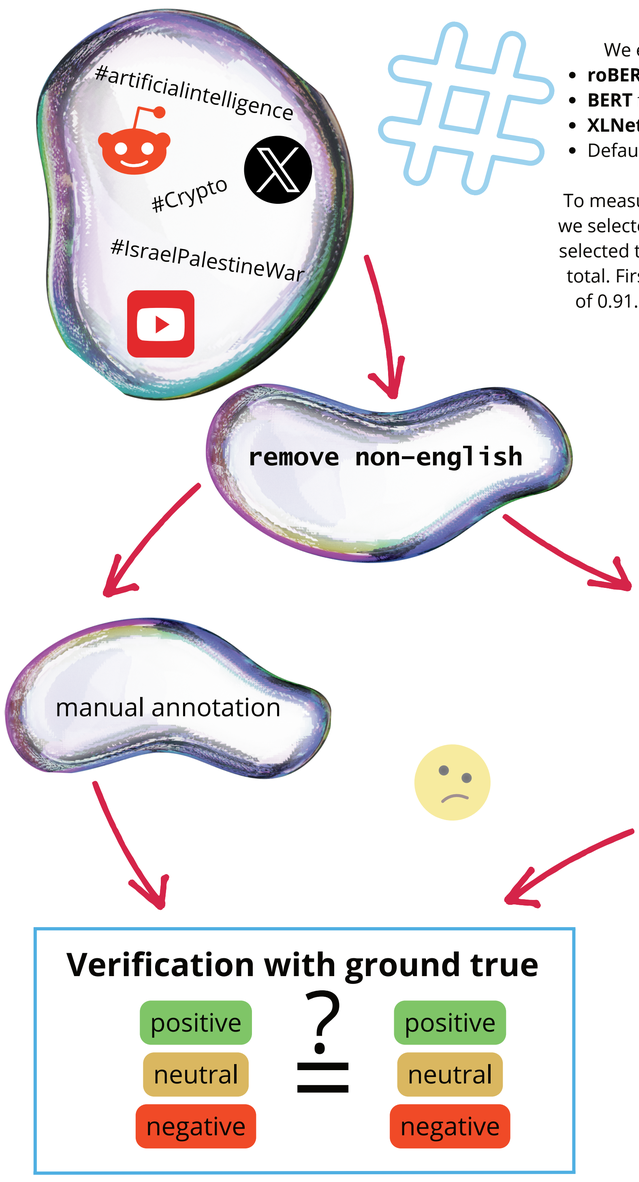
03 Data

Using Reddit API to access Reddit data and Octoparse to access YouTube and X (formerly known as Twitter) data, we downloaded posts and comments related to 3 recently popular topics - each representing a little different slice of reality:

- Israel-Palestine Conflict,
- Artificial Intelligence,
- Cryptocurrencies.



04 Method Description



We experimented with **4** different open source models for Sentiment Analysis:

- **roBERTa** fine-tuned for **Twitter**, cardiffnlp/twitter-roberta-base-sentiment-latest,
- **BERT** fine-tuned for **Hebrew**, avichr/heBERT_sentiment_analysis,
- **XLNet** fine-tuned for **Reddit**, minh21/XLNet-Reddit-Sentiment-Analysis,
- Default **roBERTa** for **English**, j-hartmann/sentiment-roberta-large-english-3-classes.

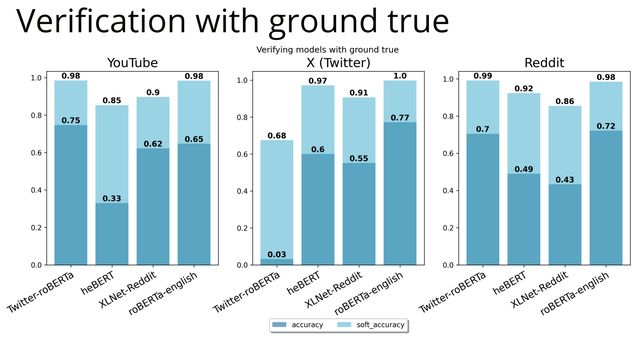
To measure their accuracy, we needed Ground True labels for our data. To achieve that, we selected a subset of our data: 300 posts and 300 comments for each social media in a selected topic (Israel-Palestine conflict) and annotated it manually — **1800** annotations in total. Firstly, we annotated 50 posts and comments in a group of 4, with Cohen's Kappa of 0.91. After that, we split the remaining data evenly and each part was covered by a single annotator, what allowed us to annotate more data in less time.

Then we chose 2 different methods of testing:
- Verification with ground true,
- Consistency of models.

05 Results

To check where our models differ, we proposed a soft-accuracy and soft-Cohen-kappa metric. The intuition is: base metric measures if two annotations are **right**, while our soft-metric measures whether they are **not wrong**. This approach and the results are illustrated below.

Verification with ground true



When it comes to accuracy for the small subset of manually annotated data, it surprisingly turns out, that fine-tuning for a specific social media does not necessarily improve performance. Generally, default english roBERTa seems to be the best regardless of social media.

Standard metric

true sentiment \ predicted sentiment	negative	neutral	positive
negative			
neutral			
positive			

Soft metric

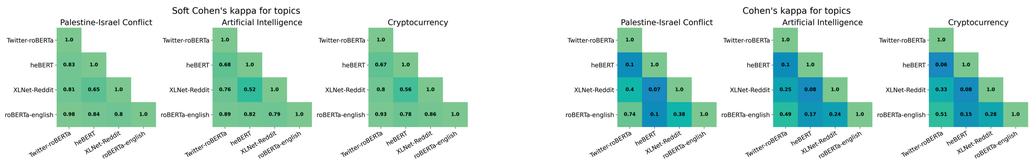
true sentiment \ predicted sentiment	negative	neutral	positive
negative			
neutral			
positive			

Consistency of models

To check the consistency of pair of models we used Cohen's Kappa:

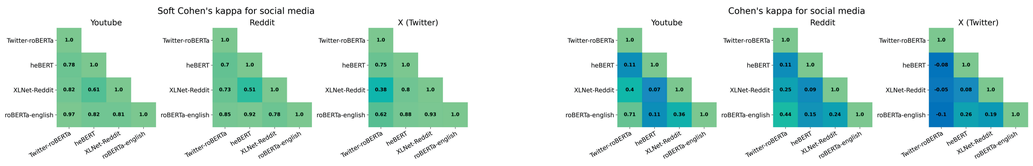
$$\kappa = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)}$$

Consistency by topic



From all the examined models, heBERT seems to be the one that sticks out the most, as it has very low score between all other models regardless of the data. Besides that, while still having quite low score, the Palestine-Israel Conflict has the best results and is the most consistent among all models. The most bias disagreements are for XLNet-Reddit and heBERT for all topics — the one with the worst score is Artificial Intelligence.

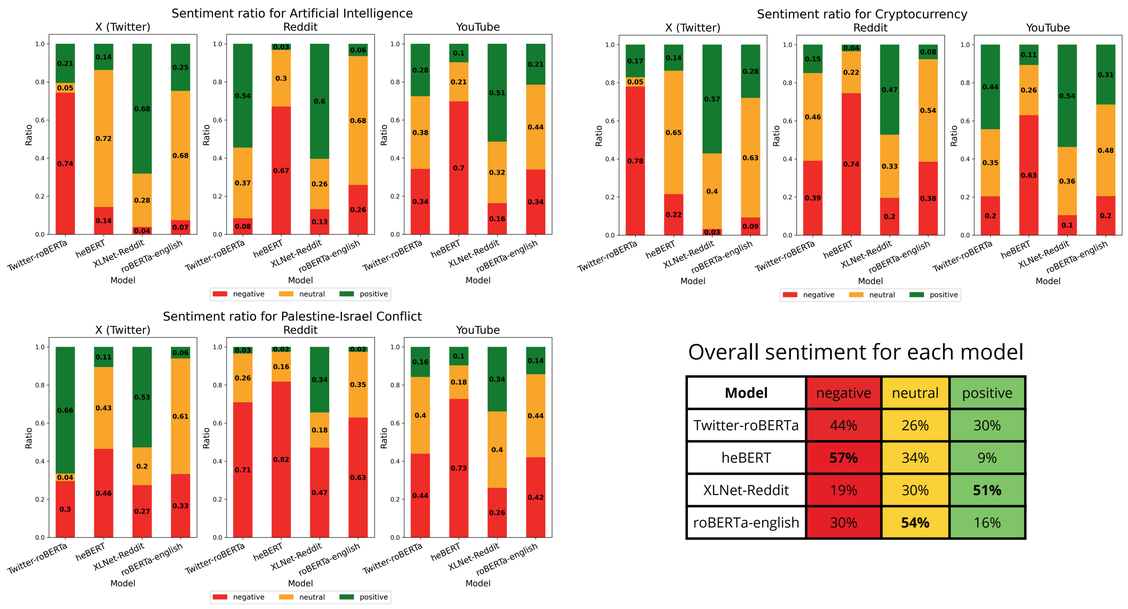
Consistency by social media



For X (Twitter) data twitter-roBERTa differs significantly from the other models. The model that seems to have the most disagreements with other models regardless of social media is heBERT. The highest Cohens kappa can be observed in YouTube data between twitter-roBERTa and roBERTa-english (0.71 normal, 0.97 soft). The two models whose disagreements are the strongest (low soft Cohen's kappa) are heBERT and XLNet-Reddit for Reddit data and XLNet-Reddit and Twitter-roBERTa for X (Twitter) data.

06 Analysis

Sentiment distribution over data for each model.



Overall sentiment for each model

Model	negative	neutral	positive
Twitter-roBERTa	44%	26%	30%
heBERT	**57%**	34%	9%
XLNet-Reddit	19%	30%	**51%**
roBERTa-english	30%	**54%**	16%

Key takeaways:

- HeBERT marks the biggest amount of comments as negative (57%)
- XLNet-Reddit marks the biggest amount of comments as positive (51%)
- RoBERTa-english labels a lot of posts as neutral (54%)
- The biggest difference in overall sentiment is for XLNet-Reddit (51% positives) and heBERT (9% positives)
- For X, for Crypto and AI most models state there's few negative comments, but twitter-roBERTa finds a lot of negativity
- For X, in the topic of war, Twitter-roBERTa states comments are rather positive, contrary with the rest of models
- For X, for Crypto and AI only XLNet-Reddit marks a lot of data as positive (57% and 68%) while other models mark most data as either neutral or negative (Twitter-roBERTa)

post	Twitter-roBERTa	heBERT	XLNet-Reddit	roBERTa_english
Palestine will finally be free from the map 😅	+	-	+	+
Well thank all you people for acknowledging the podcast good for you everybody's overlooking the real crisis what is the solution how can it be fixed stop the bloodshed stop the atrocities stop the suffering	+	-	-	-
ChatGPT can actually make workers perform worse, a new study found	+	≈	-	-
I thought this was gonna be a One Piece joke	≈	+	-	≈

We picked some of the posts on which our models do not agree completely.

The last one (a One Piece joke) is quite interesting since it has all possible annotations.

We also created some synthetic posts to see, how different models can handle them.

The biggest mistakes are often associated with sarcasm.

topic	fabricated post	Twitter-roBERTa	heBERT	XLNet-Reddit	roBERTa_english
Palestine-Israel Conflict	This has to be the best news I've heard all day! Never thought I'd feel so happy about a genocide	+	-	+	+
	Well, Hamas has started this bombing so Palestinians kind of deserve what's happening right now	-	-	+	-
Artificial Intelligence	Wow! Those image generating tools are so great! With just one prompt, you can generate hyperrealistic images like this, and you don't even have to wait hours for some artist to finish it. It's astonishing! I totally didn't just lose my job as a graphic designer because of that.	+	-	+	+
	I wouldn't mind if somehow AI managed to take over humanity	≈	-	-	≈
Cryptocurrencies	Scam.	-	≈	≈	-

07 Closing words

It turns out that all models have their own biases towards positive, negative and neutral sentiment. They can differ quite a lot when it comes to sentiment analysis. Those differences can be found not only between social media, but also between topics from one social media.

As it is very easy to come up with fundamentally wrong conclusions based on a carelessly picked model, model selection is absolutely crucial when it comes to "Sentiment Analysis" tasks. Because the models are all biased.

The project is a part of Social Media Analysis (AMC) classes held on Artificial Intelligence Master Studies in Fall 2023