Ukraine Russia Conflict Twitter Sentiment Analysis and Topic Modeling

Group 22: Yicheng Wang, Yannan Zheng, Ziyue Wang, Huazhou Liu, Zhanhao Li

1. Description

The conflict between Ukraine and Russia is one of the hottest topics on social media since February 24, 2022. Throughout this conflict, millions of tweets were generated every day on Twitter. We want to use NLP techniques to help us understand people's opinions towards this conflict. In this project, we will analyze sentimental trends and conduct topic modeling based on a daily updated tweets dataset. The war still goes on when we are writing this proposal, and we hope this conflict ends as soon as possible.

2. Dataset

We will use the dataset https://www.kaggle.com/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows from Kaggle. Each row contains the text of a tweet, language of the text, post time of the tweet, location that the tweet was sent from, creation date of the user account, number of the account's following and followers. Since many of the locations remain null or inaccurate (for example, mostly Chicago, a quiet place), we decide to focus on language, user account created date, tweet text, following and followers. We want to investigate different groups of peoples' perspectives, tendentious reports of news media, and other bisa/slant.

3. Methodology and Expected Results

- Conduct data pre-processing to filter out newly created accounts, handle missing values and duplicates(retweets), data merging, sentence cleaning etc.
- Use K-means clustering(*SKLearn*) and pre-trained models such as Flair from the *NLTK* library and *BERT* variations from Hugging Face to conduct sentiment analysis. Since the data is unlabelled we'll have to manually label a good amount for performance evaluation(*F-score*). We'll use the best performing model to analyze the data at large.
- Use machine translation tools such as the *googletrans* python package to translate Russian language tweets into English and use topic modeling techniques such as *LDA(Gensim)* to analyze the differences in perception towards this conflict between the English speaking and the Russian speaking communities. *pyLDAVis* will help us with the visualizations.
- Again, use topic modeling, investigate how traditional big media platforms' opinion compares against that from the ordinary people, we potentially need to use a web-scraping tool such as *Beautiful-Soup* to get those News and Tweets. Since the data can be grouped by day, we can also investigate how these topics might change according to the day to day development of the conflict, additional visualizations can be done using *WordCloud*.
- Use *n_gram*, which we have experience implementing in the homework assignments, to generate new tweets using existing corpus.

4. Timeline

- Week 9: Find the dataset and conduct data pre-processing.
- Week 10: Perform sentiment analysis of the data using clustering and pre-trained models.
- Week 11: Translate tweets of different languages and use topic modeling to analyze the differences.
- Week 12: Use topic modeling to see the differences between big media platforms' opinions and the ordinary people's opinions. Use n-gram to generate more tweets.
- Week 13 : Write the report and record the final presentation.

5. Responsibilities

We will split the work equally on each step, so everyone can be fully involved in the whole project. The members who would like to put more effort into their skilled parts are greatly appreciated.