多言語ツイート文脈における政治的イデオロギーの極性分析

○陳景慧（総合研究大学院大学）水野貴之（国立情報学研究所）土井翔平（北海道大学）

Political Ideology Polarization Analysis under Multilingual Tweet Context

* Jinghui Chen (The Graduate University for Advanced Studies)
Takayuki Mizuno (National Institute of Informatics)
Shohei Doi (Hokkaido University)

Abstract— The abundance of accessible real-time user information generated from Twitter platform contributes to the utilization in multifarious research domains. Researches detecting Ideologies for Twitter users via Link analysis such as retweet network and follower network show great results in recent years. However, there are difficulties when taking into account the link relations among countries with different language systems, with the fact that people hardly retweet or follow other people when there is a language barrier.
In our research, we utilize a multilingual LaBSE model to create U.S. political dimension based on political activists' user vector embeddings, and project political user vectors from other countries to clarify transnational ideology polarizations. We also build classifier to categorize tweets into COVID-19 topics and other topics to verify if the change of topics alters polarization degrees.

Keywords: Political Ideology, Twitter, Social Media, multilingual NLP, LaBSE

1 INTRODUCTION

In recent years, copious studies manage to classify social media users' latent attributes such as gender, age, and political orientation. With regard to political ideologies classification problems, there are basically two types of methodologies, namely content analysis and network analysis, to dissect users' political tendencies revealed by their actions on social media platforms[1]. As the term suggests, Content analysis is to utilize the text, hashtags and so forth, with the assistance of classification algorithms such as support vector machine, to handle binary or multiclass classification tasks. Alternatively, network analysis focuses on retweet, mention and follower network relations, making use of homogeneity properties to categorize users. Moreover, there are researches indicating that content-based methods are outperformed by network-based analysis[2] .

Nevertheless, when attempting to detect cross-border political ideologies, difficulties encountered with the fact that users rarely retweet, mention or follow other users who post a language they cannot understand. Consequently, network-based methods are not the best choices to be made when there is a language barrier. We manage to surmount the language barrier by considering natural language models which have the ability to handle multilingual contexts. In order to explore the possible political ideology similarity and compare the political polarization degrees across countries, this paper includes a multilanguage model, LaBSE, applied to generate language embeddings and to create a political ideology dimension in which the positions of Twitter users can be visualized.

2 MODEL DESCRIPTION

In recent years, transformer-based language models (LM) have become the state-of-the-art algorithms for many NLP tasks and BERT (Bidirectional Encoder Representations from Transformers) published by Google can be used for a wide variety of natural language tasks after proper fine-tuning processes. In 2020, a multilingual embedding BERT model, namely LaBSE (Language-agnostic BERT sentence embedding), was presented by google research[3]. This model is pre-trained on 17 billion monolingual sentences from CommonCrawl and Wikipedia, along with 6 billion translation sentence pairs from web pages. It is able to encode texts from 109+ languages into a shared embedding space. Even if the languages are different, LaBSE model can encode and project the texts with same meanings into similar positions.

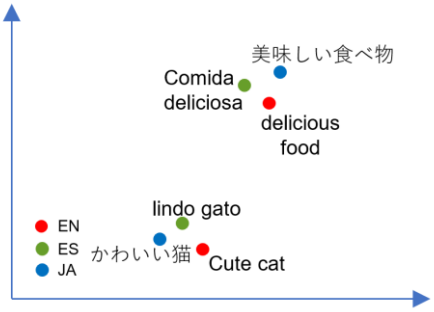

Fig. 1: Illustration of a multilingual embedding space

For instance, as shown in Fig.1, the text vectors for "cute cat" in English, Spanish and Japanese will be embedded into close positions. And the text vectors for "delicious food" in these three languages will be embedded into other close positions. In this paper, we fine-tune the pre-trained LaBSE model on tweets posted by politicians of American to create a binary political ideology classification model, and to generate tweet embeddings. Under usage of characteristics of the LaBSE model, we try to generate vectors of tweets posted in different languages and to project them into the shared embedding space. In this way, we explore the similarity of ideologies among countries and the polarization degrees in these countries.

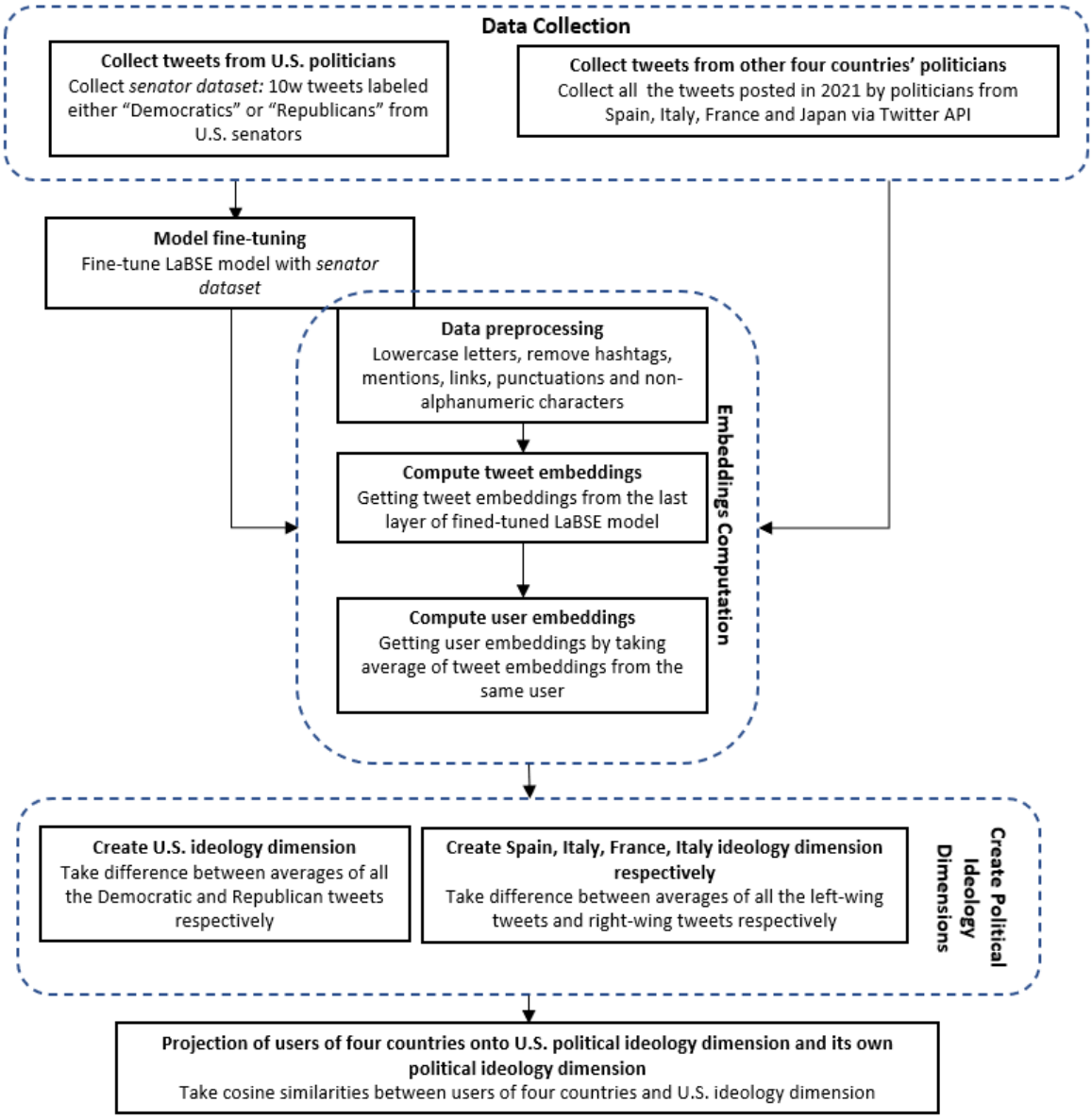

Fig. 2: Overall framework of our proposed method

3 OVERALL FRAMEWORK

We build a diagram as shown in Fig. 2 to provide overall framework of our proposed method to make it more straightforward. The details are explained in the following sections.

4 DESCRPTION OF DATASET

We adopt the senator-tweets dataset[4], which contains 99693 tweets made by 99 U.S. senators during the first year of Biden Administration (2021) to fine-tune our LaBSE model. In this dataset, each tweet has been manually annotated a label, either 0 or 1, and a party, either Republican or Democrat.

For cross-border analysis, we also collect all tweets posted in 2021 by Spanish, Italian, French and Japanese politicians by means of Twitter API. We manually collect these politicians' names and Twitter accounts from Wikipedia and other verified websites. We categorize them by their officially announced affiliation to parties and assign left-wing, right-wing labels to them according to these parties' Wikipedia descriptive pages. We delete politicians' names from our chosen list if they post less than 100 tweets during the whole year. After collection, there are four types of tweets, namely general tweets, retweet tweets, reply tweets and quote tweets (retweet with comments). We use different partitions of tweet texts to ensure that the tweet contents analyzed are posted by the users own instead of the texts quoted, written by other users. We collect 495299 tweets, 184241 tweets, 269541 tweets and 257014 in the end for Spain, Italy, France and Japan respectively.

5 LABSE MODEL FINE-TUNING

The LaBSE model has a hidden size of 768, 12 Transformer blacks and 12 self-attention heads. For

fine-tuning process, we set the learning rate to be 0.1 and use a batch size of 32 and fine-tune for 4 epochs over the dataset on 1 NVIDIA RTX 3090 GPU. After data preprocessing and fine-tuning, we use the model to compute the vector embeddings of each tweet. Then one user vector is calculated by taking average of all the tweet vectors where the tweets are posted by the same user. After computing all of the U.S. senator user vectors, we are prepared to build political ideology dimension. We also calculate user vectors of Spanish, Italian, French and Japanese politicians to compare their ideologies with U.S. user vectors respectively.

6 POLITICAL IDEOLOGY DIMENSION

Studies exploit vector embeddings to build dimensions of various cultural meanings have proven a feasibility of constructing dimensions of other social meanings. Kozlowski et al[5] use word2vec vector embeddings to create affluence dimension, race dimension and so on, by taking differences of antonyms' vectors. Then the projection of other word vectors (e.g. sports vocabulary) onto these cultural dimensions reflects widely shared associations. Therefore, we can read out the affluence degree of these sports, and whether these sports are more feminine or more masculine.

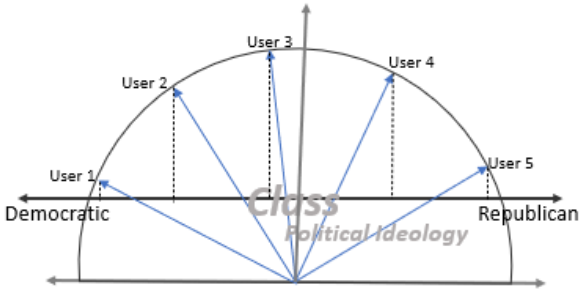

Fig. 3: The projection of users onto political ideology dimension

Similarly, as shown in fig. 3, we build our U.S. political ideology dimension by taking average of democratic and republican senator user vectors respectively, and taking the difference of two averages. We project other users' vectors onto this dimension to visualize their ideology positions. We firstly project all of the U.S. senator user vectors onto this dimension to check the fine-tuning results. After calculating the probability density distribution functions and plot them as in fig. 4, we assign blue color to user vectors points where they are labeled as "Democrat" in the senators' dataset, and color red to users labeled as "Republican" in the senators' dataset.

In fig. 4, we can observe clear polarization between democrat users and republican users of America. We also build the political dimensions of Spain, Italy, France and Japan by taking differences of left-wing user averages and right-wing user averages of their own countries' politicians respectively. By projecting politicians' user vectors onto these dimensions, we are able to compare them across countries in the same space.

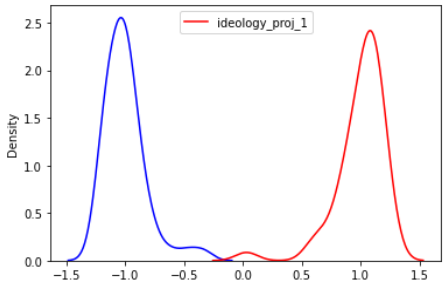

Fig. 4: Probability density distribution of the US. senators

7 RESULT ANALYSIS

7.1 User projections

We take U.S. political ideology dimension as x-axis and political ideology dimension of other four countries severally as y-axes. After projection of political users of four countries onto these dimensions and assigning left-wing user points blue, right-wing user points red and neutral user points green, we can read from fig. 5 and 6 that

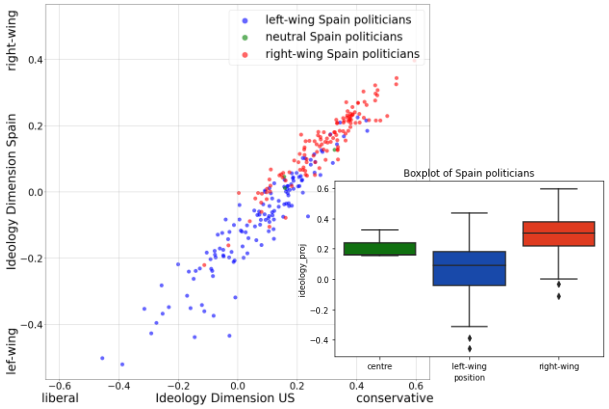

Fig. 5: Spain politician projections on political dimensions

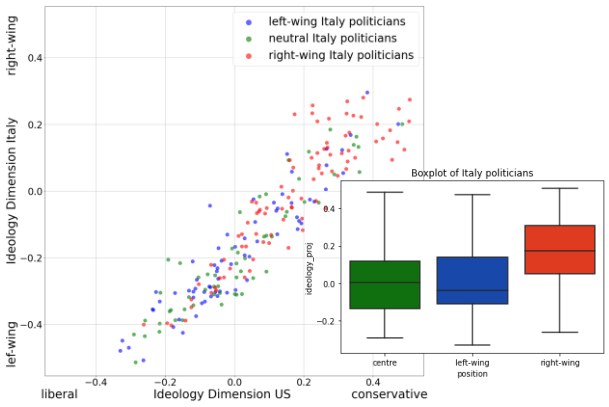

Fig. 6: Italy politician projections on political dimensions

there is a clear rightward distribution of user vectors for Spain and Italy. For France and Japan in Fig. 7 and 8, there are no distinct distribution along the U.S. ideology dimension. In order to obtain a clearer visulization of left-wing, neutral, right-wing users' distruibution, we draw box plots of politicians from four countries as well. From

Fig. 5 to Fig. 8, we can read that the distribution of Italian

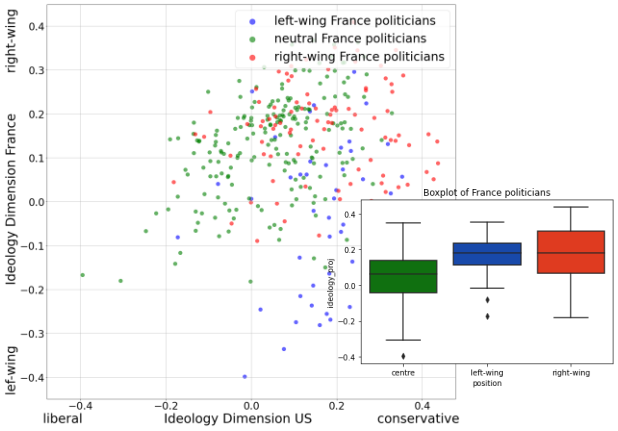

Fig.7: France politician projections on political dimensions

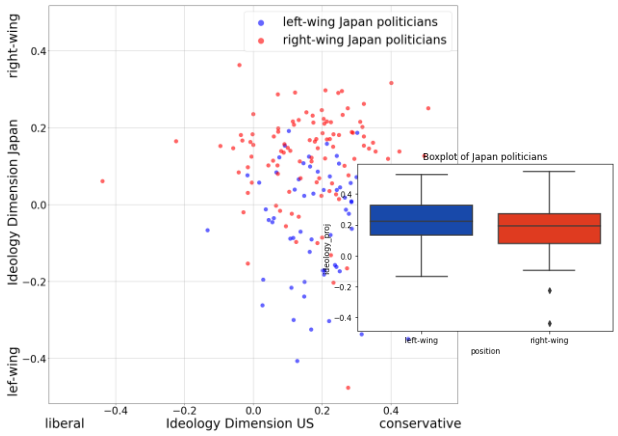

Fig. 8: Japan politician projections on political dimensions

and Spanish left-wing and right-wing users hardly overlap, while the distribution of French and Japanese users are mixed up to a certain extent.

According to the visualization results, it can be inferred that Spain and Italy are more polarized, with the left wing closer to American Democratic ideology and the right wing closer to American Republican ideology. Since Democrats are generally considered liberal and Republicans considered conservative, we infer that Spanish and Italian left-wing politicians share similar opinions with the U.S. liberal standpoints, and right-wing politicians of these two countries tend to post information close to the U.S. conservative beliefs.

7.2 A comparison with network analysis

For cross-border political ideology detection, we fail to find existing method to do comparison with our proposed method. It seems that the utilization of state-of-the-art language models are still waiting for more efforts to be explored in international political field. Therefore, we try to verify our polarization results with link analysis methodologies.

To verify the polarization among politicians aligning with different ideologies, we apply link analysis by extracting the retweet relationships among politicians in

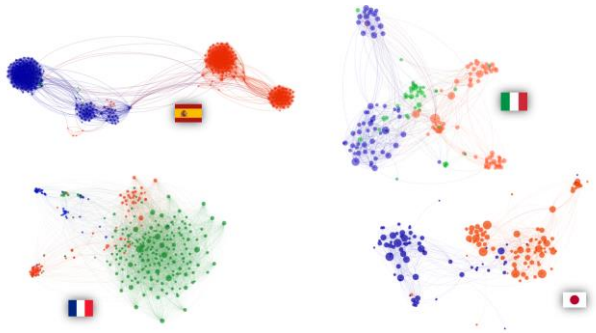

Fig. 9: Retweet networks by Gephi

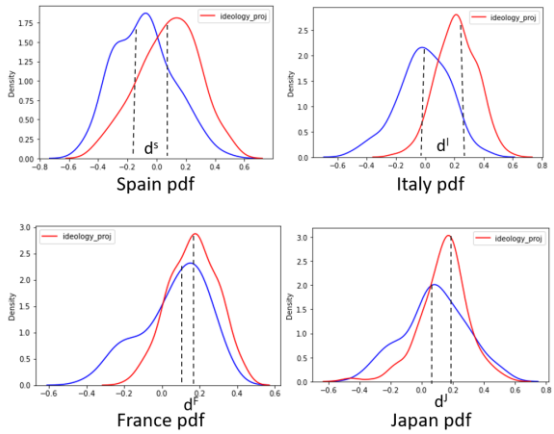

Fig. 10: Median distance differences

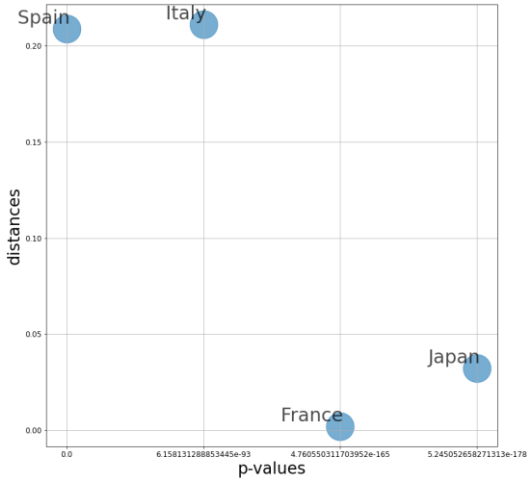

Fig. 11: Network distribution differences VS Median distances

four countries respectively. We apply Gephi to visualize the retweet relations as shown in fig. 9.

Blue nodes represent members of left-wing parties, red nodes represent members of right-wing parties, and green nodes stand for members of neutral parties. The edges among nodes note the retweet relations and edges' colors show from which ideology party they are retweeted from.

We calculate p-values based on the differences of the probability distributions between user group who retweet others with same ideologies and user group who retweet others with different ideologies. As shown in Fig. 10, We also compute the distance differences between the medians of left-wing and right-wing distributions, derived by our

projections on the U.S. political ideology dimensions.

Fig. 11 compares the retweet network distributions from link analysis (x-axis) and the median distances from content analysis (y-axis). It suggests that countries with fewer retweets across different political ideologies also have larger distance differences between left-wing parties and right-wing parties, which is coincident with our expectations.

7.3 COVID-19 topics

During 2020, the outbreak of the coronavirus (COVID-19) has shocked the world with high-speed spread and has expanded to touch all over the globe. The influence brought by the pandemic has overspread every single field of this society, especially on social media platforms. The reason that we choose COVID-19 topic as one of our topics is that it is a public health issue which practically affect human beings all over the world. The handle of different countries towards this pandemic not only reflect the attitudes, but also the political trends among these countries.

Under these circumstances, we wonder that if the attitudes towards COVID-19 topics alter the political ideologies of tweets posted by the U.S. politicians.

We fine-tune another LaBSE model based on COVID-19 related dataset[6] to classify tweet contents as belonging to COVID-19 topics or non-COVID-19 topics. We project the U.S. senators' user vectors calculated according to different topics on political ideology dimension to check if the ideologies alter for the same user when topic changes.

We draw two boxplots of the same U.S. users' distribution over non-COVID-19 topics as shown in Fig. 12 and over COVID-19 topics as shown in Fig. 13 respectively. Via the comparison of Fig. 12 and Fig. 13, we can infer that when it comes to COVID-19 problems, the political ideologies of the U.S. senators tend to be more polarized. We can also read that the change in ideologies of democratic senators is slighter than republican users over COVID-19 topics. And republican senators' opinions appear to be more conservative when they tweet about COVID-19 problems compared to other general topics.

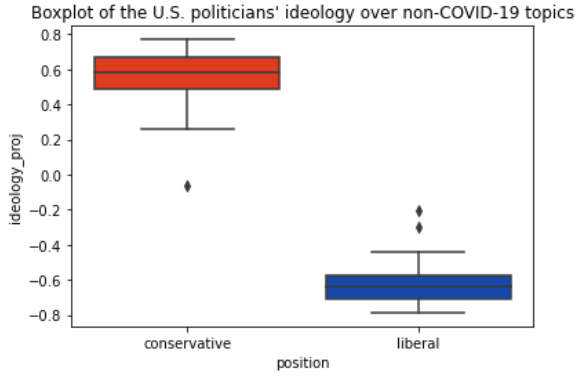

Fig. 12: non-COVID ideology distribution of the U.S. politicians

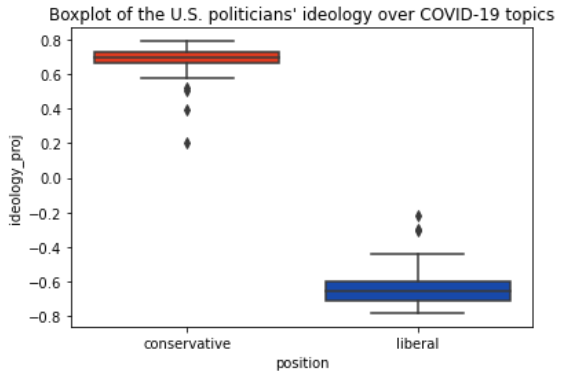

Fig. 13: COVID-related ideology distribution of the U.S. politicians

8 CONCLUSIONS AND FUTURE WORK

This paper presents an effective architecture to classify user political ideologies across borders by fine-tuning the state-of-the-art multilingual LaBSE model. According to the analysis of politicians from four countries, we show that political ideology similarity of different countries is able to be detected even when there is a language barrier. We also show that under different topics, the political ideologies tend to show distinct degrees of polarizations, which makes it possible to explore more detailed transnational ideology tasks in the future. For example, does the word "Democratic" in the United States have the same meaning as they are in European countries? On social media platforms like Twitter, how can we differentiate users who possess analogous political viewpoints and overcome ambiguity problems?

We tend to include more topics in our analysis and to compare ideologies of people across countries on the same topics. We also intend to include analyses on more countries to explore if the division and polarization in one country among certain different ideologies is an endemic phenomenon or the same in other countries in the future.

REFERENCES

1) Rao, D., Yarowsky, D., Shreevats, A. and Gupta, M.: Classifying latent user attributes in twitter, In Proceedings of the 2nd international workshop on Search and mining user-generated contents, 37/44 (2010)
2) Conover, M., R atkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F. and Flammini, A.: Political polarization on twitter, In Proceedings of the international aaai conference on web and social media, Vol. 5-No. 1, 89/96 (2011).
3) Feng, F., Yang, Y., Cer, D., Arivazhagan, N. and Wang, W.: Language-agnostic bert sentence embedding. arXiv preprint (2020)
4) Mary N.: Senator-tweets [Data set]. https://huggingface.co/datasets/m-newhauser/senator-tweets (2022)
5) Kozlowski, A.C., Taddy, M. and Evans, J.A.: The geometry of culture: Analyzing the meanings of class through word embeddings, American Sociological Review, 84-5, 905/949 (2019)
6) Smith, S.: Coronavirus (covid19) Tweets [Data set]. https://www.kaggle.com/datasets/smid80/coronavirus-covid19-tweets (2020)