On the Applicability of Features of Linguistic Complexity to the Classification of Political Speeches

Zarah Weiß
University of Tübingen
zweiss@sfs.uni-tuebingen.de

Abstract

This paper compares the performance of classical word vectors and complexity vectors on the tasks of party and government affiliation identification for political speeches from German *Bundestag* using SVMs with linear kernels. The results show that BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA

1 Introduction

The analysis of political speeches as well as the prediction of party affiliation are common targets of computational linguistic analyses. The goals of these analyses are mainly related to voting behaviour prediction or the detection of ideological stances.

- Increasing popularity of using linguistic complexity for classification
- Proficiency analysis, readability
- Also sometimes applied to stylometrics
- Here: identifying political affiliation in terms of party and government participation
- use of content neutral features
- organisation affiliation
- political speeches in Bundestag, highly formal language, which is a hybrid between spoken and written language.
- use political speeches for party prediction, since affiliation is very clear
- allows for focus on feature investigation

Can party affiliation be identified by linguistic complexity measures? That is, are there systematic differences or similarities between how members of parties phrase their discussions? And has being part of the ruling coalition or the opposition an effect on speech style?

This paper tries a novel approach, comparing the performance of classical word vectors with complexity vectors. The results show that ... BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA

The remainder of the paper is structured as follows: first, a brief overview over related work and similar approaches is given. Then, the data set used is introduced. Section 4 discusses the methods used in the classification experiments and section 5 reports the experiments. The article closes with the conclusion.

2 Related Work

One common strand in the analysis of political speeches is the analysis of political affiliations of parties or individuals, allowing to estimate similarities between them with respect to certain positions. This is highly related to political science and journalism. These analyses are often relying on word patterns (**lowe2013**). Lowe et al. (2011) identify political affiliation on a continuous left to right scale using political categories assigned to sentences based on their content. Sim et al. (2013) are identifying the ideological position of American politicians expressed in their political texts using an ideology lexicon and a Bayesian Hidden Markov Model.

Other research focuses more on the aspect of organisation affiliation: Dahllöf (2012), Koppel 2009 Dahllöf (ibid.) performs a binary classification of gender, age, and political affiliation based on Swedish parliament speeches from 2003 to 2010 using binary word vectors indicating absence or presence of words. Koppel et al. (2009) identify both, ideological and organisational affiliation for Arabic texts.

Party prediction is often performed based on social media data, in order to facilitate the analysis of voting behaviour, see Gottipati et al. (2013).

Abu-Jbara et al. (2012) use unsupervised learning methods to identify sub-group affiliation in political fora.

3 Data

For this study, German political speeches from the 13th to 17th legislative period of German *Bundestag* were analysed. The speeches held between February 8th, 1996 and September 3rd, 2013 were extracted from 985 protocols obtained from the *PolMine-Plenardebattenkorpus* (PDK) by **???**, which contains protocols from German parliaments on state and federal level. These protocols also contain information on audience comments during speeches, which were excluded for this study. Also, administrative remarks, such as the appointment of the next speaker or voting summarize, are annotated as separate speeches. These passages of moderation were excluded from the data set, too, be removing all speeches from the respective *Bundestags(-Vize)Präsident(-in)*. This reduced the amount of speeches in the data set from 215,061 to 120,331.

The PDK contains meta information on each speaker's party, function, and name, as well as on each protocol's legislative period, protocol number, and protocol date. Participation in the governing coalition was inferred by combining a speaker's party affiliation with the speech's legislative period and added as additional meta information.[1] For this, CDU and CSU were considered to be a single party as well as PDS and Linke, although PDS and Linke only merged in June 2007.

4 Methods

- Government vs. Opposition
- Party affiliation
- (Multinomial) Logistic Regression
- Stratified 10-folds cross validation
- precision recall f1 score
- svm common choice for classification: dahllöf 2012

4.1 Features

For each classification experiment, two different types high-dimensional feature vectors were employed: first, a rather traditional word vector and second, a complexity vector. In order to keep the analysis feasible, each document was presented by averaged complexity and word vectors, instead of allowing for multidimensional matrices.

GloVe Koppel 2009: Each document is represented as a numerical vector each entry of which is the frequency of some linguistic feature in the document. In our experiment, we simply chose as features the 1000 most common words in the entire corpus. We did not use stemming since this process is time-consuming and preliminary tests indicated that it is not necessary for achieving good results. Note that the total number of unique words appearing in the corpus is in the tens of thousands; we chose only the most frequently occurring 1000 words since it is well established [4] that filtering in this way increases efficiency without degrading accuracy. The 1000 words we chose include both function words and content words. (see Forman 2003)

- WordVectors `http://nlp.stanford.edu/projects/glove/`
- Size
- Where from

Complexity Features

- style-based analysis
- 218 features of linguistic complexity
- Descriptive features, Lexical, Syntactic, Morphological, coherence and cohesion, complex linguistic constructions

[1] The governing coalition of each legislative period was obtained from `https://de.wikipedia.org/wiki/Liste_der_deutschen_Bundesregierungen`.

4.2 Learning Algorithm

5 Experiment

Dahllöf als Diskussions und Design Beispiel!

5.1 Results

5.2 Discussion

6 Conclusion

References

Abu-Jbara, Amjad et al. (2012). “Subgroup detection in ideological discussions”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pp. 399–409.

Dahllöf, Mats (2012). “Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches—A comparative study of classifiability”. In: *Literary and linguistic computing*, fqs010.

Gottipati, Swapna et al. (2013). “Predicting user’s political party using ideological stances”. In: *International Conference on Social Informatics*. Springer, pp. 177–191.

Koppel, Moshe et al. (2009). “Automatically Classifying Documents by Ideological and Organizational Affiliation.” In: *ISI*, pp. 176–178.

Lowe, Will et al. (2011). “Scaling policy preferences from coded political texts”. In: *Legislative studies quarterly* 36.1, pp. 123–155.

Sim, Yanchuan et al. (2013). “Measuring ideological proportions in political speeches”. In: *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 91–101.