Text

Description automatically generated with medium confidence

ISSS609 Project Proposal

Text Analysis of   
Singapore National Day Rally Speeches

Team 2

# Team Members

|  |  |  |
| --- | --- | --- |
| Name | ID | Email |
| Chua Yan Ting (Amelia) | 01244864 | yantingchua.2021@mitb.smu.edu.sg |
| Jim Tan Jun Ze | 01259396 | jim.tan.2021@mitb.smu.edu.sg |
| Thng Corryne Nataline | 01252902 | corrynethng.2021@mitb.smu.edu.sg |
| Wesley Djingga | 01433111 | wesleyd.2021@mitb.smu.edu.sg |
| Zhang Yi | 01444133 | yi.zhang.2021@mitb.smu.edu.sg |

## Introduction

The National Day Rally (NDR) is an annual address delivered by the Prime Minister of Singapore. Since 1966, it has served as a platform to review the country’s progress in the past year and share the direction for the future (Min, 2016). In these speeches, leaders have defined the challenges faced by the nation, the goals set to transform the country as well as the strategies and policies that have shaped Singapore (National Archives of Singapore, n.d.). These speeches injected a sense of patriotism and Singaporeans are usually reminded to set their differences aside and unite as Singaporeans, to help the nation to advance (Seng, 2014). With this as the backdrop, the main motivation of the project is to analyse the speech content and derive valuable insights.

## Purpose / Objectives

Given a dataset of NDR speeches, our team wishes to identify and retrieve those speeches where a certain topic was covered. For example, the user may be interested to know which speeches covered the topic of housing. After retrieving these documents, the user can perform further analysis such as understanding when the topic of housing was first brought up and how the situation evolved over time. Therefore, the first objective of this project is to allow for efficient document retrieval in the dataset.

Political Speeches are usually long and span various topics. Citizens may wish to get a gist of the subject matters covered in the speech without going through the entire speech. This is especially the case when there are multiple speeches involved. Therefore, the second objective of this project is to identify the key topics in a political speech. Having identified the key topics in a given speech, our team endeavours to perform a comparison of speeches across time. This provides insights into how the topics of concern have evolved throughout Singapore’s development and growth as a nation.

Apart from identifying the key topics in a speech, it is important to understand the general sentiment of the speech. This provides further insights into how the speech was delivered, the emotions it wishes to stir in its audience and the general climate during that period. Consequently, the third objective of this project is to perform sentiment analysis on the political speech. In addition, for speeches that are given in the era of social media, our team will perform sentiment and emotion analyses on social media texts (e.g., Tweets from Twitter) to understand the public’s response towards the speech. This would provide insights into whether the sentiment of the speech matches the public’s sentiments and provides the speaker with an idea of how the speech was received by the public.

## Literature Review

Miranda and Bringula (2021) studied the annual State of Nation Address of 13 past Philippine presidents to determine the sentiments and emergent topics of the speeches. Using the National Language Toolkit’s sentiment intensity analyser, VADER (Valence Aware Dictionary and sEntiment Reasoner), they attempted to identify change in sentiment during each president’s term in office and make comparisons between their first and last speech in office. A sentiment score was assigned to each sentence and aggregated for each speech. They also utilised Latent Dirichlet Allocation (LDA) to identify topics and multidimensional scaling to select the number of topics for each speech. They subsequently aggregated speeches made to identify topics that emerged from each president.

Social media is a useful source of data for observing public opinion. O’Conner et al. (2010) found a correlation between traditional political opinion surveys and sentiment word frequencies in Twitter messages.

## Data Sources

The NDR speeches were retrieved from the National Archives of Singapore (years 1982 to 2003) and Prime Minister’s Offices’ (Years 2002 to 2022) websites. Apart from 2020 when NDR was cancelled due to Covid-19, the NDR was held annually. From 1970, the NDR address was also delivered in the Chinese and Malay language, and the English-translated versions of the speech were also retrieved. A total of 61 speeches were retrieved, each stored in a text file format.

|  |  |  |
| --- | --- | --- |
| **Language of Speech** | **Years** | **No. of Files** |
| English | 1982, 1992, 2002, 2004 to 2019, 2021 to 2022 | 21 |
| Chinese (English-translated) | 1992, 2002, 2004 to 2019, 2021 to 2022 | 20 |
| Malay (English-translated) | 1992, 2002, 2004 to 2019, 2021 to 2022 | 20 |

Table : NDR Speech Years by Language

## Methodology

We will be working on three tasks: Document Retrieval, Topic Modelling, and Sentiment Analysis. Please refer to the table below for the details.

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Task** | **Years** | **No. of Files** |
| 1 | Document Retrieval | 1982, 1992, 2002, 2004 to 2022  (All documents used in project) | 30 |
| 2 | Topic Modelling | 1982, 1992, 2002, 2012, 2022  (Every 10 years) | 13 |
| 3a | Sentiment Analysis   * Speaker’s sentiment | 1982, 1992, 2002, 2012, 2022  (Every 10 years) | 13 |
| 3b | Sentiment Analysis   * Public’s online sentiment * Public’s emotions | 2022 | -  (Using Twitter data) |

Table : Analyzed NDR Speech Year by Task

Our text files will be organised into two folders: Folder 1 for Task 1, and Folder 2 for Tasks 2 and 3.

### 1. Document Retrieval

#### 1.1 Pre-processing

The first folder of text files will be read as a corpus. Each document will go through the following steps: tokenisation, conversion to lowercase, removal of punctuation, stopwords and non-alphabets, and lemmatisation.

#### 1.2 Constructing Retrieval System

Using the resulting bags of words, a dictionary is created to represent the unique words in the corpus in alphabetical order. A document-term matrix (DTM) is then created. From the DTM, we can see the term frequencies (TF) in each document. We can also create a document-frequency vector, followed by an inverse document frequency (IDF) vector based on the following formula. The product of TF and IDF would produce a term frequency-inverse document frequency (TF-IDF) matrix.

When a term is entered as a search term, there will be a TF-IDF vector for the search term as well. Cosine Similarity between the search term and each document will be computed and used to rank the resulting documents from most to least relevant (similar).

### 2. Topic Analysis

#### 2.1 Pre-processing

Follow same steps as in Section 1.1 but using the text files (every 10 years speech data) in Folder 2.

#### 2.2 Topic Learning

Similar to steps in Section 1.2, TF-IDF matrix will be generated. This matrix will then be fed in to the LDA model. To pick the ideal number of topics, the coherence and perplexity scores can be computed. The higher the coherence score, the better the topic model. On the other hand, the lower the perplexity score, the better the topic model. Based on the top words, the topic can ideally be inferred.

Next, we will apply the LDA model to the documents to identify the topic(s) based on the speech content.

### 3. Sentiment Analysis

#### 3.1 Pre-processing

The second folder of text files will be read as a corpus. Each document will go through the following steps: Split into paragraph within document, tokenisation, conversion to lowercase, removal of punctuation, stopwords and non-alphabets and lemmatisation.

#### 3.2 Sentiment Analysis on Speech

Sentiment analysis will be performed on the speeches to understand the opinion or attitude of the overall speeches as well as on a specific topic. The resulting bag of words from the pre-processing step in section 3.1. can be used as an input to the sentiment analysis libraries / tools such as VADER, textblob, flair, HF, etc.

VADER relies on a dictionary which maps lexical features to emotion intensities called sentiment scores, that is measured on a scale from -4 (most negative) to +4 (most positive). On a sentence level, these scores will be summed up and normalized to give a score that ranges between -1 to +1.

Similarly, textblob, flair, HF and other sentiment analysis libraries will also return a score that represents positive or negative sentiments. By using multiple libraries to evaluate sentiments, consensus could be achieved by taking majority vote as the final output.

#### 3.2 Sentiment and Emotion Analysis on Online Response to Speech

Sentiment analysis is also performed on the public response to the speech. Twitter data will be used as the source of public response in which can be pulled by using Twitter API such as Tweepy. Every tweet data collected will be evaluated on two different categories: sentiments and emotions.

Similar to the speech sentiments analysis, every tweet will also go through similar text pre-processing which resulted in a bag of words. This bag of words will then be fed in into the same libraries / tools for the evaluation of sentiments.

In addition, we will also perform emotions analysis by using text2emotion python package. The package helps to process any textual message and recognizes the emotion embedded in it. The emotions that will be recognized are categorized as happy, angry, sad, surprise and fear.

## Assumptions & Limitations

The transcripts for speeches before 1972 are not available, thereby limiting the temporal comparison of speeches. Many speech transcripts before 1990 were also incomplete or unavailable in text format. Secondly, some of the English-translated versions of mandarin speeches are summarised versions of the full speech, which may pose challenge(s) for the topic modelling task.

# Appendix

## Project Timeline

|  |  |
| --- | --- |
| **Week of** | **Tasks** |
| 12 Sep (T1W4) | Project Topic Brainstorming |
| 19 Sep (T1W5) | Finalise project topic after inputs from Prof. |
| 26 Sep (T1W6) | Project Proposal  Dataset extraction |
| 3 Oct (Recess) | Data pre-processing  Division of Work |
| 10 Oct (T1W7) | Data Processing |
| 17 Oct (T1W8) | Data Analysis |
| 24 Oct (T1W9) | Data Analysis  Project Presentation Preparation & Finalisation  Project Report Preparation & Finalisation |
| 31 Oct (T1W10) | Project Presentation & Submission of Report |

Table : Estimated Project Timeline

## Document Retrieval Flow Diagram

A picture containing text, sign, screenshot, vector graphics

Description automatically generated

Figure : Document Retrieval Flow

## Topic Analysis Flow Diagram

Icon

Description automatically generated

Figure : Topic Analysis Flow

## Sentiment / Emotional Analysis Flow Diagram

### Speech (Sentiment)

Icon

Description automatically generated

Figure : Speech Sentiment Analysis Flow

### Public (Sentiment)

Graphical user interface, text

Description automatically generated

Figure : Public Sentiment Analysis Flow

### Public (Emotional)

Icon

Description automatically generated

Figure : Public Emotional Analysis Flow

# References

Min, C. H. (2016, Aug 20). *From babies to casinos: 11 memorable National Day Rally speeches*. Retrieved from The Straits Times: https://www.straitstimes.com/singapore/from-babies-to-casinos-10-memorable-national-day-rally-speeches

Miranda, J., & Bringula, R. (2021). Exploring Philippine Presidents’ speeches: A sentiment analysis and topic modeling approach. *Cogent Social Sciences*, 7:1. doi:10.1080/23311886.2021.1932030

National Archives of Singapore. (n.d.). *National Day Rally Speeches*. Retrieved from National Archives of Singapore: https://www.nas.gov.sg/archivesonline/NationalDayRally/

O'Connor, B., Balasubramanyan, R., Routledge, B., & Smith, N. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the International AAAI Conference on Weblogs and Social Media*, (pp. 122-129). Washington, DC.

Seng, L. T. (2014). *Parades, Flags and Rallies.* Singapore: National Library Board.