Exploring UN speeches and their relation to Life Ladder

Francisco Pereira[13911376], Gijs Gubbels[11408707], Dragos Pop[14007584], Lan Chu[13966618], and Robin van den Berg[11317345]

University of Amsterdam

Abstract. The UN speeches provide valuable information regarding the political landscape of each country. In this work, we analyse the speeches from each year and match them with the World Happiness Report. We provide evidence of how events over the years impact the sentiment of the speeches and the most relevant words are changed by them by focusing on the Syrian Civil War. The results of the sentiment analysis show that the negative sentiment increases at the beginning of the civil war. Also, a Light Gradient Boosting Machine (LGBM) was used on attempts to make regression on the 'Life Ladder' index based on the most important features of speeches obtaining an R2 score of 33%. The results from the LGBM show that the features were not informative enough to estimate the life ladder of a country. However, a LGBM classifier was used to predict the region of origin from the speeches, and a model with 73% of precision was obtained. This means that the classifier was able to correctly describe the region using a speech.

1 Introduction

Since 1946, spokespersons from every United Nations (UN) member state assemble yearly and give speeches on their government's outlook with regards to primary political affairs during the General Debate (UNGD). As such, the speeches provide valuable information regarding the political landscape during the years. The speeches from 1970 until the present were collected by [1] and made publicly available. The digitised version of the speeches allows for computational analysis and, as such, could provide opportunities for quantitative comparison of political perspectives both on the spatial as well as the temporal dimension. Through analysis of the speeches, we aim to answer the following research questions:

1. **Research Question:** Is it possible to uncover important events based on word usage and is this evident from the sentiment of the speeches? As an example, we investigate the specific case of the Syrian Civil War.
   Hypothesis: The negative sentiment (NS) in Syrian speeches is increased during The Syrian Civil War and the most characterising words are related to war.
   Approach: Word importance was characterised using TF-IDF and the sentiment of the speeches was uncovered using VADER. We expect that deviations in speech sentiment will often occur as a result of important events.
2. **Research Question:** Can we infer the happiness level of a country from its speech?
   Hypothesis: The characteristics of the speeches, such as sentiment and relevant words, will allow us to roughly estimate the life ladder (LL) of a country.
   Approach: A Light Gradient Boosting Machine was used in an attempt to predict LL given in the World Happiness Report (WHR) given a speech's characteristics.
3. **Research Question:** Can we infer the region of origin from a speech?
   Hypothesis: Inferring the origin region from the most characterising words and other features of the speeches lead to accurate results.
   Approach: A similar setup to 2 was used to predict the origin countries of different speeches. A multi-object classifier was used for this purpose.

In this report, the methods will be discussed first. An overview of the data set is given, after which the exploratory data analysis process is described and followed by the statistical learning methods used. Subsequently, the results of the methods are visualised and discussed in the results and discussion section. Also in this section, the limitations of our analyses are acknowledged and possible follow-up steps are suggested to further explore this subject. We conclude by revisiting the hypotheses.

2 Methodology

In this section, an overview of the datasets is given, after which exploratory data analysis will be described. The set-up of the statistical learning methods is also outlined.

2.1 Datasets

In this report, two datasets were used. The first dataset entails the speeches given during the GD of the UN from 1970 to 2020 [1]. The second dataset contains data from the WHR from 2005 to 2020.

The UNGD dataset contains the speeches per year per country. During the years, the countries present in the dataset change as a result of the increase in UN member states and the varying country composition, e.g. the separation of one country into two or vice versa. The WHR report features information on 'Life Ladder', 'Log GDP per capita', 'Social support', 'Healthy life expectancy at birth', 'Freedom to make life choices', 'Generosity', 'Perceptions of corruption', 'Positive affect' and 'Negative affect'. The columns contain respectively 0%, 1.8%, 0.7%, 2.8%, 1.6%, 4.6%, 5.6%, 1.1%, and 0.8% missing values. Next to the missing values from the existing rows, the countries present per year also differ. A summary of the numerical data is given in table 1.

Table 1: Description of the numerical columns in the data.

	year	positive sentiment	negative sentiment	Life Ladder	Log GDP per capita	Social support	Healthy life expectancy at birth	Freedom to make life choices	Perceptions of corruption
count	8498	8498	8498	1915	1880	1903	1875	1886	1810
mean	1997.3	0.169855	0.080484	5.461753	9.356195	0.812104	63.358227	0.742588	0.750282
std	14.3	0.029006	0.027136	1.121065	1.151276	0.119182	7.524648	0.141947	0.184048
min	1970	0.056	0.003000	2.375092	6.635322	0.290184	32.299999	0.257534	0.035198
25%	1986	0.15	0.061000	4.633863	8.459277	0.747869	58.680000	0.647427	0.694766
50%	1998	0.169	0.078000	5.374446	9.456262	0.835527	65.199997	0.763883	0.803661
75%	2010	0.189	0.097000	6.290220	10.326396	0.906053	68.599998	0.855859	0.873493
max	2020	0.294	0.268000	8.018934	11.648169	0.987343	77.099998	0.985178	0.983276

2.2 Exploratory Data-Analysis

Cleaning and Preprocessing To make the speech data suitable for analysis, we preprocessed the data to various extents depending on the extracted features. For determining the sentiment of the speeches, we consider the unaltered speeches as it takes into account punctuation and word order within a sentence. Average sentence length was also directly deduced from the unaltered speech data. Before taking a word count of the speech, the speech was tokenised and non-alphabetic characters were removed. To prepare for the extraction of the characterising words from the speeches using TF-IDF, the speeches were split up into separate sentences after which the words were lemmatised. This was implemented per sentence as the function of the words within the sentence is important for the lemmatisation. Subsequently, all non-alphabetic characters were removed as well as the stop words present in the text.

The next step is to relate the WHR to the UNGD speeches. As the WHR uses country names while the UNGD dataset uses Iso Alpha-3 Code, a third dataset, UNSD, containing both country names and Iso Alpha-3 Codes was used to relate the two

datasets. Inconsistencies were observed between the codes used in the UNGD and UNSD dataset as well as the country names used in the WHR and UNSD dataset. The inconsistencies were manually resolved to facilitate alignment and prevent information loss.

After the full data set was obtained, we tended to the missing data in WHR. Linear Interpolation (LI) was employed to fill in the missing values, where LI was chosen because of two reasons: (1) WHR consists of time-series. This allows us to make the assumption that a missing value will be between the prior and subsequent point; (2) considering that the happiness metrics values per country are closely spaced across years, linear interpolation is accurate enough. A first limitation is that LI won't be able to approximate a missing value when the first or last available values are missing, as the data does not show consistent trends rendering linear extrapolation insufficient, a second disadvantage is that LI does not capture possible outliers. Considering that the WHR data does not often miss data at the extremes and volatility is relatively contained, both limitations are acceptable for this specific case.

Analysis and Visualization After preprocessing the dataset, several features were extracted from the text. First, speech length and average sentence length were determined. Furthermore, the most characterising words were determined using TF-IDF. This method was applied to compare both speeches across one year, as well as the characterising words across the years by looking at the speeches of one year as one document. Lastly, the sentiment was extracted using VADER. All features were then cross-correlated to examine the relations. In this correlation analysis, the LL data available in the WHR was also taken into account. After the extraction of the features, the sentiment of the speeches was searched for deviations. At the point of these deviations, the characterising words for that year were examined in an attempt to elucidate a possible cause of the deviation. The words were then visualised through word clouds. Lastly, scatter plots were used to study the relationship between sentiments in speeches and happiness metrics in detail.

2.3 Statistical Learning - Speech Origin and LL prediction

Having examined a set of characteristics of the speeches, we explore the possibility to predict the speech's origin region and the LL from the characteristics of speech through supervised learning. To this end, we utilise the Light Gradient Boosting Machine models for classification and regression, respectively. This method was chosen as it provides fast training speeds with efficient memory usage and it provides better accuracy than other boosting algorithms due to the leaf-wise split approach versus a level-wise split approach since it creates more complex trees. As a result, the algorithm is more prone to overfitting, especially when compared to other bagging algorithms such as random forests, which is taken into account in the optimisation process [2].

The process of optimisation and evaluation for the regression and classification are similar. However, for the regression, only samples for which the LL in WHR is available can be used, while for the classification, the region, i.e. target variable, is available for all speeches. After selection, we divide the dataset into 80% training and 20% test set. We evaluate the performance of the base regressor and classifier through 5-fold cross-validation (CV) on the training set. Subsequently, we perform a randomised search using 5-fold CV to obtain the best hyperparameters using the training partition. Lastly, we evaluate the performance of the models on the test set. To combat overfitting an 80/20 training/validation split is used, where the fitting is halted when the performance on the validation set doesn't increase for 50 rounds.

As features, we use the sentiments extracted from the speeches, the word count, the average sentence length and the year of the speech. Furthermore, the most characterising words per speech, previously extracted with TF-IDF, are used through OneHotEncoding. For the latter, a selection was made of words that occurred at least in 20 of the samples in order not to fit the model on single occurrences. An additional reason for this choice is countries referring to themselves. By setting a threshold of 20, we remove the cases where countries were only mentioned by themselves (max. 15 speeches of each country). This yielded a list of 147 words to be included as features.
Additionally, in the case of the multiclass classification of speech' origin region, the number of records from each of the five regions was counted and it was discovered that there are substantially fewer observations from Oceania, namely 28, in comparison with the others who have over 300 instances each. Therefore, the 28 speeches from Oceania were removed in order to deal with class imbalance.

3 Results and Discussion

Evolution of most characterising words over time in relation to speech sentiment and WHR metrics Figure 1 gives an example of how most characterising words (with highest tf-idf score) in speeches of Syria and the United States evolve over the course of 11 years, starting in 2010. As can be seen, for the US, while some words including "peace", "human", "security" are constantly repeated over time, there seems to be a decrease in the frequencies of terms such as "war", "nuclear". Meanwhile "peace", "security", "war", "support", "resolution" are most frequently used and high-weighted words from speeches of Syrian leaders.

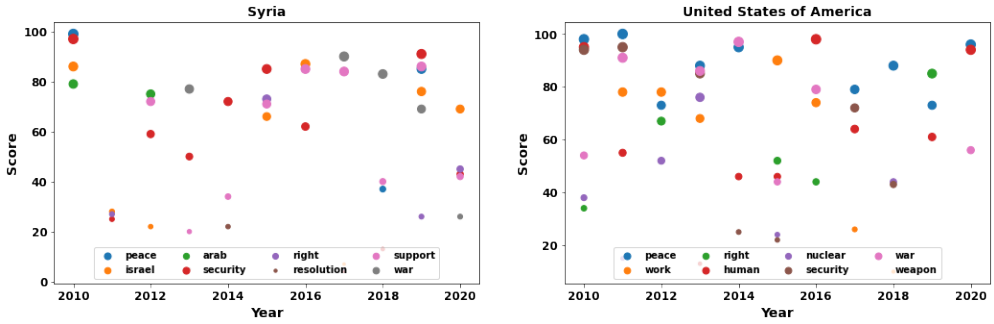

Fig. 1: Evolution of most characterising words over time.

Having a closer look at Syria as an example, we explore how the sentiment in the country's speeches evolves over time and whether there is any impact of the civil war on speech's sentiments and happiness metrics. From figure 2, a stronger NS is evident in speeches of Syrian's leaders after the civil war started in 2011. As seen in figure 3, there seems to be a relation between war and LL. After 2011, we see stronger NS and a lower LL. The years before 2011 witnessed lower NS and a higher LL. We can conclude that, in this case, the events in Syria do manifest in the characteristic of the speech and that an image of the countries state was computationally retrievable.

The word cloud of Syria in 2011 (fig.4) let us reflect back then the political and social crises the country faced right before the war. A strong demand for political reforms and rejection of foreign interventions are well illustrated in the word cloud. In 2014, Syrian's speech focused on the threat of terrorism, extreme ideology and its commitment to eliminate chemical weapon as part of OPCW-UN joint mission in Syria (fig.5).

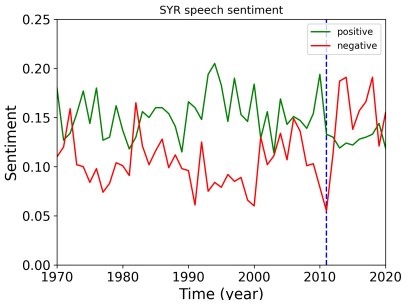

Fig. 2: How sentiment in Syrian's speeches develop over time, before and after civil war

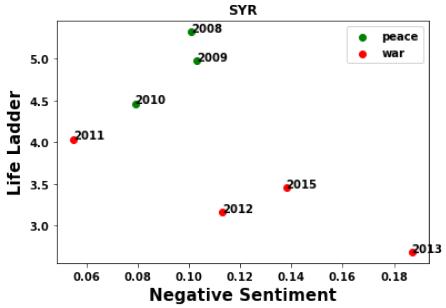

Fig. 3: Speeches negative sentiment and Life ladder in Syria

Fig. 4: Syria's Word Cloud, 2011

Fig. 5: Syria's Word Cloud, 2014

Predicting LL We elaborate on the results from predicting LL based on the characteristics of the speech. Firstly, we optimise the LGBM regressor through random search and find the following parameters to function the best: subsample: 1, reg_lambda: 0, reg_alpha: 0, num_leaves: 12, n_estimators: 2000, min_child_samples: 1, max_bin: 255, learning_rate: 0.01, colsample_bytree: 0.64, boosting_type: 'gbdt'. Examining the accuracy of the regressor, we apply the model to a test set, the results of which are given in table 2. Table 2 illustrates the increase in accuracy after optimisation of the model. However, the current R2 score is indicates that only $\approx 33\%$ of the variance in LL is explained by the current features. The gain over each iteration is visualised in figure 6, demonstrating that the performance on the validation set doesn't increase after approximately 1000 iterations. At this value, we find the optimal trade-off between variance and bias, where further training does not improve the predictive capabilities of the model and would unnecessarily increase its complexity. Figure 7 shows the feature importance given as the number times the parameter is used in the model. It is evident that the sentiments and the length of the texts are the most descriptive.

Predicting speech origin region Following a similar approach as in the case of predicting the LL, the best hyperparameters of the classifier are found through random search. These are; subsample: 0.75, reg_lambda: 0.5, reg_alpha: 0, random_state: 42, num_leaves: 4, n_estimators: 3000, min_child_samples: 10, max_bin: 510, learning_rate: 0.005, colsample_bytree: 1, boosting_type: gbdt.

Accordingly, the precision of the model on the test data improves from 0.67 to 0.73 after the hyperparameters are tuned. When it comes to the gain over each iteration, it highly resembles figure 6, while the figure 8 shows the feature importance of the classifier. As one can see, the length and sentiments of the speech have the most predictive power. Lastly, the figure 9 represents the confusion matrix, with true labels on the y-axis and predicted regions on the x-axis.

The analysis shown in this section can be extended in multiple ways. Firstly, to better investigate the relationship between speech sentiment and events, the analysis should be extended to other cases to confirm the hypothesis. Secondly, to improve the predictions of the models, other features should be extracted from the speeches, such as the type of greeting or the mentioned subjects found through clustering.

Table 2: R2 score for unoptimised (UO) and optimised (O) regression. 5-fold CV score is mentioned with the standard deviation.

R2	5-fold CV	Test set
UO	0.20 +- 0.05	0.23
O	0.48 +- 0.13	0.33

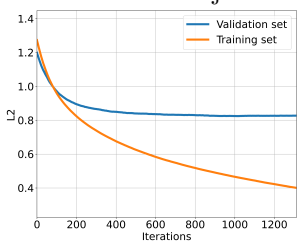

Fig. 6: L2 scores over the iterations.

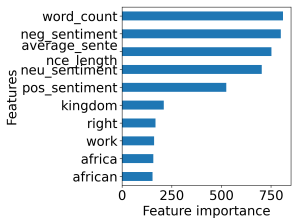

Fig. 7: Feature importance.

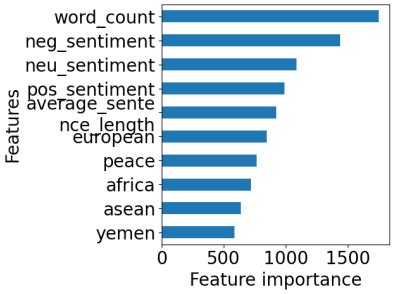

Fig. 8: Feature importance classification

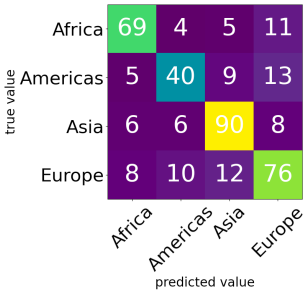

Fig. 9: Confusion matrix

4 Conclusion

In this section, the initially established hypotheses are revisited. We provide an answer based on our analyses and summarise the main key-points of our work.

- **Hypotheses 1:** The negative sentiment (NS) in Syrian speeches is increased during The Syrian Civil War and the most characterising words are related to war.
  Answer: From the analysis it was possible to prove that there was an increase of the NS of the speeches from 2011 (start of the civil war). By visualising the most relevant words from the Syrian speeches before and after the war, it was proven that in fact these were related. Therefore we accept our hypothesis.
- **Hypotheses 2:** The characteristics of the speeches, such as sentiment and relevant words, will allow us to roughly estimate the life ladder (LL) of a country.
  Answer: From the results obtained from the LGBM, we conclude that the features extracted from the speeches were not informative enough to infer the LL index. We can not accept the hypothesis.
- **Hypotheses 3:** Inferring the origin region from the most characterising words and other features of the speeches lead to accurate results.
  Answer: The LGBM classifier was able to correctly infer the region of the speech from most of the countries. A precision of 0.73 was achieved, therefore, we accept our hypothesis.

With this work we explored the speeches from the U.N. since 1970 to 2020 and related them to the World Happiness Report. We performed Exploratory Data Analyses and applied Statistical Learning methods to solve the questions of interest.

References

1. Baturo, A., Dasandi, N., Mikhaylov, S.J.: Understanding state preferences with text as data: Introducing the un general debate corpus. Research & Politics **4**(2), 2053168017712821 (2017). https://doi.org/10.1177/2053168017712821, https://doi.org/10.1177/2053168017712821
2. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: NIPS (2017)