Fundamentals of Data Science: Assignment 1

Adam Walsh, Alexander Koch, Efstathios Trigkas, and Hendrik Jilderda

University of Amsterdam, 1012 WP Amsterdam, Netherlands
http://uva.nl

Abstract. In this paper we analyze the speeches of the United Nations General Assembly between 1970 and 2020. We consider if the United States became more or less vocal about war participants over the years in their UN speeches. We find that specific events influence the amount of mentions heavily making it difficult to conclude the question. However, there are some signs that it is the case. We also try to determine a connection between the wording and sentiment of a country's speech and the country's arms imports. We find that the wording can be applied generally across countries, while the sentiment analysis is only useful for select countries to predict arms imports.

Keywords: United Nations (UN) · War dataset · UN debates dataset.

1 Introduction

While the general purpose of the United Nations (UN) is to maintain international peace and security [5], war and the threat of military conflict has become more apparent in Europe with the Russian invasion into Ukraine in February 2022 and the very recent conflict in Azerbaijan. Especially with the war in Ukraine, military aid and arms imports have become a prominent topic in political circles in Europe and the US. Since the start of the Russian invasion, arms imports of European countries have surged as the countries reevaluate their national security situation [3]. The UN General Assembly provides a place for countries to voice their international security concerns. Driven by the current relevance of the topic, we analyze how members of the UN mention conflict partners and allies during times of war and peace. Furthermore, we determine how sentiment and specific wording in speeches can be used to predict a country's arms imports.
More specifically, we answer the following research questions:

- **Did the US become more or less vocal about war participants over the years in their UN speeches?**
- **Can the wording and sentiment in General Assembly speeches be used to predict arms imports of the corresponding country?**

The first research question will be answered with an exploratory analysis. In this analysis we go over multiple different datasets and conclude the question with visualizations. Because of the large amount of data we focus on the United States. The second research question will be answered by means of a predictive analysis, employing information on arms imports, a sentiment lexicon and linear regression models.

2 Methodology

The main dataset of this assignment and the answering of our two research question is the Harvard UN debates dataset [2]. This dataset contains speeches of all the UN Nations from 1946 until 2022. While during time of writing the UN meeting of 2023 has been held we decided not to include these speeches into our data. Next to this dataset the research questions make use of different datasets. These datasets will be highlighted in their respective sections of this methodology.

2.1 Exploratory Analysis

In combination with the Harvard UN debates dataset we used a WAR conflicts dataset [1]. This dataset contains two tables, Conflict participants and Participants. The first of the two tables contains a list of all the different wars over the years, reaching from 01/07/1946 until 02/08/2022. For each of these wars the start and end dates are listed in combination with all of its participants. It is important to note that the side which a country took in that war is not listed. Next to that we noticed that the end dates of the wars were not correctly listed and were just a copy of the start dates. To counter this problem we used ChatGPT [7]. The participants table mentions the same data as the first table but in a different layout. However, this table is not used in our analysis.

Before being able to execute the analysis the multiple tables had to be combined in several ways to prove useful for our research question. After having used ChatGPT to get the correct start and end dates of all the wars, we extracted all the unique war participants over all the wars from the conflict participants dataset and remove the 'United states of America'. This allows us to take this as the columns for one of our final datasets. We combine this list of countries together with all the years which had wars and made a binary matrix with all the values preemptively set to False. After that we iterate through conflicts participants dataset, in cases where a country was indeed in some way related to a conflict together with the US the value in the binary matrix is converted to True. Now that the table is filled it provides us a lookup table for the conflict participants in the correct year.

For the Harvard UN debates dataset we did not do a lot of data processing before it was useful for our research question. We imported it with help of the LAB3 aux [10]. After importing we selected only the speeches from the US and dropped all the columns except for the year and the speech itself.

Now that the correct tables have been acquired the last step consists of looping over all the speeches and counting the amount of times countries have been mentioned by the US. In cases where the mentioned country is not related to war with the US in that year the count is set to 0. This resulted in one large table with the years of speeches made by a representative of the United States (1970-2022) as the rows and the countries which at some point were at war with the US as the columns. In addition to that we also added an extra column which contains the total of the respective row.

2.2 Predictive Analysis

For our predictive analysis, we use a dataset of the arms imports of countries (or paramilitary groups) between 1950 and 2020 [6]. The dataset includes data from paramilitary groups and countries that no longer exist or have become a part of other countries today. To ensure that the data that we use is correct, we decided to remove countries and groupings that are not part of the United Nations today. To make the dataset usable for further analysis, the table also needed to be melted. To make the yearly arms imports of differently-sized countries comparable, we create a normalized table, in which the yearly arms imports of a country are divided by the total arms imports of the country over the observed period. We employ this normalized imports table in all further analysis.

To answer the question of whether or how wording and sentiment in General Assembly speeches can be used to predict arms imports of the respective countries, we employed three methods.

We started by creating a list of keywords related to war in each speech, with the goal of using the frequency of these keywords as predictors for arms imports. Our assumption is that speeches containing a higher number of keywords associated with war, arms, conflict, and similar themes may relate with increased imports of arms. In the absence of scientific sources that could provide a list of keywords, we employed ChatGPT due to its ability to create a list of keywords far more extensive than could be created manually. ChatGPT [8] created a list of 660 keywords related to these topics. After importing the keyword dataset, we computed the frequency of each word within the dataset for every UN speech and then aggregated these counts to calculate a total sum. We merged the results with the imports dataset and we were left with a table with 2 columns with every row containing the frequency of keywords and the corresponding import of arms for a country. Since we have two continuous variables (keyword frequency and arms imports in USD), the appropriate method for predicting our target value (the imports) is to use a linear regression model. For this purpose we use the LinearRegression class from the scikit-learn Python package [9] to test polynomials with up to 20 degrees. As a second method, we

employed the "term frequency - inverse document frequency" (TF-IDF) method to calculate the relative importance of the words used in the General Assembly speeches, which we then combine with the arms imports dataset to analyze the connection between certain wording in a speech and a country's arms imports. To implement the TF-IDF method we pre-processed the speeches by performing tasks such as lemmatization, removing stop words and cleaning the text of non-words or irrelevant characters. To establish a connection between speech and arms imports, we merged the TF-IDF table containing the relative importance of all words across all relevant speeches (speeches in years 1970-2017 from countries with data available for arms imports) with the arms imports dataset. Relevant speeches from the years 2018 and 2019 were kept as a testing dataset. This gives a training dataset of 2563 speeches for training and 151 speeches for prediction. To stabilize the variance of the arms imports from different countries, we perform a logarithmic transformation on the import values. Again, we use a linear regression model for the prediction.

As the third method we determined the voiced emotions for each speech individually. Our assumption is that the use of words related to feelings of anger, fear or trust may correlate with the arms imports of a country, since they relate to a country's feeling of security on the world stage. We use the NRC-Emotion Lexicon [4] which contains a list of 14154 words associated with 8 different basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust). To determine the emotions of each speech, we again pre-process the speeches as described previously and then count the occurence of relevant emotional words in each speech. To make speeches of different length comparable, we only use the share of emotional words in the speeches moving forward. Again, we try to use the share of emotional words in the speeches to predict the normalized arms imports of the countries using linear regression. Since some emotions in a speech are highly correlated (for example, the share of words signifying anger, disgust and fear typically correlate with over 0.75 across all speeches), we build our linear regression models using the bottoms-up approach. We test each emotion individually as a predictor and only keep the it if all coefficients are significant at the 5% level. Another emotion is added as a predictor if all coefficients as well as the F-statistic remain significant and the R-squared of the whole model is increased. We do this for all possible combinations of emotions and for all countries individually, keeping the model for each country with the highest R-squared. For this analysis we use the Python package statsmodels [11].

3 Results

3.1 Exploratory Analysis

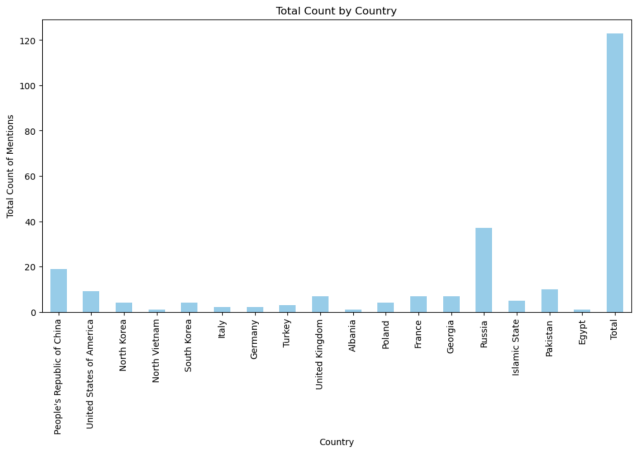


Fig. 1. This exploratory data analysis (EDA) plot highlights the trends mentions of all war participants in United States United Nations speeches. War participants are other participant in wars the United States were involved in. This is taken over a range of years(1970-2020). The plot provides insights into how often different states and entities, including the "Total" or a sum of all mentions were referenced in these speeches. By comparing the lines representing different actors, it is evident that the mentions often differ country to country. This suggests that certain countries/entities may have had more significant roles or received more attention in these speeches during certain years.

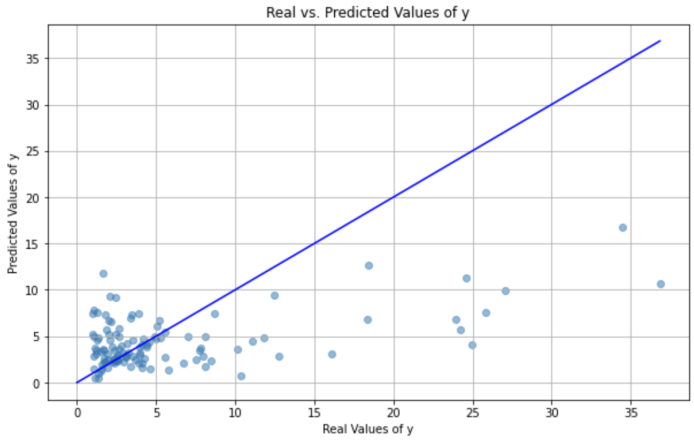
Fig. 2. This line plot presents the yearly mentions of various war participants in United States United Nations speeches. The plot reveals distinct temporal trends in the mentions of various war participants. For instance, there is a clear increase in mentions of "Islamic State" during the 2010s, indicating a period of heightened attention to this entity. Peaks and valleys in the plot lines indicate years of major changes in mentions. Researchers can delve into the historical context to understand the reasons behind these points of inflection. They may highlight the United States political motives when giving their respective UN speech each year. The "Total" line represents the combined mentions of all war participants, which provides a look at the general trends in speech content over time.

3.2 Predictive Analysis

For the first method, we use all speeches combined. A visual analysis of the data quickly showed that there was no clear relationship between the frequency of keywords and (normalized) arms imports. The linear regression confirmed this assumption. It had a mean squared error of 22.38, while the variance for the true values was 22.58. This, as well as an R-squared of 0.008, shows that the keywords can not be used in a linear regression to predict the values of arms imports.

The TF-IDF method proved more successful with the linear regression achieving an R-squared of 0.25 on the test set.

Fig. 3. real vs pred



In Figure 3 we compare normalized imports predicted by the linear regression and the real values for the test set. It is clearly visible that the predictions are more accurate for the lower real values, while the model tends to under-predict higher arms imports. Higher values for the imports are rare in both the test and train sets, which can explain this under-prediction. Nevertheless, the TF-IDF method can be used to calculate estimates of arms imports based on a given speech. The last method was successful only in select cases. Analyzing

the speeches of all countries combined proved unsuccessful, achieving R-squared values of less than 0.01. However, when analyzing countries individually, we observed large differences in the model's success. For many countries, no significant regression models could be determined. For some countries, only one predictor is significant, usually leading to R-squared values between 0.05 and 0.15. For 5 countries, namely Belgium, Brasil, Myanmar, South Africa and Zambia, linear regression models with an adjusted R-squared between 0.2 and 0.35 could be fitted. The highest amount of significant predictors is always two, although there is no common list of emotions that can be used for all countries. Nevertheless, analyzing the emotion in speeches can help estimate arms imports for select countries.

4 Discussion

4.1 Dataset limitations

As stated in the methodology we used ChatGPT [7] to counter the biggest flaw of the war participants dataset. While ideally the dataset would be complete, the use of ChatGPT made sure we could quickly continue our analysis instead of completing the dataset manually.

It is also important to note that potentially not all the wars are registered in the dataset. The primary reason for this could be the differences in definition of 'war'. A good example of this is the Cold war, this war is not mentioned in the dataset while it did have a significant impact on speeches. Since we are not qualified to determine what can be counted as a war or who can be a participant, we chose to work on the basis of the existing dataset.

4.2 Research limitations

In the Exploratory analysis we only look into the wars the countries actively participate in and are registered in the dataset. This however does not need to be the case always. An example of this is the current Ukraine-Russia war where the US and many other countries do not actively participate in but do have an influence because of weapon delivery.

As previously mentioned in the methodology the dataset also does not distinguish between allied and non-allied countries. This makes it difficult to infer any further correlation between speech mentions and 'criticism'.

The biggest risk to the validity of our exploratory research is the the code of conduct of the UN found in the UN charter [5]. this states in Article 2.4 and many other articles that "All Members shall refrain in their international relations from the threat or use of force against the territorial integrity or political independence of any state, or in any other manner inconsistent with the Purposes of the United Nations." [5]. Resulting in a large chance that that countries part of the UN, in our case the US, are softening their speeches and mentions of their conflict participants.

5 Conclusion

In conclusion, our research has cast light on the intricate relationship between UN speeches, countries at war and arms imports. The exploratory data analysis, reveals insights into temporal trends and international discourse. Variability of mentions over time was a prominent feature in the research. Specific years or events led to increases or declines in speech mentions, emphasizing the changing nature of international diplomacy. Making it difficult to conclude the exploratory question. It brings into question the United Nations' responsibility in addressing current global conflicts. This initial research served as a guide for the predictive research. The insights gained from the exploratory data analysis assisted in our understanding of diplomatic discourse, while also informing the design and interpretation of subsequent predictive research.

These observations underscored the need for more advanced analysis. Approaches such as TF-IDF and country-specific analyses were used in an effort to capture the dynamics of diplomatic discourse.

Our work contributes to the wider dialogue on the role of diplomacy in shaping global peace initiatives. Its intention is to provide a foundation for future research investigations into evolving relationship between diplomatic discourse, active conflict and the international arms trade.

References

1. Barash, G.: WAR! conflicts and nations who took part in them — kaggle.com. https://www.kaggle.com/datasets/guybarash/war-conflicts-and-nations-who-took-part-in-them, [Accessed 24-09-2023]
2. Dasandi, N.: United Nations General Debate Corpus 1946-2022 — dataverse.harvard.edu. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/0TJX8Y, [Accessed 23-09-2023]
3. Institute, S.I.P.R.: Surge in arms imports to europe, while us dominance of the global arms trade increases. https://sipri.org/media/press-release/2023/surge-arms-imports-europe-while-us-dominance-global-arms-trade-increases (2023), [Acessed 02-10-2023]
4. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon. Computational Intelligence **29**(3), 436–465 (2013)
5. Nations, U.: Chapter I: Purposes and Principles (Articles 1-2) — United Nations — un.org. https://www.un.org/en/about-us/un-charter/chapter-1, [Accessed 30-09-2023]
6. Oh, J.: Arms Imports Per Country (1950-2020) — kaggle.com. https://www.kaggle.com/datasets/justin2028/arms-imports-per-country, [Accessed 24-09-2023]
7. OpenAI: Chatgpt: A large-scale generative language model. https://www.openai.com/research/chatgpt (2021), prompt: Provide a list of end dates to the following conflicts, if still ongoing, insert 2022 as the end date
8. OpenAI: Chatgpt: A large-scale generative language model. https://www.openai.com/research/chatgpt (2021), prompt:Create a dataset with x words related to war, arms and conflict
9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
10. Santos, F.: $Lab3_assignment1_aux--.ipynb, accessed25-09-2023$
11. Seabold, S., Perktold, J.: statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference (2010)