Predicting Perceptions of Government Corruption Using Sentiment Analysis of UN General Debate Speeches

Matej Lončarić, Ana Negru, Alvaro Millan Ruiz, and Klara Willberg

Abstract. Sentiment analysis is an increasingly popular automated method that provides the user with classification of data into positive, neutral, or negative sentiment. This paper intends to explore whether it is possible to build a model that predicts perceptions of government corruption based on the sentiment analysis output from the United Nations General Debate Corpus (UNGDC). Oftentimes, sentiment analyses have been conducted on social media content – this paper further builds upon existing research by applying the method to longer documents. Results indicate that building a model predicting perceptions of government corruption based on sentiment output of UNGDC is possible, however the inclusion of sentiment did not improve predictions in our context.

Keywords: UN General Debate, World Happiness Report, Perceptions of Corruption, Sentiment Analysis.

1 Introduction

Since 1946, representatives of the UN member states have come together annually for the United Nations General Assembly (UNGA), which is comprised of all 193 members of the United Nations. The focal point of the UNGA is the General Debate (UNGD), during which Heads of State and Government are invited to deliver and discuss world issues in global politics through their perspectives (United Nations, 2022). The statements delivered during the UNGD often reveal valuable insights into policy preferences regarding a wide range of topics. The most discussed topics include terrorism, nuclear non-proliferation, development and aid, and climate change”– all which share a mutual concern among the member states (Baturo et. al, 2017). Based on the range of topics discussed, we expect that sentiment – referring to whether a speech is regarded as positive, neutral, or negative – will vary across speeches depending on the relevant topic. Moreover, sentiment analyses are increasingly popular; “sentiment analysis, or opinion mining, is an active area of study in the field of natural language processing that analyses people's opinions, sentiments, evaluations, attitudes, and emotions via the computational treatment of subjectivity in text" (Hutto & Gilbert, 2014, p. 217).

The World Happiness Report is an annual collection of country-wise surveyed statistics regarding variables such as overall happiness, confidence in national government, social support, and perceptions of government corruption (to name a few). Our research paper will focus on perceptions of government corruption specifically, and whether the sentiment of UNGD speeches may influence such perceptions. To research the possible effects of sentiment is relevant on both a scientific and societal perspective – it gives an indication on how positive/negative tones may affect society and adds to existing literature by expanding the length of the analysed textual data. Thus, we propose the following research question: Is it possible to build a model to predict perceptions of government corruption based on the sentiment analysis output from the United Nations General Debate Corpus (UNGDC)?

2 Methodology

Answering the research question required the use of two datasets. The first dataset comprised UN General Debate Speeches from 1970 to 2020, as collected by Jankin et al. (2017). The dataset was further appended with all speeches from 2021 which were translated to English (United Nations, 2022). The second dataset was the 2022 World Happiness Report, as collected by Helliwell et al. (2022).

Some preprocessing was required on both datasets to meet the requirements for answering the research questions. First, a new variable *Regions* was created in the United Nations General Debate Corpus (UNGDC). It directly corresponded to the variable *Region Name* from the original dataset, apart from observations from the Americas, which were split into *North America* and *Latin America*. This was done to account for the economic, social, and historical differences between the two regions while retaining the sub-regional division proposed by the United Nations (Jankin et al., 2017). Moreover, in line with the recommendations by Zhou (2019), other preprocessing such as removing stop words or lemmatizing the speeches was not conducted as to avoid potential impairments in the sentiment analysis. Second, where possible, missing values from the World Happiness Report were imputed by the respective country's mean values. Otherwise, they were replaced by 0 to permit the predictive analysis.

The exploratory data analysis of the UNGDC was carried out through VADER's *SentimentIntesityAnalyzer* module (Hutto & Gilbert, 2014), as implemented by the natural language toolkit (NLTK) (Bird et al., 2009). The lexicon uses a standard list of lexical features to assign a sentiment classification; namely, positive, negative, or neutral, combined with five general rules representing grammatical and syntactical conventions for expressing sentiment intensity to assign an intensity score to each input text (Hutto & Gilbert, 2014). The resulting compound scores were used to produce visualizations depicting changes in sentiment trends for each of the six regions defined in the previous paragraph. Moreover, the compound scores were later extracted and used as features for predicting perceptions of government corruption. While the use of VADER in this context raises important limitations, namely because the model is tailored for analyzing the sentiment of micro-blog-like posts on social media, we have not found a suitable alternative providing numerical sentiment intensity scores.

On the other hand, the exploratory data analysis of the World Happiness Report involved creating scatter plots and correlation heatmaps of all features from the report against the variable of interest, *Perceptions of corruption*. Afterwards, a stepwise regression with backward elimination was conducted to estimate the relevant features from the Happiness Report in predicting corruption. Next, the two datasets were merged, and a linear regression model was trained using 80% of the observations from the joint dataset, in line with common practice. The training included five-fold cross-validation, as well as testing for second-degree polynomial features, normalized data, and intercept fitting. Linear regression was chosen because the goal was not to develop a complete model for predicting perceptions of government corruption, as they are likely also influenced by factors excluded from the Happiness Report, but to estimate whether the sentiment of speeches in the UN General Debate can be used to gather insight into the political climate of participating countries. Finally, model performance was evaluated using estimates for the mean squared error (MSE) and the R^2 coefficient of determination. If the inclusion of the compound sentiment score would improve the predictive power of a linear model estimating the perceptions of government corruption, one would expect to observe a lower MSE and a lower value for the R^2.

3 Results

By visualizing VADER's *SentimentIntesityAnalyzer*, results revealed that Africa, Europe, Oceania, and Latin America (see Appendix 1) have continuously had compound scores above 0.65, indicating overall positive sentiment. It should however be noted that the compound scores have fluctuated along the timeline – especially around the new millennium.

Asia's compound scores (see Appendix 1) were in general positive, with compound scores approaching neutrality around the millennium – similarly to the aforementioned regions. Moreover, both Asia and Latin America seem to vary considerably in relation to compound scores and thus, sentiment.

Finally, North America (see Appendix 1) reveals some interesting visuals. Prior to the end of the 1990's, North Americas compound scores were almost perfectly (and consistently) positive. After the millennium, the compound scores decreased and reached the lowest of all scores among the six regions. Within twenty years, the compound score was as low as zero on three different occasions. However, between these three instances, the sentiment increased back to highly positive. An explanation behind the drop in compound scores around the millennium for most of the regions could be explained by the 9/11 terror attacks in 2001. As mentioned earlier, terrorism is a highly prevalent topic in the UNGD – therefore, the compound scores could have been influenced by the event.

Next, the World Happiness Report data was analyzed. Figure A in Appendix 2 visualizes the relationships between all variables in the dataset. Most of the variables show no clear relationship with *Perceptions of corruption*, the variable of interest. The correlations between *Perceptions of corruption* and the other variables are shown in Figure B. On the one hand, the variables *Health expectancy life at birth*, *Confidence in the national government*, *Social support*, *Positive affect* and *Generosity* have a weak negative correlation with *Perceptions of corruption*. These correlations lie between -0.13 and -0.25. Moreover, the variables *Life ladder* and *Freedom to make life choices* have a moderate negative relationship with *Perceptions of corruption*, while *Negative affect* has a weak positive correlation with the variable.

Following the exploratory data analysis, a step-down regression was performed to evaluate whether the variables from the World Happiness Report have explanatory power in predicting the variable of interest, *Perceptions of corruption*. The model presented an adjusted R^2-value of 0.33 and all the variables, besides *Log GDP per capita* and *Healthy life expectancy at birth*, had p-values lower than 0.001. The former, having a higher p-value was therefore removed from the model. The subsequent model showed no changes in the adjusted R^2, but *Healthy life expectancy at birth* showed a p-value above 0.05 and was therefore removed from the model as well. The final model included all the other variables from the World Happiness Report.

The model was first trained using all of the variables which had explanatory power over the *Perceptions of corruption*. After five-fold cross-validation, the Grid Search found that the best model contains no polynomial features, fits the intercept, and normalizes the data. Model validation estimated the mean squared error to 0.03, while the variance score equaled 0.30. Ultimately, the Vader sentiment scores were added to the model, and the process was repeated. The best-performing model used non-normalized data and had an increase in the variance score of 1 percentage point.

4 Discussion

As the analysis above has revealed, adding the compound score *VADER* compound, based on sentiment analysis does not attribute significant predictive power to the regression model, as compared to the model without the sentiment scores constructed solely using Happiness Report dataset variables. This was indicated by a minimal 1% change in the R^2 and the insignificance of the *VADER compound* variable (p-value of 0.361). Thus, to answer the research question, yes, it is possible to build a model to predict perceptions of government based on the sentiment analysis output from the United Nations General Debate Corpus (UNGDC). However, the sentiment analysis component does not improve the predictive power of such a model.

This may be explained by further examining the nature and purpose of the data obtained from the UNGDC. GD statement speeches are delivered using a “strategic signaling” tactic, to indicate certain political preferences and influence other states to further their own agendas (Baturo et al., 2017). A main purpose of the GD is cited as "providing governments with the opportunity to influence international perceptions of their state, aiming to position their states favorably, as well as to influence the perception of other states". Bearing this in mind, it is not possible to assume that the sentiment of all speeches in the dataset is an organic or accurate representation of the political state or views of actions/statements at a government level of every state that has given a speech at the GD. Hence, the desire of states to depict themselves in a way that does not necessarily represent reality may be a reason as to why the *VADER compound* variable proved to be statistically insignificant in the context of the constructed model.

On the other hand, there are certain noteworthy limitations to this study worth mentioning. In relation to the scatterplots of the variable of interest *Perceptions of corruptions,* against all variables in the Happiness dataset, the plots may at times show seemingly contradictory results. For instance, examining the plot between *Perceptions of corruption* & *Social support* shows that high values of perception of corruption are also associated with higher values of social support. There are similar contradictory relations between other predictor variables in the Happiness dataset as well. When rationalizing this, it is important to consider that this comparison between each pairing of variables is averaged across all countries, and what the visualizations are showing is essentially a compound of that. However, it would be cumbersome to make such a plot for every country (or even region) individually, hence this is a limitation of the visualization portion of the EDA for this study.

Furthermore, this study made use of backward elimination in stepwise regression. In relation to this, Smith (2018), claims that deciding model explanatory variables based on R^2 or statistical significance is one of the greatest issues with stepwise regression, and a reason why researchers should reconsider using it. Following automated rules based on statistical correlations without considering whether it makes sense to include a specific variable may lead to nuisance variables being included in the model. Thus, for this study, terminating the step-down regression process with 7 remaining variables that had significant explanatory power may have led to a nuisance variable being included in the model. Smith (2018) suggests using recursive feature elimination with cross-validation or regularized tree ensembles as suitable alternatives to stepwise regression. Therefore, scope for future research lies in using these alternative methods to conduct this study to test whether different results would be obtained.

Finally, in reference to the methodology section, it was specified that the scope of this paper was to estimate whether the speech sentiment component can be used to gain insight into the political climate of participating countries, focusing on perceptions of corruption, not a complete model for general prediction of government corruption. Further datasets with factors directly influencing opinion on national government may be incorporated into this study and used in combination with data used here to develop a more complete model aimed at specifically predicting government corruption perceptions.

5 Conclusion

In conclusion, it is possible to build a model predicting perceptions of government corruption based on the sentiment analysis output from the United Nations General Debate Corpus (UNGDC). However, as mentioned before, the compound score did not improve the model, and therefore was rather insignificant to our results. As mentioned, this may be due to member states or representatives avoiding pushing their agenda or opinion too far to not appear too forthcoming. Consequently, VADER's *SentimentIntesityAnalyzer* may be more suitable for corpora comprised of more politically divisive text.

References

1. Baturo, A., Dasandi, N., & Mikhaylov, S. J. (2017). Understanding state preferences with text as data: Introducing the UN General Debate corpus. *Research & Politics*, *4*(2). https://doi.org/10.1177/205316801771282
2. Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.
3. Helliwell, J. F., Layard, R., Sachs, J. D., De Neve, J.-E., Aknin, L. B., & Wang, S. (2022). *World Happiness Report 2022*. Sustainable Development Solutions Network. Retrieved 1 October, 2022, from https://worldhappiness.report/ed/2022/
4. Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 216-225. Retrieved 1 October, 2022, from https://ojs.aaai.org/index.php/ICWSM/article/view/14550

5. Jankin Mikhaylov, S., Baturo, A., & Dasandi, N. (2017). *United Nations General Debate Corpus* (Version 6) [Data set]. https://doi.org/10.7910/DVN/0TJX8Y
6. Smith, G. (2018). Step away from stepwise. *Journal of Big Data, 5*(1). Retrieved 1 October 2022, from https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0143-6
7. United Nations (2022). *General Debate of the 76th Session*. Retrieved 1 October, 2022, from https://gadebate.un.org/generaldebate76/en/
8. Zhou, M. (2019, July 23). *StopWords and Lexicon Normalization for Sentiment Analysis*. Medium. Retrieved 1 October, 2022, from https://medium.com/data-science-blogs/stopwords-and-lexicon-normalization-for-sentiment-analysis-f9f10f0d4108

Appendix 1

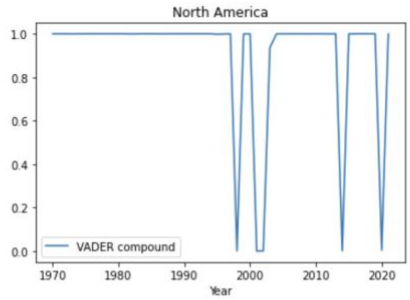


Fig. 1. Sentiment Scores in North America

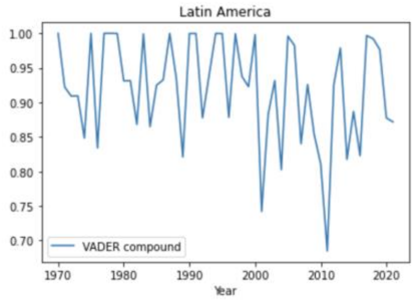


Fig. 2. Sentiment Scores in Latin America

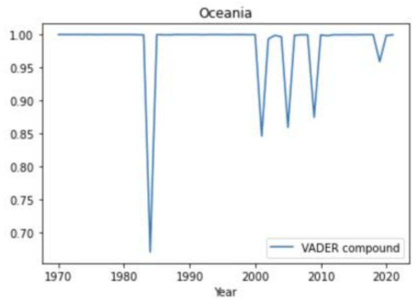


Fig. 3. Sentiment Scores in Oceania

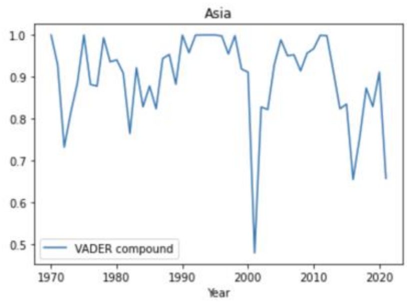
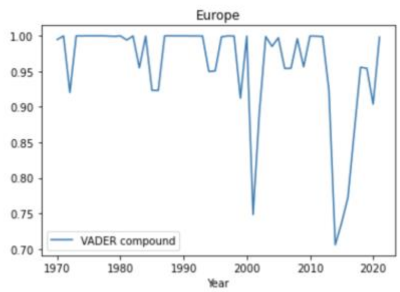


Fig. 4. Sentiment Scores in Asia



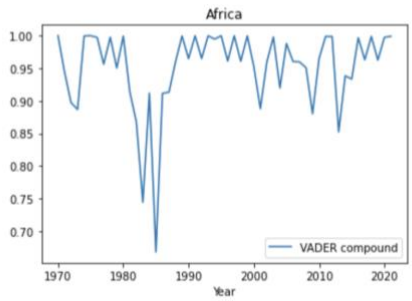
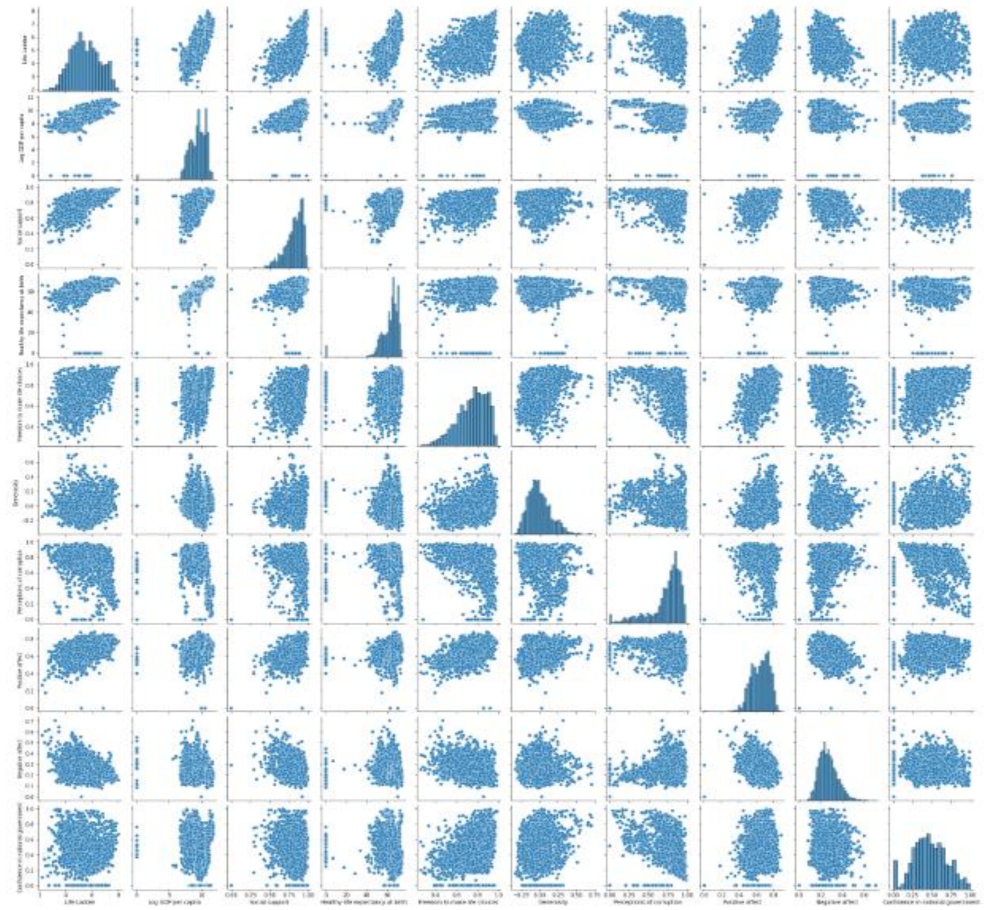


Fig. 5. Sentiment Scores in Europe **Fig. 6.** Sentiment Scores in Africa



Appendix 2

Fig. A. Plot of relationships in the World Happiness Report

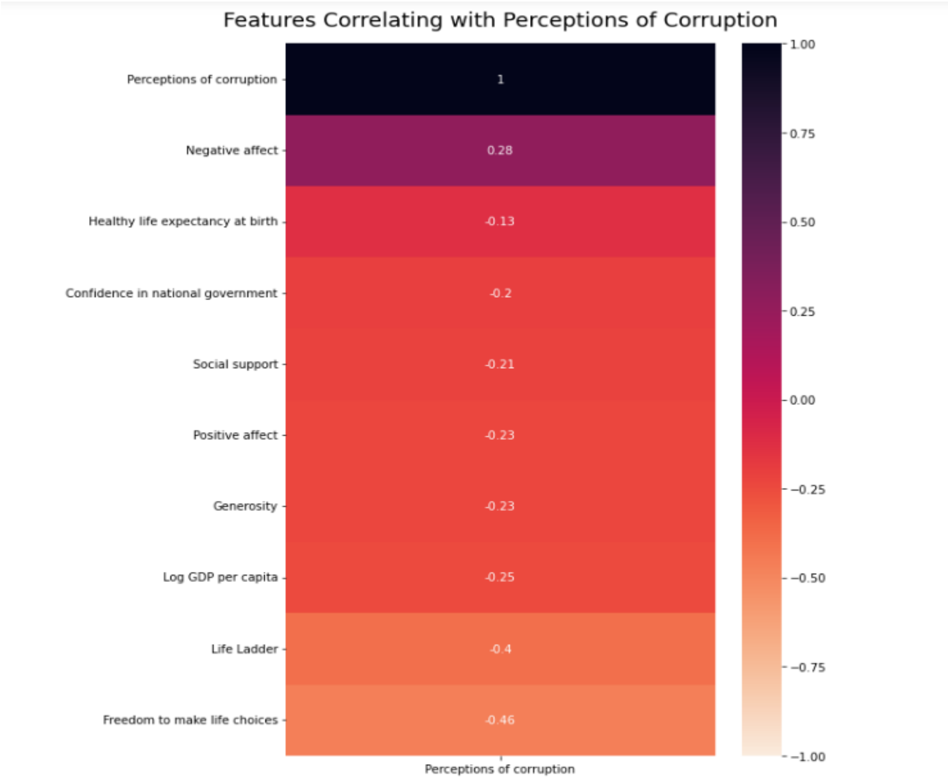


Fig. B. Features correlating with *Perceptions of Corruption*