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

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 hegemons 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 or 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.

The basic approach is implemented in `R` (R Core Team, 2021) and thus easily integrated with statistical education at LMU. For the BERT solution we resort to `Python` (van Rossum and Drake, 2011) which is all but standard for deep (NLP) modeling.

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:

Word. Words w are sequences of characters and represent the smallest unit of text we consider.

Vocabulary. The aggregate of unique terms present in a collection of text constitutes a vocabulary of length $V \in \mathbb{N}$ from which a one-hot encoding for words can be derived: for the v-th instance of the vocabulary, $v \in \{1, 2, \ldots, V\}$, this is a length-$V$ vector with all but the v-th entry, which is one, equaling zero. Note that pre-processing (discussed in chapter 4) might result in a vocabulary that is smaller than the total number of distinct words occurring across all texts.

Document. Documents $d \in \{1, 2, \ldots, D\}$, $D \in \mathbb{N}$, are generally understood to be sequences of $N_d \in \mathbb{N}$ words, and, in our case, tweets.

Corpus. Lastly, the set of all D documents considered makes up a corpus.

3.2 Theoretical Concepts

3.2.1 Topic Modeling

Idea. 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, which suggests potential gains from clustering tweets prior to sentiment analysis (see, for example, Ficamos and Liu (2016), Bhatia and Padmanabhan (2018), or Jang et al. (2021)).
Grouping texts into semantic clusters, or topics, is generally referred to as *topic modeling* and typically an unsupervised learning task. The idea is to uncover latent structures in a corpus along which documents can be characterized. Topic modeling is essentially a means of dimensionality reduction. Text analysis requires text to be cast to numerical representation, the simplest form of which is to represent documents by counts of vocabulary instances. The dimension of the resulting document-term matrix increases exponentially in the number of documents and words contained in them, making an urgent case for dimensionality reduction (Vayansky and Kumar, 2020).
Topic modeling results in two types of output: one that links words with their propensity of occurring within a topic $k \in \{1, 2, \ldots, K\}$, $K \in \mathbb{N}$, and one stating the extent to which documents discuss each topic. This projection of texts into a K-dimensional latent space ($K \ll V$) is a purely mathematical operation and the assessment of interpretability is up to human judgment. Usually the resulting topics are then examined with respect to their most characteristic terms, according to an appropriate measure, in the attempt to find a meaningful description. In particular, the number of topics K is a hyperparameter that must be specified a priori (Aggarwal, 2018).

Approaches. Topic modeling approaches roughly decompose into deterministic and probabilistic, or generative, approaches. The former are based on factorizing the document-term matrix $M \in \mathbb{R}^{D \times V}$ (or a weighted version that takes into account prior probabilities of term occurrence) into two low-rank matrices U, W^T, $U \in \mathbb{R}^{D \times K}$ and $W^T \in \mathbb{R}^{K \times V}$, whose product approximates M loss-minimally. Probably the most prominent methods from this category are *latent semantic analysis (LSA)* (also known as *latent semantic indexing*), which performs singular value decomposition and thus projects the data into a subspace spanned by M's principal eigenvectors, and *non-*

negative matrix factorization (NMF), a constrained version that often yields better interpretability (Aggarwal, 2018).
Non-probabilistic models suffer from limitations in inference and out-of-sample extension, which is why generative approaches, addressing these issues, have become widely popular. Generative models hail from the Bayesian paradigm. Loosely speaking, they seek to reverse-engineer the imaginative process of how documents generation: first, for each document d in a corpus we draw a length-K vector of topic proportions from some distribution; then, for each word position in $\{1, 2, \ldots, N_d\}$, assign it to a topic with probabilities depending on the sampled topic proportions, and then draw a word from the distribution associated with this topic (Vayansky and Kumar, 2020). *Latent Dirichlet allocation (LDA)* by Blei et al. (2003), employing Dirichlet and multinomial distributions, pioneered this approach to topic modeling. We revisit LDA in section 4.2.1 as it also provides the foundation for the STM.

3.2.2 Sentiment Analysis

Idea. Sentiment analysis is a standard classification problem and as such an inherently supervised learning task. Under the assumption that the data can be categorized into $g \in \mathbb{N}$ discrete classes we predict for each document its associated class from a set of features. More formally, we find a model $f : \mathcal{X} \to \mathbb{R}^g$, $\mathcal{X} \subseteq \mathbb{R}^p$ for $p \in \mathbb{N}$, that maps from the space of input features into g-dimensional Euclidean space. Each observation is assigned a vector of continuous class scores or probabilities (depending on the classifier). This mapping must be learned from a set of labeled training data, which marks a fundamental difference to the topic modeling task. The actual class labels $y \in \mathcal{Y}$ are then found by thresholding or an *argmax* operation on the score/probability vectors (Bishop, 2006).
Our case addresses the binary task of classifying tweets into positive and negative sentiment instances. This set of labels, with $g = 2$, is also referred to as sentiment *polarities* and typically encoded as $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \{-1, 1\}$.

Approaches. As mentioned above, casting textual data into a design matrix of numeric features must be handled with care. Once the data are available in tabular form we can, in principle, use any type of learner suited to classification. We specifically consider *random forests* and *regularized logistic regression* for the standard machine learning solution and turn BERT into a classifier by fine-tuning it to sentiment analysis; details are given in section 4.
It should be noted that sentiment analysis can also be performed in a rule-based manner. Such methods abstain from fitting a statistical model, instead classifying documents by summing the number of terms associated with each sentiment and assigning the class with the highest count. Prior polarities are typically taken from large dictionaries (Sidarenka, 2019). In our binary task this corresponds to identifying whether tweets contain more words with positive or negative connotation. We do incorporate this type of analysis but use the respective numbers of positive- and negative-polarity terms as an input feature rather than as the sole grounds for sentiment classification.

A Appendix

B Electronic Appendix

Data, code and figures are provided in electronic form.

References

Aggarwal, C. C. (2018). *Machine Learning for Text*, Springer.

Bhatia, S. and Padmanabhan, D. (2018). Topic-specific sentiment analysis can help identify political ideology, *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 79–84.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Springer.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation, *Journal of Machine Learning Research* **3**: 993–1022.

Bode, L. (2017). Gateway political behaviors: The frequency and consequences of low-cost political engagement on social media, *Social Media + Society* **3**.

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.

Jang, H., Rempel, E., Roth, D., Carenini, G. and Janjua, N. Z. (2021). Tracking covid-19 discourse on twitter in north america: Infodemiology study using topic modeling and aspect-based sentiment analysis, *Journal of Medical Internet Research* **23**(2).

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).

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).

R Core Team (2021). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Roberts, M., Stewart, B., Tingley, D. and Airoldi, 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 Wiegrebe, S. (2020). Twitter in the parliament - a text-based analysis of german political entities, *Technical report*, Ludwig-Maximilians-Universität, Munich.

Schulze, P., Wiegrebe, S., Thurner, P. W., Heumann, C., Aßenmacher, M. and Wankmüller, S. (2021). Exploring topic-metadata relationships with the stm: A bayesian approach.

Sidarenka, U. (2019). *Sentiment Analysis of German Twitter*, PhD thesis, University of Potsdam.

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.

Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N. and Fox, E. A. (2020). Natural language processing advancements by deep learning: A survey, *CoRR* .
URL: *https://arxiv.org/abs/2003.01200*

van Rossum, G. and Drake, F. L. (2011). *The Python Language Reference Manual*, Network Theory Ltd.

van Vliet, L., Törnberg, P. and Uitermark, J. (2020). The twitter parliamentarian database: Analyzing twitter politics across 26 countries, *PLoS ONE* **15**(9).

Vayansky, I. and Kumar, S. A. (2020). A review of topic modeling methods, *Information Systems* **94**.

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.