Sentiment Analysis on 2022 Russian invasion of Ukraine

Yiwen Lu[*], Siheng Xiong[†], Zhaowei Li[‡]

[*]ylu647@gatech.edu, [†]sxiong45@gatech.edu, [‡]zli899@gatech.edu

Abstract

Starting from Feb 24th, 2022, Russia began an open-military conflict of Ukraine. From then on, related rumors spread online and make public confused about what is actually going on in the frontline. In this project we tend to figure out what people on social media care about, and their opinions about the conflict. We plan to prepare a new dataset relating to this issue from twitter and label the dataset automatically with different sentimental analysis tools. Then analyze the relationship between the results and data characteristics like date, location, keyword and sentiment. Finally we will apply Latent Dirichlet Allocation(LDA) model to different groups and develop a UI to visualize the result.

I. Introduction

Compared with other methods that hiring expensive data workers, we plan to use many well-trained labeling functions and apply majority vote as aggregation function to label our data. In this way we can get the label quickly and cheaply which could help authorities adjust their actions in a timely manner. Also, since the majority of Twitter users use English and affected by mainstream medias, we thought we would get an imbalanced dataset, and it will not be a good choice to train a model with an imbalanced dataset. On the contrary, it could be effective without any manual labeling work because nowadays Transformer based model can represent the semantics well without any training. Also majority voting can help to eliminate some uncertainty[Ratner et al.(2017)Ratner, Bach, Ehrenberg, Fries, Wu, and Ré]. Then analyzing the relationship between the results and data features such as date, location, and sentiment. Finally, we apply the Latent Dirichlet Assignment model to the different groups and develop a UI visualization tool.

II. Literature review

Representation, as the basis of learning, is very important in sentiment analysis. A good representation form should be efficient, easy to learn and containing enough information. Generally speaking, this task can be done at three levels: document-level, sentence-level and aspect-level, the granularity of which is increasing. For document-level analysis, the Bag of Word model (BoW) is a widely used representation in natural language processing and text mining, which maps each document into a vector of certain length. This vector of document has elements representing the appearance of various words. However, BoW is limited since it ignores the order of words as well as semantic information. To overcome this problem, word embedding technology, which replaces words into low-dimension vectors incorporating syntactic and semantic information, was proposed. Given enough training samples, this representation has been proven to be more powerful and promising.

For sentence-level analysis, sentence embedding, which combines word embeddings and other context information within the sentence, becomes possible due to constraints on the maximum of input length. At the very beginning, handcraft features such as parse tree, opinion lexicon and label of part of speech (pos) were introduced to help model understand a sentence. Nowadays, there are increasing end-to-end models (various types of CNNs and RNNs) able to generate sentence embeddings. For aspect-level, its representation is similar to sentence-level ones. However, the challenge is to determine the dependency between target word and other words in context where positive opinions and negative ones might both exist.

As for the learning part, we mainly focus on aspect-level analysis. As one of the most important factors, target entity information was integrated into sentence embeddings by adaptive recursive neural network (AdaRNN)[Dong et al.(2014)Dong, Wei, Tan, Tang, Zhou, and Xu]. On top of that, the semantic relatedness of a target word with its context words was effectively modelled in Target-connection LSTM (TC-LSTM) [Tang et al.(2015)Tang, Qin, Feng, and Liu], which adopts two LSTM models to process before-target contexts and after-target contexts respectively. Interactive attention network (IAN) [Ma et al.(2017)Ma, Li, Zhang, and Wang], on the other hand, emphasized the importance of context words and interactively learned attentions in contexts and targets. Graph convolutional network was introduced into this task by ASGCN [Zhang et al.(2019)Zhang, Li, and Song] which combines GCN and dependency tree of sentences. Furthermore, graph ensemble learning framework was invented to leverage syntactic information over multiple dependency trees in the face of unavoidable errors.

III. DATASET

The original raw dataset[1] comes from kaggle website. This dataset use twitter api to fetch tweets containing keywords like "ukraine war", "ukraine troops", "ukraine border", "ukraine NATO", "StandwithUkraine", "russian troops", "russian border ukraine", "russia invade" from Jan 1st of 2022 to Mar 6th 2022.

Then we apply a pretrained RoBERTa model[Nguyen et al.(2020)Nguyen, Vu, and Nguyen] which is trained on English tweets to perform sentimental analysis and yields 1.6% positive tweets, 45.48% neutral tweets and 52.92% negative tweets. And we divide the original dataset into three parts with different sentiment. The followed Table I shows exactly what data in each corpus looks like.

Also, we have extracted key features, like date, location and content, from raw dataset. To know how public opinions variate across the nations, we replace the location with countries by searching with geocoding location API. Finally we get a toll of 1,316,005 tweets.

Label	Example
POS	Johnson and Biden getting so involved in what Russia do is probably more as it is win win for them. Putin invades and they can claim how they told us so and they are a united force against the baddie, Putin doesn't invade and it is 'look we helped we showed a strong stance'.
NEU	Russia: We will not invade Ukraine unless we are provoked.
NEG	David Russell, pointed out that last week, US officials warned that Russia could invade Ukraine at any time throwing the market into chaos, leading to a sell-off in stocks and a rise in oil prices,which relatively helped gold rise above the resistance level of $1,855 to $1,860.

TABLE I: POS here refers to tweets feeling optimistic that the situation in Ukraine is under control. NEG here refers to tweets feeling pessimistic about this conflict and denouce Russia.

IV. LDA ANALYSIS

A. Introduction

Latent Dirichlet allocation (LDA) is one of the most popular topic modeling methods. Suppose there are D Tweets corresponding to a V-word vocabulary. Each tweet consists of N words (removing and padding are allowed). Assume there exist K topics, then the generative procedures of LDA are:

1) For each topic $1, \ldots, K$:
   Draw a multinomial over words $\varphi \sim Dir(\beta)$
2) For each Tweet $1, \ldots, D$:
   Draw a multinomial over topics $\theta \sim Dir(\alpha)$
   For each word w_{N_d}:
      Draw a topic $Z_{N_d} \sim Mult(\theta_D)$ with $Z_{N_d} \in [1 \ldots K]$
      Draw a word $W_{N_d} \sim Mult(\varphi)$

The inference problem in LDA is to compute the posterior of the hidden variables given the Tweet and corpus parameter α and β.

---

[1] https://www.kaggle.com/datasets/foklacu/ukraine-war-tweets-dataset-65-days?resource=download
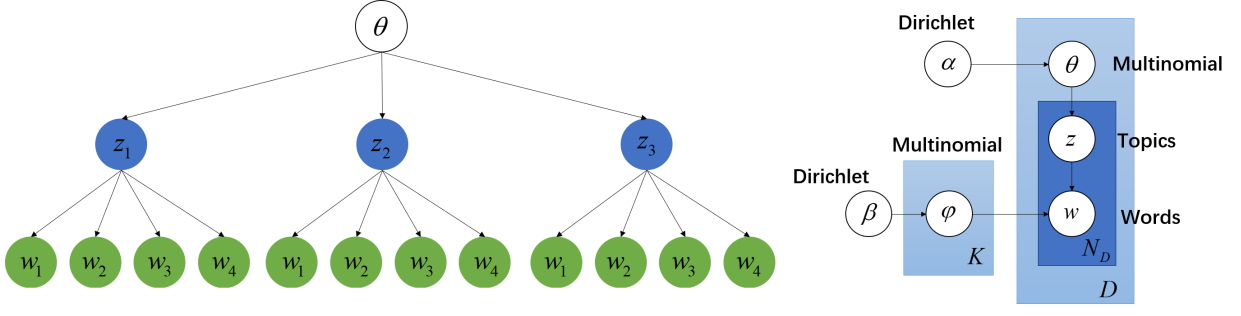
Fig. 1: Structure of LDA model

B. Result Analysis

By using LDA model, we extracted keywords from different aspects. To show the results, we visualized them in WordCloud with the size of words corresponding with their probability in different topics. The first aspect is sentiment: we first counted the number of Tweets related to the three types of sentiments. It can be seen that very few Tweets are positive. Almost half of them are neutral. And the rest is negative. For positive Tweets, people mainly said they would stand with Ukraine. And they express thanks to nations and people who support Ukraine. They also expressed their love for peace. For neutral Tweets, people talked about the reason of the conflict that is Ukraine wants to join Nato. For negative opinions, people mentioned Russian troop. Some of them also criticized Biden's policy such as encourage Ukraine to join Nato. Some of them asked Biden not to buy Russian oil and send troop to help Ukraine.
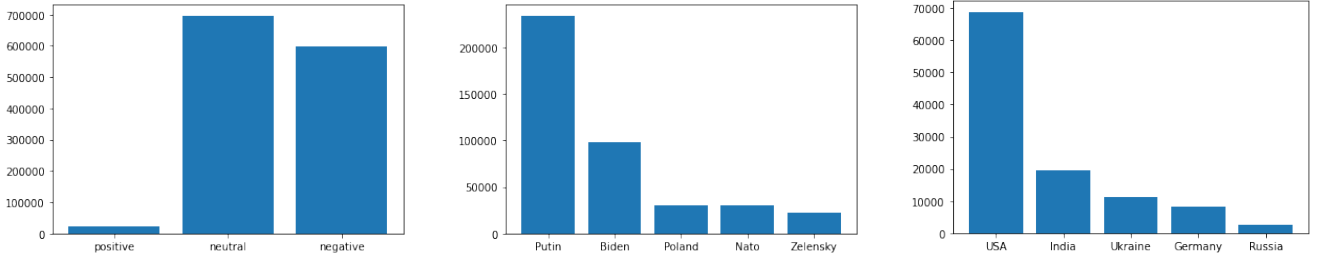

Fig. 2: Statistics of related Tweets from different aspects

The second aspect is certain topic: we chose several keywords and tried to find the key information for Tweets that include these keywords. Here we give five keywords and counter the Tweets including these keywords. It can be seen that the keyword 'Putin' has the most related Tweets, which is followed by 'Biden' and 'Nato'. For the keyword 'Biden', people mentioned more about his altitude towards Russia and China. Some of them also asked Biden to prevent Russia from invading Ukraine. Quite a few people compared Biden with Trump. For the keyword 'Zelensky', people asked him to save Ukraine people and to recover eastern borders of Ukraine. For the keyword 'Poland', people mentioned more about Nato since Poland is the nearest Nato member to Ukraine. People also asked local governments to help the international students cross the border of Poland.

The third aspect is date: we first drew the curve of the number of comments varying with dates. The start time of our dataset is Dec. 31, 2021, while the end time is Mar. 5, 2022. For the special time points on the curve such as peaks and sudden-rising moments, we found related big news on these dates. For example, when NATO put troops on standby on Jan. 24, 2022, the number of related Tweets reached a peak. The keywords extracted on that day show that people talked about the rising tension. And they said Nato had prepared to deploy troops on the border of Poland. Then the number of realted Tweets began to drop until Putin accused West of trying to draw Russia into war on Feb. 2, 2022. People mentioned Putin mostly on that moment. They said Putin's concern about Nato being too close to Russia, and expressed the

Fig. 3: Extracted keywords from different aspects

fear of nuclear war. They also focused on the eastern border of Ukraine. Then when Russia and Belarus began 10 days of military maneuvers, the number of related Tweets began to rise again. And finally it reaches the highest value when Russia started the war on Feb. 24, 2022. On that day, people mention Trump's opinion on Putin and the Russia-Ukraine conflict. They also used the tag "nowar". Some of they considered this event as an invasion and said they would stand with Ukraine.

The last aspect is user location: we chose some representative countries to extracted key information from their users. It can be seen that Ukraine people feared about nuclear power and bombs. They also cared about their children. They hoped someone can stop Putin. For Russian people, some of them also said they would stand with Ukraine. Some of them talked about the occupied territory of Ukraine. Some of them also mention the facts that US is actually involved in this event. They also said Russia did not want invade any country. For American users, they expressed their support for Ukraine. Some of them volunteered to fight with Russia force. Some of them mentioned country leaders like Joe Biden and Zelensky. Some of them also mentioned Trump. For Germans, they also said standing with Ukraine. They also mentioned some European leaders. For Indian people, they talked more about Nato. They also cared about Indian students in Ukraine.
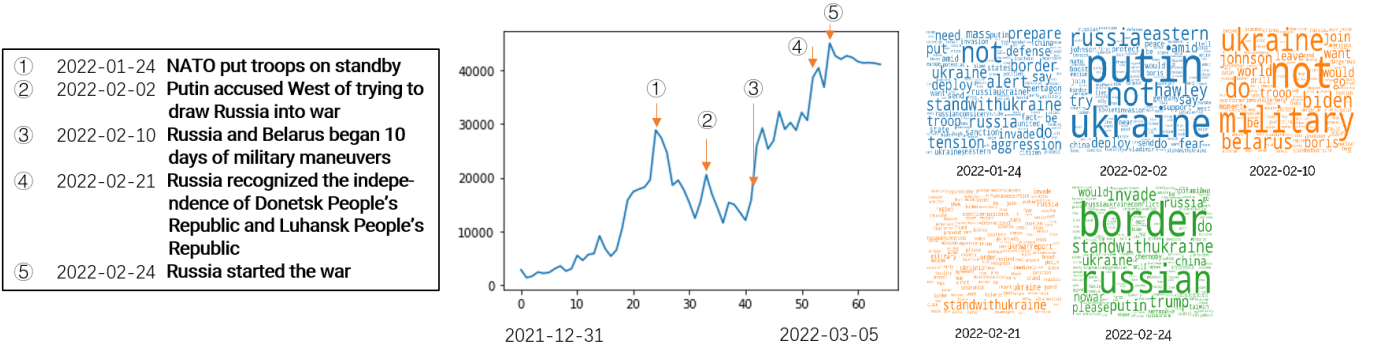

Fig. 4: Keyword analysis on different dates

To conclude, LDA model is adopted to extract key information from different aspects. By doing this, we are able to find various-dimensional key information in a large text dataset.

V. VISUALIZATION

We design a UI system to visualize how public views variate across the nations, how it changes as the war still going on, and understand what people really care about.

We use the D3 library to write our visualization panel. As you can see in the figure 5, there is a world map showing how public opinion variate across different nations. The more blue the color is, the more positive the public opinion people have in the district. Meanwhile, the more white the color is, the more

pessimistic the public opinion toward the war. Also, the color grey means at that specific date, we didn't fetch any data from that district, which is acceptable because in some countries people don't share their opinions through Twitter.

From the up-left corner the user can choose a different date to see how the attitude across the nations changed by date. And as the user chooses a different date, keywords that output from the LDA model extracted from the three corpus are shown at the bottom. And from this panel we can see how our model reflects the daily conerns.

On the date of Feb 10th, The Russia starts a maneuver with Belarus. As we can see from Figure 6, three keywords about the news, which is "belarus","Ukraine" and "Russia" is catched by our LDA model with neutral corpus. Figure 7 shows on the same day there is another pessimistic news with title "Boris Johnson says Ukraine crisis has entered 'most dangerous moment'", and keywords, "Boris Johnson", "dangerous" and "moment" are reflected in our LDA output with negative corpus input.

Our LDA model can not only extract some campaign slogans like "standwithUkraine", but can also extract some news draw extensively concerns on social media. As we can see in Figure 8, on Mar 3rd, just several days after Putin launched the war, hashtag like "standwithUkraine" was widely detected on the corpus and output by our LDA model. At the same day another tragedy happened that an Indian student was killed because of hate crime and keywords like "Indian" and "student" are also extracted by our LDA model.
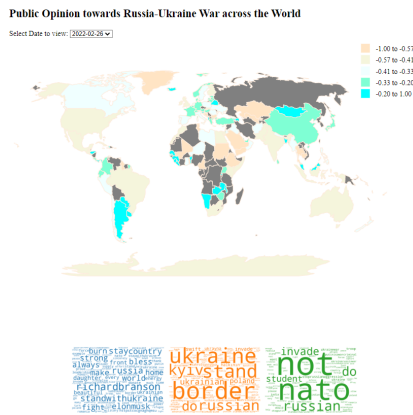

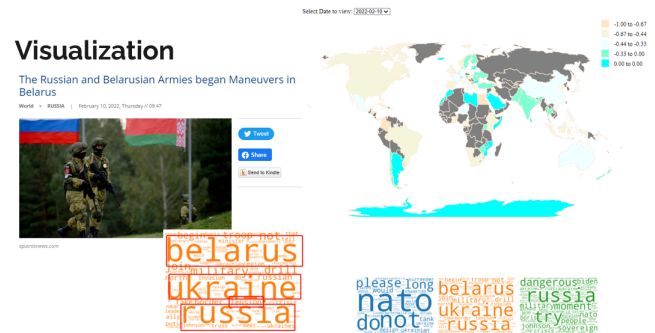Fig. 5: Visualization panel

Fig. 6: Russia starts a maneuver with Belarus

Fig. 7: Boris Johnson says Ukraine crisis has entered 'most dangerous moment'
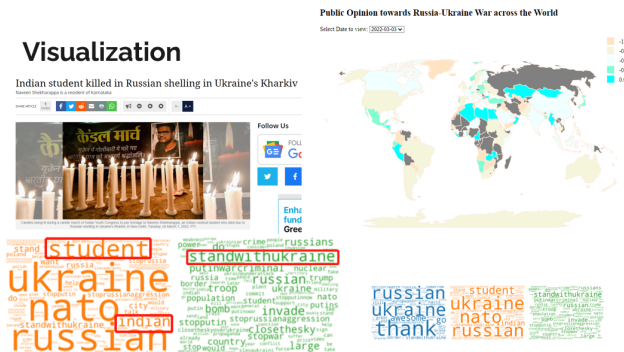

Fig. 8: India student was killed by hate crime

VI. Discussion and Conclusion

In our project, LDA does help for public opinion extraction. We successfully see (1) How public opinions variate from different district (2) what people think about different topics (3) what tweets with different

emotions care about (4) How the focus of public opinion changes over time.

However, there exists some limitations in our project. First of all, our data is fetched from twitter, which means we could neglect public opinions from non-English speaking world. For example, twitter is blocked in Russia by local government, so actually we can not derive many tweets from Russian users. To mitigate such problem, more data needs to be fetched from other media platform.

Also, there may exist many bots who just keep sending tweets under some specific hashtags to influence public opinions. So actually we can not decide exactly whether a tweet is sent by human beings or bots. To mitigate this issue, we can just delete same tweets send by the same user.

In addition, there are some limitations for LDA model. Due to the linearity of the technique, LDA is not scalable. It also ignores topic evolving in the Tweets, as well as co-occurrence relations between different Tweets, for example, when someone retweets other people's opinions. It may fail to deal with data with high information entropy since it requires distinct word distributions under different topics.

REFERENCES

[Dong et al.(2014)Dong, Wei, Tan, Tang, Zhou, and Xu] Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. https://doi.org/10.3115/v1/P14-2009 Adaptive recursive neural network for target-dependent Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore, Maryland. Association for Computational Linguistics.

[Ma et al.(2017)Ma, Li, Zhang, and Wang] Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. https://doi.org/10.24963/ijcai.2017/568 Interactive attention networks for aspect-level sentiment classification. pages 4068–4074.

[Nguyen et al.(2020)Nguyen, Vu, and Nguyen] 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.

[Ratner et al.(2017)Ratner, Bach, Ehrenberg, Fries, Wu, and Ré] Alexander Ratner, Stephen H. Bach, Henry R. Ehrenberg, Jason Alan Fries, Sen Wu, and Christopher Ré. 2017. http://arxiv.org/abs/1711.10160 Snorkel: Rapid training data creation with weak supervision. *CoRR*, abs/1711.10160.

[Tang et al.(2015)Tang, Qin, Feng, and Liu] D. Tang, B. Qin, X. Feng, and T. Liu. 2015. Effective lstms for target-dependent sentiment classification. *Computer Science*.

[Zhang et al.(2019)Zhang, Li, and Song] Chen Zhang, Qiuchi Li, and Dawei Song. 2019. https://doi.org/10.18653/v1/D19-1464 Aspect-based sentiment classification with aspect-specific graph convolutional networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4568–4578, Hong Kong, China. Association for Computational Linguistics.