

Statistical Consulting

Topic-Specific Sentiment Analysis for Tweets by German MPs

Department of Statistics
Ludwig-Maximilians-Universität München

Munich, month dayth, 2021



Authors Asmik Nalmpatian
Lisa Wimmer

Project Partner Prof. Dr. Paul Thurner
Department of Political Science

Supervisors Matthias Aßenmacher, Ph.D.
Prof. Dr. Christian Heumann
Department of Statistics

Abstract

Contents

1	Introduction	1
2	Project Outline	2
3	General Theoretical Context	3
3.1	Terminology	3
3.2	Topic Modeling	3
3.3	Sentiment Analysis	3
3.4	Topic-Specific Sentiment Analysis (TSSA)	3
4	Analytical Proposal	3
4.1	Data	3
4.1.1	Data Collection	3
4.1.2	Data Pre-Processing	3
4.2	Standard Machine Learning Solution	3
4.2.1	Methodology	3
4.2.2	Results	3
4.3	Deep Learning Solution	3
4.3.1	Methodology	3
4.3.2	Results	4
5	Knowledge Transfer	4
5.1	Static Material	4
5.2	Live Teaching	4
6	Discussion	4
6.1	Analytical Proposal	4
6.2	Knowledge Transfer	4
7	Conclusion	4
A	Appendix	V
B	Electronic Appendix	VI

List of Figures

List of Tables

List of Abbreviations

MP	Member of Parliament
NLP	natural language processing
STM	structural topic model
TSSA	topic-specific sentiment analysis

List of Symbols

A	identity matrix with s^2 entries
0	s -dimensional zero vector

1 Introduction

The advance of social media has sparked a fundamental change to public debate. Political coverage and discourse has spilled from traditional outlets over to online locations whose accessibility has lowered entrance barriers to welcome a wide audience (Bode, 2017). Social media exhibit certain properties that make them attractive to politicians seeking to broadcast their message. From a supply-side perspective it is easy to publish content: posting on social media is cheap, does not require approval of any authority, and allows for full control over the intended presentation. In an environment that spins information at enormous speed the resulting ability to react to events in real-time offers a distinct advantage over the inertia of traditional channels (Stier et al., 2018). On the receiving end, social media grant unprecedented access to target audiences. The industry’s oligopolistic structure sees people from heterogeneous backgrounds convene on few global platforms to exchange their views. Research suggests that the low cost of online engagement causes political content to also circulate among users with less political affinity (Jost et al., 2018). A contrary but equally important aspect is the evolution of echo chambers. Users view content according to their perceived preferences and thus often end up in community niches populated by like-minded people. This clustering process creates groups disproportionately receptive to certain messages and has played an important role in large-scale propaganda. Echo chambers are particularly suited for the direct communication among users social media offer: they enable a dialogue between politicians and their electorate that is hard to achieve via traditional channels, and allow to deploy the power of emotion to shape opinion (Hasell and Weeks, 2016).

These opportunities have propelled internet platforms to a position at least level with the former hegemon of political debate. Twitter, in particular, has emerged as a medium for political information. In what might have been inconceivable a few years ago politicians actively convey messages to the public via tweets (van Vliet et al., 2020). The impact of this change in the political environment is complex and most certainly has positive as well as worrisome aspects. Yet, from a purely scientific point of view, activity on social media creates on its way a vast amount of publicly accessible data that benefits the research community with a constant source of information.

A question frequently posed in political analysis is the assessment of public opinion toward a particular matter. However, the textual data gathered from social media that might hold the answer command the use of specific tools subsumed under the field of *natural language processing* (NLP). NLP has gained much traction with the rise of deep learning methods and become virtually ubiquitous, techniques ranging from simple heuristics to gigantic neural networks powering search engines and the like (Torfi et al., 2020). Statistically speaking, the above problem translates into the classification of texts into instances of certain sentiments (typically, *positive* and *negative*). Building upon the assumption that sentiment may be expressed differently in varying contexts, such *sentiment analysis* is often combined with some form of *topic modeling* (see, for example, Ficamos and Liu (2016)).

It is the goal of this project to make analysis of social media texts in a political context more easily accessible to researchers. We focus on the analysis of public sentiment in a topic-aware manner for texts collected from Twitter posts by German Members of Parliament (MPs). Our contribution is two-fold:

1. We explore how topic-specific sentiment analysis can be implemented, considering (1) standard machine learning techniques and (2) more complex deep learning models.
2. We provide extensive teaching material on both approaches, composed as a coherent online course, to educate researchers on addressing NLP problems in their own work.

The remainder of this report is organized as follows. First, we outline the project in more detail in section 2. Section 3 provides some theoretical context for topic modeling and sentiment analysis from which we derive what we call *topic-specific sentiment analysis (TSSA)*. We proceed in section 4 by sketching our data collection and cleaning process and then present our proposal for conducting TSSA, laying out for both approaches the underlying methodology and discussing the results of applying them to the data at hand. Section 5 outlines how the findings from this analysis translate to the proposed teaching material. Afterwards, in section 6, we critically assess the findings and limitations of the project, and conclude with a brief summary in section 7.

2 Project Outline

Before we outline the scope of our work it should be noted that parts of it are based on a predecessor project. Schulze and Wiegrebe (2020) studied how German MPs’ Twitter data can be modeled with a *structural topic model (STM)* (Roberts et al. (2013)), and have since engaged in follow-up research on the STM (Schulze et al., 2021). Much of the data procurement and topic modeling process is adopted from their work. The project at hand encompasses two subsequent steps. In a pioneering mode we first explore the overall feasibility of TSSA with both basic and more advanced statistical techniques from the NLP toolbox, and then, based on our findings, propose a collection of material to support fellow researchers in conducting similar studies.

Topic-specific sentiment analysis

Regardless of how the downstream task is solved, the first challenge to address is data collection. The idea here is to retrieve information from the web in an automated and resource-efficient manner that results in a suitable data structure. Afterwards, we pursue two fundamentally different approaches toward performing TSSA.

1. The first approach applies standard machine learning tools which require the input data to be of tabular form. Obviously, texts are complex constructs and not arbitrary sequences of interchangeable characters that can simply be cast into tabled variables (section 4.1.2 will address the challenges arising from Twitter data in more detail). It is nevertheless possible, as general research and also our own results suggest, to obtain fairly good performance with this reduction of complexity.
2. State-of-the-art approaches, by contrast, avoid such blunt simplification and attempt to teach the entire concept of language to machines. This comes at the expense of large computational requirements but achieves promising results in a variety of NLP tasks. We therefore build a deep bidirectional Transformer architecture (BERT, Devlin et al. (2019)) as a second approach and examine whether the additional complexity is justified by better performance.

Knowledge transfer

Based on the results of this exploratory analysis we propose teaching material devised to support research in similar applications. The acquired collection is organized as a coherent and self-contained tutorial composed of basic theory, code demonstrations, and exercises (including solutions). While the course materials are primarily aligned to solve the TSSA task, the covered components are certainly also instructive for other types of applications. We have made the material available for both live teaching purposes and self-study on a public website. First experiences from a live workshop held in April/May 2021 for researchers from the Department of Political Science at LMU will be discussed in section 6.

3 General Theoretical Context

3.1 Terminology

In this section we briefly review the general theoretical concepts behind our analysis; the actual methods we employ are described in chapter 4. Throughout the report we will make use of the following terminology:

Document Foo

Corpus. Foo

3.2 Topic Modeling

Recall that the ultimate goal is the classification of tweets into groups signaling a specific sentiment. It is reasonable to assume that sentiment, and the way of expressing it, is susceptible to context (SOURCE), which suggests potential gains from clustering tweets prior to sentiment analysis. Grouping texts into semantic clusters is generally referred to as *topic modeling* and typically constitutes an unsupervised learning task. The idea is to uncover latent structures in a corpus

3.3 Sentiment Analysis

3.4 Topic-Specific Sentiment Analysis (TSSA)

4 Analytical Proposal

4.1 Data

4.1.1 Data Collection

4.1.2 Data Pre-Processing

4.2 Standard Machine Learning Solution

4.2.1 Methodology

Automated machine learning pipeline

foo

Methodological concepts

foo

4.2.2 Results

4.3 Deep Learning Solution

4.3.1 Methodology

Deep transfer learning

foo

BERT

foo

4.3.2 Results

5 Knowledge Transfer

5.1 Static Material

5.2 Live Teaching

6 Discussion

6.1 Analytical Proposal

6.2 Knowledge Transfer

7 Conclusion

foo

A Appendix

B Electronic Appendix

Data, code and figures are provided in electronic form.

References

- Bode, L. (2017). Gateway political behaviors: The frequency and consequences of low-cost political engagement on social media, *Social Media + Society* **3**.
- Choi, Y., Kim, Y. and Myaeng, S.-H. (2009). Domain-specific sentiment analysis using contextual feature generation, *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*.
- Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.
- Ficamos, P. and Liu, Y. (2016). A topic based approach for sentiment analysis on twitter data, *International Journal of Advanced Computer Science and Applications* **7**.
- Hasell, A. and Weeks, B. E. (2016). Partisan provocation: The role of partisan news use and emotional responses in political information sharing in social media, *Human Communication Research* **42**: 641–661.
- Jebbara, S. and Cimiano, P. (2016). Aspect-based sentiment analysis using a two-step neural network architecture, *Proceedings of the Third SemWebEval Challenge*, pp. 153–167.
- Jost, J. T., Barbera, P., Bonneau, R., Langer, M., Metzger, M., Nagler, J., Sterling, J. and Tucker, J. A. (2018). How social media facilitates political protest: Information, motivation, and social networks, *Advances in Political Psychology* **39**(1).
- Naskar, D., Mokaddem, S., Rebollo, M. and Onaindia, E. (2016). Sentiment analysis in social networks through topic modeling, *Proceedings of the Tenth Language Resources and Evaluation Conference*.
- Pontiki, M., Galanis, D. and Papageorgiou, H. (2016). Semeval-2016 task 5: Aspect based sentiment analysis, *Proceedings of the Tenth International Workshop on Semantic Evaluation*.
- Qiang, J., Qian, Z., Li, Y., Yuan, Y. and Wu, X. (2019). Short text topic modeling techniques, applications, and performance: A survey, *Journal of LaTeX Class Files* **14**(8).
- Roberts, M., Stewart, B., Tingley, D. and Airolidi, E. (2013). The structural topic model and applied social science, *Advances in Neural Information Processing Systems Workshop on Topic Models*, pp. 1–20.
- Schulze, P. and Wiegerebe, S. (2020). Twitter in the parliament - a text-based analysis of german political entities, *Technical report*, Ludwig-Maximilians-Universität, Munich.
- Schulze, P., Wiegerebe, S., Thurner, P. W., Heumann, C., Aßenmacher, M. and Wankmüller, S. (2021). Exploring topic-metadata relationships with the stm: A bayesian approach.
- Sivakumar, M. and Reddy, U. S. (2017). Aspect based sentiment analysis of students opinion using machine learning techniques, *Proceedings of the International Conference on Inventive Computing and Informatics (ICICI)*.
- Stier, S., Bleier, A., Lietz, H. and Strohmaier, M. (2018). Election campaigning on social media: Politicians, audiences, and the mediation of political communication on facebook and twitter, *Political Communication* **35**(1): 50–74.

- Thelwall, M. and Buckley, K. (2013). Topic-based sentiment analysis for the social web: the role of mood and issue-related words, *Journal of the American Society for Information Science and Technology* **64**(8): 1608–1617.
- Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N. and Fox, E. A. (2020). Natural language processing advancements by deep learning: A survey, *CoRR* **abs/2003.01200**.
URL: <https://arxiv.org/abs/2003.01200>
- van Vliet, L., Törnberg, P. and Uitermark, J. (2020). The twitter parliamentarian database: Analyzing twitter politics across 26 countries, *PLoS ONE* **15**(9).
- Zhao, Y., Qin, B., Liu, T. and Tang, D. (2016). Social sentiment sensor: a visualization system for topic detection and topic sentiment analysis on microblog, *Multimedia Tools and Applications* **75**: 8843–8860.
- Zhao, Y., Qin, B., Liu, T. and Tang, D. (2017). Social sentiment sensor: a visualization system for topic detection and topic sentiment analysis on microblog, *Sustainability* **9**(10).

Declaration of Authorship

We hereby declare that the report submitted is our own unaided work. All direct or indirect sources used are acknowledged as references. We are aware that the report in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of our work with existing sources we agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future reports submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.