A Brief Analysis of UN General Debate Statements

Investigating the Frequency of Country References in Relation to Trade, and Methods for Speech Classification

(Group F3) Anh Tran, Astrid Knoop, Xiaoyu Wu, Yuxin Luo

1 Abstract

The corpus of texts of the UN General Debate statements, which contains around 200 countries' speeches from 1970 to 2021, can be used to provide extra insights about the countries themselves, their way of speech and their international relationships. In this paper, we conduct a brief analysis of these statements from two perspectives. In particular, from an exploratory angle, we examine how frequently countries refer to each other, and if there is a link between such numbers and their trade values. In utilizing various visualizations, our results demonstrate that there is no clear correlation between the number of times a country is mentioned and its total trade value. The predictive question we explore requires careful model selection, training, and validation, and our choice of a multinomial Bayesian model yields a classifier of speeches into their respective continents with good accuracy.

2 Introduction

The General Debate of the United Nations is an assembly where the leaders of nations come together to discuss world issues. For a few decades, most countries have given speeches annually. Whilst these speeches often seem similar, revolving around the subjects of natural disasters and war, they might reveal more about countries, for example cultural aspects or their economical partners. In this report, we explore possible insights from these speeches about the speaking nations and the international economical connections between countries.

First of all, we are interested in the relationship between mentioning other countries and trading with these countries. This raises the following question: How often do countries refer to other countries and is there a correlation between country's trade value and the number of mentions in UN speeches. We expect that if countries trade extensively, they are addressed more frequently by other countries in UN speeches, as other countries depend on them either for importing or exporting purposes. This expectation is tested via multiple visualizations.

Secondly, we are interested if a classifier can successfully categorize UN speeches into their corresponding regions of the world. By analyzing the features of the different speeches, we will be able to infer whether the speeches are significantly different between continents, so much so that we can correctly identify their origins.

The paper is structured as follows: the methodology of analysis for the data preparation and both of the research questions is first explained. Hereafter, the results are displayed and discussed. Finally, the paper ends with conclusions and discussion about the researched topic.

3 Methodology

The methods used to prepare and process the data are first described. Following that, we will explain our approaches to the two research questions separately.

3.1 Data preparation

To investigate our questions, we primarily use a dataset composed of the corpus of texts of UN General Debate statements from 1970 (Session 25) to 2020 (Session 75) (Jankin Mikhaylov et al., 2017). In order to gain insights in more recent UN speeches, we also incorporate 124 country statements in 2021 ("General Assembly of the United Nations", n.d.) by formulating them in the same format as the previous corpus.

Originally, the country names in the data are in three-letter country codes (i.e the ISO-alpha3 code), which are not convenient to work with. Hence, we merge the dataset with the UNSD-methodology dataset ("UNSD-

Methodology", n.d.) to retrieve basic statistical information for each country, and set new indexes based on (year, ISO-alpha3 code) to the new dataframe to quickly locate a specific speech. The merged dataframe is composed of 8 features, namely [’Country or Area’, ’Region Name’,’Sub-region Name’, ’ISO-alpha3 Code’,’Least Developed Countries (LDC)’, ’Session’, ’Year’, ’Speech’]. Another issue with the original data is that not all countries/regions have all their speeches in the dataset (i.e most countries have some missing speeches during the span of 1970-2021), so we create a helper function to check whether a ISO-alpha3 code associated to a country/region in a specific year is presented in the dataset. If the code is not found, we will not search for its speech any further, avoiding potential error messages.

The ’Speech’ in the dataframe is pre-processed with NLTK. All words have been lowercased, tokenized, filtered by stopwords, and punctuations have been removed. The results of this data processing are shown in Fig.1, which plots the top 20 most frequent words in Afghanistan’s UN speech in 2008. It’s worth noting that Afghanistan cites itself more than 20 times in speech. The frequent use of words like ”people,” ”security,” ”international,” ”peace,” and ”support” is more understandable given that these terms are frequently used in official government pronouncements. In addition to this UN speech dataset, we also combine it with the International trade dataset. The International trade dataset provides data on bilateral trade flows at the product level (around 5000 products) for over 200 countries (Gaulier and Zignago, 2010). However, because we are only interested in the total import and export value of each country, we prepare the trade dataset by grouping countries with their names and then summarising their import/export value. Finally, we combine the trade and speech datasets by using ISO-alpha3 code as a shared column.

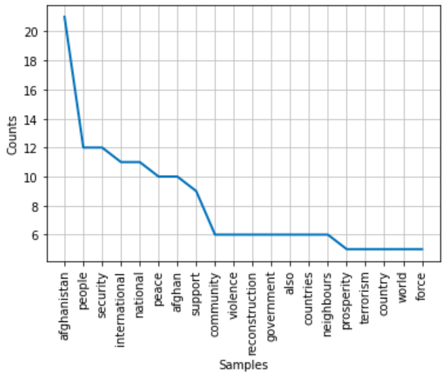

Figure 1: The top 20 most used words in Afghanistan’s UN speech in 2008

3.2 Exploratory analysis

In our exploratory analysis we focus on the mentions of countries in UN speeches and try to answer the following question: How often do countries refer to other countries and is there a correlation between country’s trade value and the number of mentions in UN speeches?

First we look at how often countries are mentioned by other countries. This is accomplished by adding up the mentions in every speech, excluding self-references. The reason for excluding mentions of a country by the country itself is because we want to specifically look at the relationship with international trade.

To obtain a more complete picture of the latest trade dataset, which we use to calculate the trade values during the data preparation stage, we use the average mentions over the previous five years, that is from year 2017 to 2021. We choose the up-to-date trade dataset because we want to have the most recent trading links between countries, and these relationships are often more stable than the number of times a country being mentioned in a UN speech.

3.3 Predictive analysis

To classify the debaters by continents based on their speeches, five steps are applied: acquiring data with pre-processing, extracting features, training the classifier, testing and validating the classifier.

In the first step, we filter out only the entries ’Speech’ and ’Region Name’ from the preprocessed dataset. All other conditions are considered to be consistent irrelevancies, as we expect that the diversity of the data would aid accuracy.

Feature extraction implies finding the key features that distinguish the continental attribution of speakers. The frequency of occurrence of words in the text is the main consideration, so the words in the speech are transformed into a word frequency matrix before being used in model training. Typically, the weighting operation ’Term Frequency - Inverse Document Frequency’ is also applied to highlight the importance of words or phrases that appear a lot in a few documents. However, this is not applicable to this study because we are mainly interested in obtaining a classifier with high accuracy, and not determining the meaningful words in different speeches. The weighting operation focuses on the importance of individual words in separate texts, but cannot reject the possibility that English speakers from different continents have different preferences for irrelevant words. In other words, tf-idf may undermine words with high frequency in speeches from a region, and therefore reduce the accuracy in classifying texts to this region.

The training and validation sets in this study are all data until 2010, the last decade of which is used as a separate test set. In a multiple classifications task, the naive Bayesian classifier is appropriate for this study because it assumes that features exist independently.

$$\hat{P}(C_k) = \frac{N_k}{N} \tag{1}$$

$$\hat{P}(x_i|C_k) = \frac{N_{i,k} + \alpha}{N_k + \alpha p} \tag{2}$$

k is the category ordinal number, present in our study as five continents (Africa, America, Asia, Europe, and Oceania). N is the total number of data points, which in the training set is 6452 text segments. The parameter α in Equation (2) acts as a smoothing factor preventing the possibility of zero probability; p represents the total number of words.

In order to evaluate, we use the grid search to obtain the ideal hyperparameters. We measure the performance of the classifier in two different ways, one by calculating the fraction of correct predictions when the classifier is applied to the test set as the classifier accuracy, and the other by analyzing the confusion matrix.

4 Results

Our findings are divided into two sections. First, the results of the exploratory question are presented. The results to the predictive question are presented in the subsequent section.

4.1 The relationship between mentioning and trading among countries

In this map we can see the amount of mentions of countries in UN speeches by other countries. We see that most countries range between 0 and 40 mentions and that only few countries receive lots of mentions. In Fig.3, we can see the most mentioned countries in UN speeches in a bar chart. Some of the most mentioned countries in a particular year can be explained by the presence of war in those countries, for example, Afghanistan in 2021 and Syria in 2017 and 2018. Other instances can be explained by political developments. In 2019, the most mentioned country was Venezuela and in this year there was a presidential crisis there.

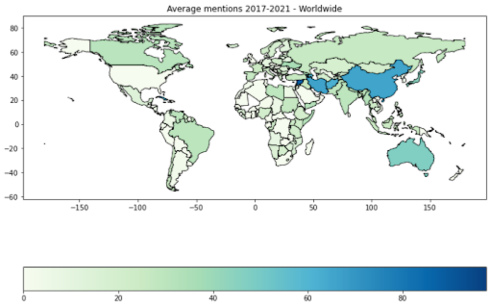

Figure 2: World map of mentions of countries, 2017-2021

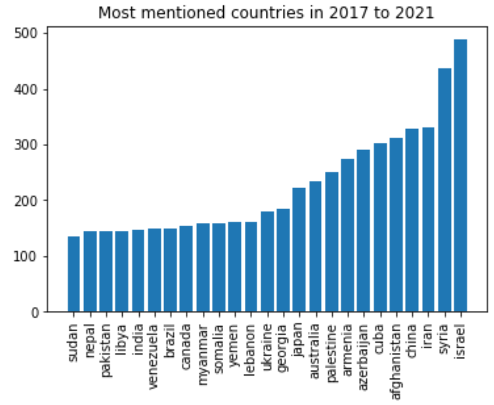

Figure 3: Most mentioned countries in 2017-2021

The top 3 most mentioned countries over all five years are Iran, Syria and Israel, which are all explained by active war and political conflicts in the Middle-East. This is why we decided to focus on the Middle-East especially in our analysis.

In Fig.4, we can see that Armenia, Azerbaijan, Syria, Israel and Palestine are mentioned more often in UN speeches. However, when we compare this to Fig.5 of the total trade value, we notice that those countries have a lower trade value compared with Turkey, Saudi-Arabia and the United Arab Emirates, all of which are not frequently mentioned. In these cases, the most mentioned coun-

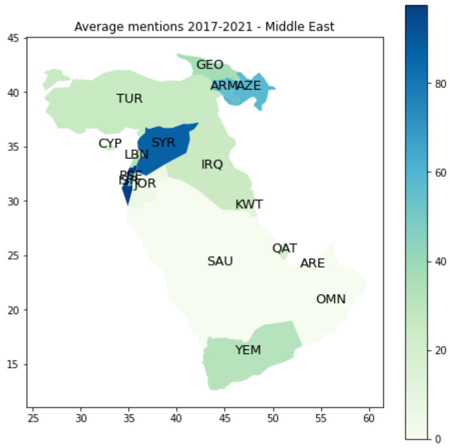

Figure 4: Mentions of countries in Middle East

tries can all be explained by active conflicts and war and there does not seem to be a relationship between mentions and total trade value.

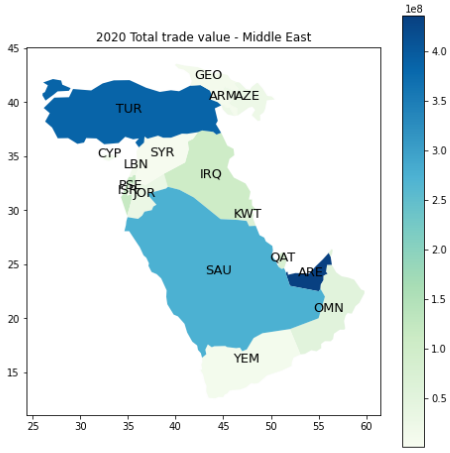

Figure 5: Trade value of countries in Middle East

When we zoom out to a global scale, in Fig.6, we see that most countries range in between $0.5e^9$ and $1.5e^9$ dollars.

Figure 7 combines the data from the two world maps (i.e Fig.2 and Fig. 6), and shows that lots of countries have next to no mentions in UN speeches. The Pearson correlation coefficient between the amount of mentions by other countries and the value of import and export is 0.14486428, which means that there is no clear correlation between these two variables. Thus, our hypothesis is rejected.

4.2 Classifying geographical regions based on speeches

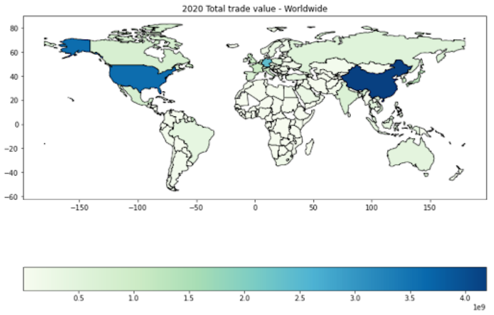

Figure 6: World map of trade value per country

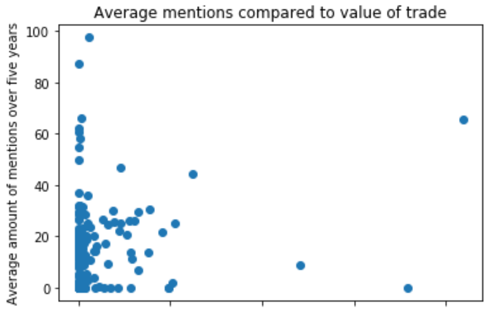

Figure 7: The relationship between number of mentions and country's trade value

The multinomial Bayesian classifier has an accuracy of 0.87 in classifying speeches in the test data to different continental regions. A visualization of this categorization is shown in the confusion matrix in Fig.8.

Adding the tf-idf weighting scheme to the pipeline for classifying speeches reduces the model accuracy to 0.44. This further consolidates our decision of not employing tf-idf in the final model.

Finally, in tuning the model parameters, it is found that the highest accuracy the model can achieve is 0.87 - the default model. It examines the speeches via a matrix of counts of single words (unigrams, ngram-range = (1,1)), and uses a smoothing parameter alpha = 1.0 for the multinomial Naive Bayes model.

5 Discussion

There's a reason that may raise the bias that different names can be used to address the same country. In our analysis, it shows that the United States of America has never been mentioned. This might be true, but would be unrealistic since they are the biggest economy in the world and have an important role in geopolitics. We suspect that the country may be referred to by other, more common names, such as the States, the USA, or the US. Fu-

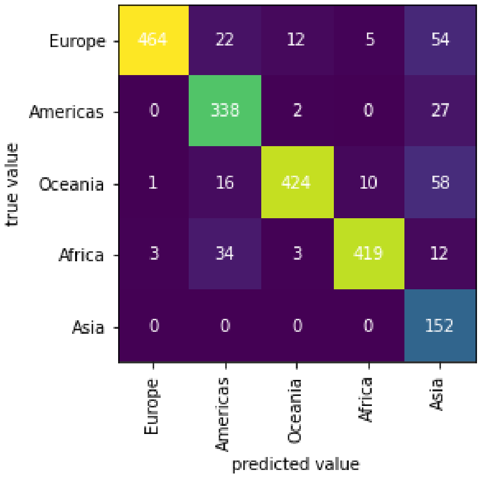

Figure 8: The Performance of multinomial Bayesian in Confusion matrix

ture analyses of these speeches should take steps to process such variations in nations' names for a more accurate analysis.

Another limitation of our current study is that in constructing the classifier we used all the data from the training set at the same time, ignoring the possibility that the debaters' language usage patterns were influenced by the time period in which they were present. In addition, the process of parameter tuning did not return a predictive model with improved accuracy in our case, and this could be because we only gave a small number of values for the parameters (3 to 4 values each), due to limited computational capacity. Therefore, in future studies, a wider range of parameter values should be given to optimize the result.

Despite its shortcomings, the paper produces novel insights. Regarding the exploratory analysis, it can be seen that the trade relationships among countries do not drive the number of times they refer to each other in UN speeches. As aforementioned, the higher number of mentions in certain years can be due to geopolitical events - a subject to explore in subsequent research. At the same time, we were able to classify speeches to different regions with good accuracy, which indicates that there may be distinct patterns in each region's way of speech. These patterns can be a topic worth investigating for future studies.

6 Conclusion

In conclusion, our exploratory analysis shows that only a few countries are mentioned a lot by other countries in UN speeches. The top three of these are Iran, Syria, and Israel in the span from 2017 to 2021. A closer analysis of the Middle East shows that the high numbers of mentions of these nations are not correlated with a higher trade value in 2020. This relationship is in fact true on a global scale; thus our hypothesis is rejected. Next to this, our predictive model was able to classify speeches of countries to continents with good accuracy, implying distinct features of speeches from different regions.

References

Gaulier, G., & Zignago, S. (2010). *Baci: International trade database at the product-level. the 1994-2007 version* (Working Papers No. 2010-23). CEPII. http://www.cepii.fr/CEPII/fr/publications/wp/abstract.asp?NoDoc=2726

General Assembly of the United Nations. (n.d.). https://gadebate.un.org/generaldebate76/en/

Jankin Mikhaylov, S., Baturo, A., & Dasandi, N. (2017). *United Nations General Debate Corpus*. https://doi.org/10.7910/DVN/0TJX8Y

UNSD-Methodology. (n.d.). https://unstats.un.org/unsd/methodology/m49/overview/