USA INAUGURAL SPEECH: A STATISTICAL APPROACH

Dimitri Calipolitis[1], Carla Pleștiu[2]

Abstract: The aim of this study is to determine the statistical characteristics of the terms present in the inaugural speeches of the elected American presidents, from the year 2000 to date. The methods applied belong to text mining and they determined the existence of a specific distribution of terms (Zipf and Heaps law), the correlation between certain words, the frequency of terms (alongside TF–IDF), the lexical variation within the speeches (Type–Token Ration, Hapax Richness), as well as the type of sentiment extracted by comparing specific sentiment lexicons (Afinn versus Bing). Thus, from a quantitative point of view and based on the results obtained in R, it is possible to deduce which terms will be used in the following inaugural speeches of future American presidents. This paper was co-financed by the Bucharest University of Economic Studies as part of the PhD programme.

UDC Classification: 308, 311.1, 324, **DOI**: https://doi.org/10.12955/pss.v4.393

Keywords: text mining, presidential inaugural speech, term frequency, sentiment analysis, pairwise correlation, quantitative methods.

Introduction

Text mining, also known as text analytics or natural language processing (NLP), is a powerful method for extracting valuable insights from large amounts of textual data. Using a combination of computational techniques drawn from natural language processing, machine learning, and statistical analysis, text mining enables researchers to systematically explore, analyse, and interpret textual information in a structured way (Manning, Raghavan, & Schütze, 2008). At its core, text mining involves the automated extraction of patterns, trends, and relationships from unstructured text, transforming it into actionable knowledge (Feldman & Sanger, 2007). Using sophisticated algorithms and computational models, text mining algorithms can process and analyse large volumes of text, ranging from documents, articles, and social media posts to emails, transcripts, and speeches (Jurafsky & Martin, 2008).

Given that text mining methods can be applied to a wide variety of fields in the humanities, since these fields mainly rely on written sources as working tools, we considered studying, from a quantitative point of view, the links between the existing words in the inaugural speeches of US presidents and whether there are differences between them at this level. Naturally, given the large number of American presidents, we thought that the analysis would be relevant to presidential speeches from 2000 to the present, since the geopolitical problems of the beginning of this millennium are still present today (armed conflicts, epidemics, economic crisis, social problems, etc.). This type of analysis can reveal whether the political trajectory of the state that a particular president has followed during his term of office is already reflected in his inaugural speech.

Literature Review

There are two main trends highlighted in the previous research on public speeches of political leaders. English linguists focused mainly on urgent and relevant socio-political issues, analysed through a pragmatic approach, while Russian linguists conducted a more detailed analysis of language with a more theoretical approach (Bulgakova, Shabanova & Eliseeva, 2020). The study highlights the importance of cross–cultural collaboration and comparative research in this area, suggesting that language policy could benefit from a more generalised approach and better collaboration between linguists from different cultural contexts.

Other studies examine the sentiments expressed in US and Ukrainian presidential speeches. The study shows that the speeches of Presidents Trump and Poroshenko are subjective and positive in polarity, with the predominance of positive attitudes expressing affect (Dilai, Onukevych & Dilay, 2018). The main contribution of the study is the adaptation of the SentiStrength programme for the automatic analysis of Ukrainian political texts, with the integration of a lexicon of political terms in the

[1] Bucharest University of Economic Studies, Faculty of Economic Cybernetics, Statistics and Informatics, Department of Statistics and Econometrics, calipolitisdimitri20@stud.ase.ro
[2] Bucharest University of Economic Studies, Faculty of Economic Cybernetics, Statistics and Informatics, Department of Statistics and Econometrics, plestiucarla18@stud.ase.ro

Ukrainian language. The results can serve as a basis for the further development of the methodology of sentiment analysis in Ukrainian texts.

After examining the types of speech acts used by American presidents in their speeches and the influence of socio–political events on them, it was found that presidential speeches can have different functions, such as emotional, informative, persuasive or motivational, depending on the event and the political context (Saeed, 2014). The author analyses the speeches of four American presidents and finds that different types of speech acts fulfill several functions simultaneously, contributing to the achievement of the objectives of the presidential speech.

Data and methodology

The presidential inaugural speeches were exported in full from https://millercenter.org/the–presidency/presidential–speeches and saved in separate text files, which were later imported into RStudio for processing. Two types of data processing were used: one that retained the stop words in order to determine the type of distribution of the data, and a second that eliminated this category of terms and on which the rest of the analysis was carried out. In the case of both processes, once the files were imported into the application, they were embedded in a volatile corpus (VCorpus), which was later converted into a term document matrix (TDM), and on which several transformations were applied (removal of numbers, punctuation and document titles, conversion of uppercase to lowercase, removal of spaces and lemmatisation). After making the transformations, the new object was converted into a data frame on which several functions were applied, so that the final table (Table 1) contains the following columns: "word" – the terms from the speeches, "speech" – the reference to speech, "count" – the number of occurrences of the term in a speech and "total" – the total number of words in a speech. Where necessary, other tables were created from the main one, adding different columns for the calculation of the indicators, needed for the analysis.

Table 1: Headline of the main dataframe

word	speech	count	total
freedom	Bush Inaugural Speech 2005	27	810
america	Bush Inaugural Speech 2005	20	810
america	Biden Inaugural Speech 2021	20	796
america	Trump Inaugural Speech 2017	18	527

Source: Authors

The methods were adapted according to the theory found in the scientific literature. In order to check the distribution of the data, it was tested whether Zipf's law, according to Zipf (1949), or Heaps' law, according to Heaps (1978), applied. For the calculation of specific indicators, such as term frequency times inverse document frequency (TF–IDF), type–token ratio (TTR), or hapax richness (HR), the formulas and interpretations of Rajaraman and Ullman (2011), Thomas (2005a), Jockers and Thalken (2020) were used. For sentiment analysis, the lexicons of Hu and Liu (2004) and Nielsen (2011) were integrated into the analysis. For the calculation of correlations between word pairs, concepts from Cramer (1946) and Matthews (1975) were adopted. At the code level, mechanisms were adapted from the documentation of dedicated R packages, as well as from other scientific works, such as Silge and Robinson (2017), Zhang (2018), Jockers and Thalken (2020). The choice of topics was inspired by the case studies and methods found in Niekler and Wiedemann (2020).

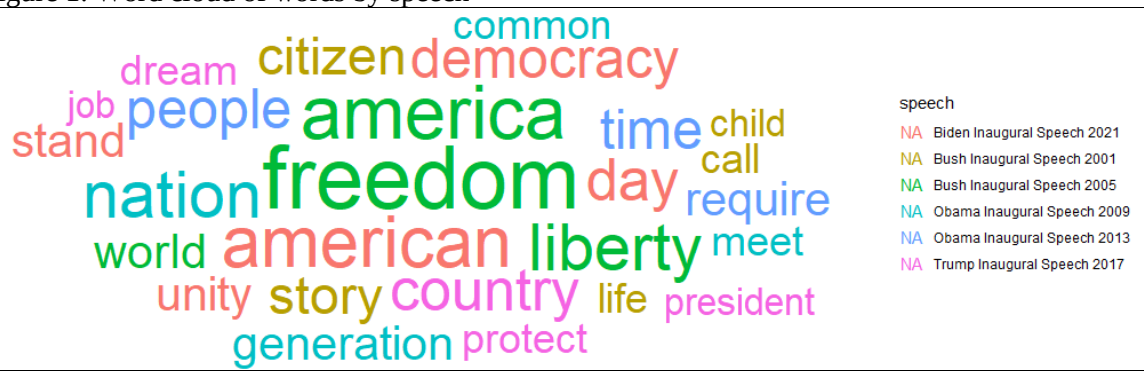
Results and Discussion

After applying the methods described in the previous section, we observed a number of quantitative differences between the presidential inaugural speeches, both at the level of the terms used and in the presentation of some defining characteristics of text mining, such as the type of feeling conveyed or the correlation between pairs of words. It should be noted that the distribution of the data in the dataframe follows up Zipf's law, since the information, that the frequency of a word is inversely proportional to its statistical rank (Zipf, 1949) is validated on the basis of the specific linear model obtained from the econometric modelling of the data (where $\log_{10}(\text{word_freq}) = (-1.268211) - 0.737753 * \log_{10}(\text{rank})$ and $R^2 = 0.9276$). According to the theory, if the existence of the Zipf distribution is demonstrated in a corpus of documents, then the Heaps distribution is also present (Kumar and Paul, 2016), but for confirmation, we created an econometric model, based on the data

from the speech document–term matrix, which validates this statement (where distinct_words = 1.7996249 + 0.6248062 * total_words).

When analysing the data frame with the stop words removed, the word frequency values, for the first ten terms in each inaugural speech, are noted to be between 0.006 (e.g.: "common", "meet" – Obama Inaugural Speech 2009) and 0.033 (e.g.: "freedom" – Bush Inaugural Speech 2005). This low frequency of the most frequently used terms is explained by the simple fact their number of occurrences is low, ranging from five to twenty–seven, out of a total number of words present in the speeches ranging from 527 (Trump Inaugural Speech 2017) to 877 (Obama Inaugural Speech 2009). From the word cloud of frequently used terms (Figure 1), we can see an affinity of the speeches towards the protection of the American nation and the values that define it, as words such as "america", "american", "democracy", "liberty", "freedom" or "nation" appear.

Figure 1: Word cloud of words by speech



Source: Authors

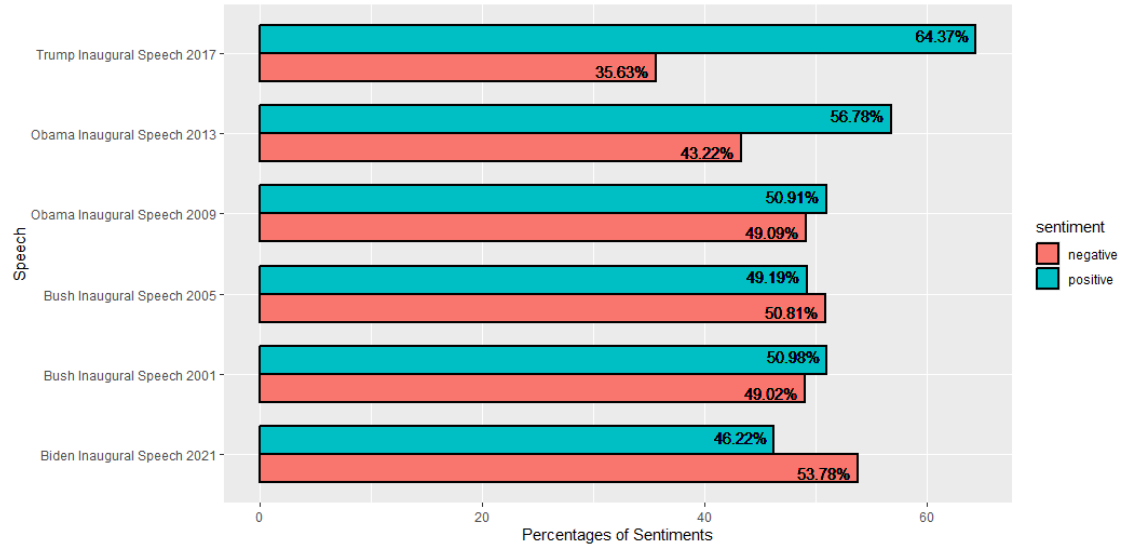
The lexical variation within the inaugural speeches is high, as the value of the Type–Token Ratio indicator is between 0.55 and 0.68, while for Hapax Richness, it is included between 0.37 and 0.50. On the basis of these two indicators and from a lexical point of view, it can be seen that the most varied speech is the one of Barack Obama from 2009 (TTR = 0.6818700, HR = 0.5062714), at the opposite pole being that of Joe Biden (TTR = 0.5590452, HR = 0.3793970). In addition, the hypothesis that the number of *hapax legomena* increases with the length of the text (Jockers and Thalken, 2020) is confirmed in the speeches, since that the correlation coefficient between the total number of words and the rare terms has a value of 0.8279153, underlining a direct and high–intensity relationship between the two variables mentioned.

At the level of term frequency times inverse document frequency (TF–IDF), which was calculated for the existing words in the speeches, it is noted that, for the first five important terms according to this indicator and related to each speech, their value is between 0.0040 (e.g.: "muslim", "conflict", "cooperation" – Obama Inaugural Speech 2009) and 0.0186 (e.g.: "story" – Bush Inaugural Speech 2001). However, it should be noted that the words with a high TF–IDF value actually represent the main ideas of the speeches, such as: compassion towards the nation (e.g.: "compassion", "community", " civility" – Bush Inaugural Speech 2001), the unjustified tyranny applied to human beings (e.g.: "tyranny", "human", "excuse" – Bush Inaugural Speech 2005), the common fight against terrorism (e.g.: "muslim", "conflict", "mutual" – Obama Inaugural Speech 2009), the completion of a successful journey towards the happiness of Americans (e.g.: "journey", "happiness", "succeed" – Obama Inaugural Speech 2013), the attack on politics, immigrants and on the predecessor of the office (e.g.: "obama", "foreign", "politician" – Trump Inaugural Speech 2017) or the triumph of democracy despite the pandemic (e.g.: "prevail", "democracy", "virus" – Biden Inaugural Speech 2021).

Sentiment analysis shows that the number of positive terms is greater than the number of negative terms, both at the level of the utterance and in relation to the total number of words collected from the processed texts. Obviously, these terms vary from speech to speech, and the results of this type of analysis vary depending on the lexicon used. In the analysis carried out using the Bing lexicon (Figure 2), it can be seen that Donald Trump's speech is the one with a high percentage of positive terms (positive = 64.37%, negative = 35.63%), followed by Barack Obama's 2013 speech (positive = 56.78%, negative = 43.22%), while Joe Biden's is the only one where the percentage of negative terms

is much higher than that of positive terms (positive = 46.22%, negative = 53.78%). It should also be noted that for the other texts, the sentiment percentages for the terms show insignificant differences.

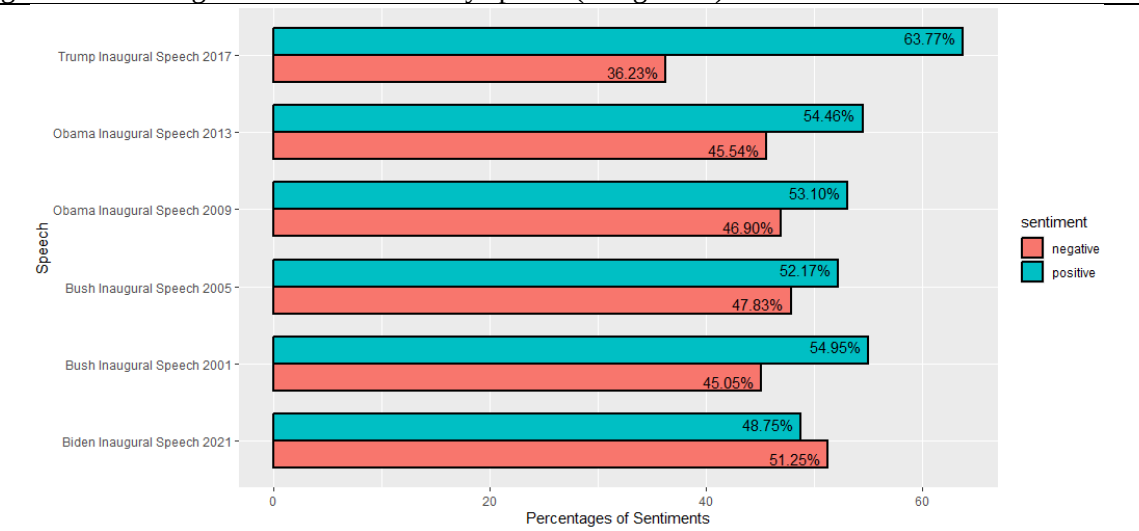
Figure 2: Percentages of total sentiment by speech (using bing)



Source: Authors

In the case of the analysis based on the Afinn lexicon, several features stand out. First, when the sentiment score is summed up for each text, it turns out that Donald Trump's speech has the highest score (sentiment_score = 50), implicitly conveying a positive sentiment, while Joe Biden's speech has the lowest score (sentiment_score = -3), conveying a negative sentiment. Secondly, although the sentiment score varies from one speech to another, it does not show any statistical relationship with the other indicators of the data frame (rank, term frequency, TF–IDF), since their correlation coefficient is included between -0.11 and 0.16. Thirdly, after dividing the words into two categories (positive and negative) and calculating the corresponding percentages at the text level (Figure 3), it can be observed that Donald Trump's speech remains the one with the highest percentage of positive terms (positive = 63.77%, negative = 36.23%), while Joe Biden's remains the only speech where the rate of negative terms is higher than that of positive terms (positive = 48.75%, negative = 51.25%).

Figure 3: Percentages of total sentiment by speech (using afinn)

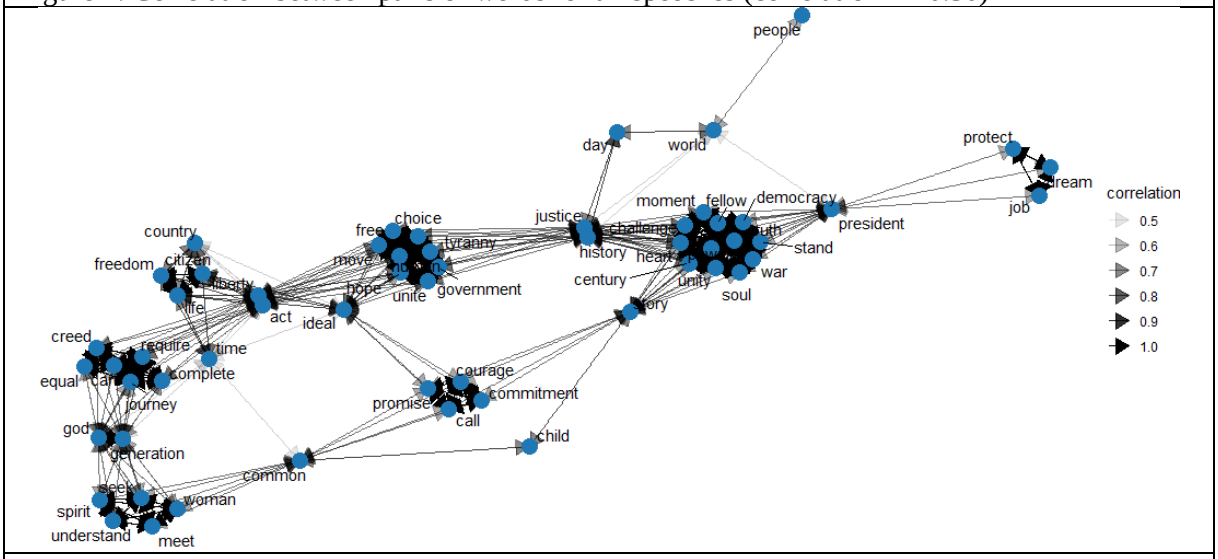


Source: Authors

The correlation between the pairs of words was only applied to the total level of terms, collected from all the texts, since the analysis applied to each individual speech did not provide any concrete information (all the cases showed only two types of links between words and the same values for the phi coefficient). After calculating the phi correlation coefficient, we created a model that includes the pairs of terms, where this indicator is equal or greater than 0.50, in order to strictly include the words

between which there is a medium or strong intensity connection. On the basis of the graph (Figure 4), which represents these links in the form of Markov chains, it is possible to identify several groups of terms with a high degree of interdependence, each consisting of between two and eleven words. It can also be seen that, for each individual term, the links between the other words are exercised in several directions, which is reflected by the direction of the arrows shown in the graph. At the same time, this statement justifies the idea that all the pairs of words between which there is a high and positive value of the correlation coefficient are related and that they constitute, from a quantitative point of view, the lexical core of the collection of analysed speeches.

Figure 4: Correlation between pairs of words for all speeches (correlation >= 0.50)



Source: Authors

Conclusion

The inaugural speeches of US presidents must necessarily contain terms referring to the American nation or to the values shared by Americans, so that the new president can give the public the impression that they are interested in the fate of the nation they represent. According to the analysis carried out, these terms are present in all speeches and constitute an important quantitative weight, regardless of the lexical variation present in the texts. However, if one wants to identify the words that have a certain importance in a speech and that indirectly reveal its main ideas, the TF–IDF indicator is a great help in this case. At the same time, in order to determine the type of feeling conveyed, it should be noted that the use of any lexicon implies the omission of many words from the initial data frame. For example, when using the Bing lexicon, between 19% and 26% of the terms found in the original dataframe were retained, while for the Afinn lexicon, these percentages are between 17% and 22%. With regard to the correlation between pairs of words, the division of each speech into several sections and the calculation of the phi coefficient do not give conclusive results, due to the low total number of terms present in each text. However, at the level of the collection of speeches, this indicator can be calculated, but the representation of the model for the data should be done in a filtered way, because it requires a large amount of computing power to include all the input data. Also, if more speeches were included in the analysis, complex machine learning models could be built that would predict the delivery of this type of speech for a future president.

References

Bulgakova, Y., Shabanova, V., & Eliseeva, A. (2018). The Language of Leadership: Analyzing Public Speeches of Presidents (Literature Review). Springer Proceedings in Business and Economics, 69–77.

Cramer, H. (1946). Mathematical Methods of Statistics. Princeton: Princeton University Press.

Dilai, M. , Onukevych, Y., & Dilay I. Sentiment Analysis of the US and Ukrainian Presidential Speeches. Proceedings of the 2nd International Conference Computational Linguistics And Intelligent Systems, Vol. 2, 60-70.

Feldman, R., & Sanger, J. (2007). The text mining handbook: Advanced approaches in analyzing unstructured data. Cambridge University Press.

Heaps, Harold Stanley (1978), Information Retrieval: Computational and Theoretical Aspects, Academic Press.

Hu, Minqing, and Bing Liu. (2004). Mining and summarizing customer reviews. KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 168–77. https://doi.org/10.1145/1014052.1014073.

Jockers, Matthew L., and Rosamond Thalken. (2020). Text Analysis with R: For Students of Literature. 2nd ed. Cham, Switzerland: Springer Nature Switzerland.

Jurafsky, D., & Martin, J. H. (2008). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Pearson Prentice Hall.

Kumar, Ashish, and Avinash Paul. (2016). Mastering Text Mining with R. Birmingham, United Kingdom: Packt Publishing.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. 405 (2): 442–451. doi:10.1016/0005-2795(75)90109-9.

Niekler, A. and G. Wiedemann. (2020). Text mining in R for the social sciences and digital humanities. Retrievable from https://tm4ss.github.io/docs/index.html

Nielsen, Finn Årup. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In CEUR Workshop Proceedings, 718:93–98. http://arxiv.org/abs/1103.2903.

Rajaraman, A., & Ullman, J. D. (2011). Data Mining. Mining of Massive Datasets, 1–17. doi:10.1017/cbo9781139058452.002

Saeed, R. (2014). A Speech Act Analysis of American Presidential Speeches. Arts Journal, 110, 1-40.

Silge, J and D. Robinson. (2017). Text Mining with R: A Tidy Approach. Sebastopol, California: O'Reilly Media.

Thomas, D. (2005a). Type-token Ratios in One Teacher's Classroom Talk: An Investigation of Lexical Complexity.

Zhang, Z. (2018). Text Mining for Social and Behavioral Research Using R: A Case Study on Teaching Evaluation. Retrievable from https://books.psychstat.org/textmining.

Zipf, George K. (1949). Human Behavior and the Principle of Least Effort. Cambridge, Massachusetts: Addison-Wesley.