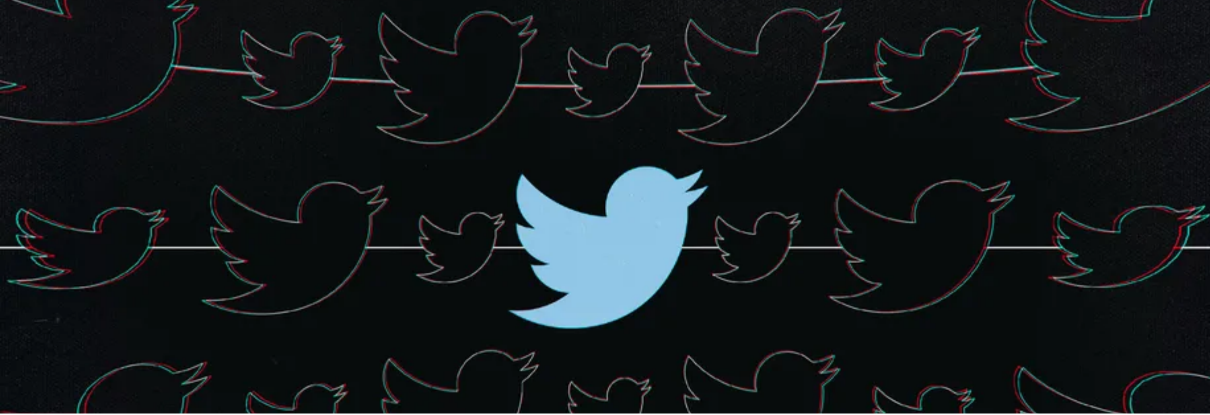

Topic-specific sentiment analysis for tweets by German MPs

Statistical consulting

Asmik Nalmpatian & Lisa Wimmer | July 12th, 2021

REFERENCES

1 INTRODUCTION & PROJECT OUTLINE

1 INTRODUCTION

- Social media: constant stream of publicly available **text data**
- **Twitter** established as a medium for political discourse and constant source of information
- Frequently resurfacing **research questions:**
  - Which **topics** are being addressed?
  - What kind of **sentiment** is expressed about these topics?

1 PROJECT OUTLINE

- **Primary goal:** analysis of public sentiment in a topic-aware manner for posts scraped from Twitter by German Members of Parliament (MPs)

  → Explore how **topic-specific sentiment analysis** can be implemented with (1) standard ML techniques and (2) more complex DL models.

- **Secondary goal:** make analysis of social media texts in a political context more easily accessible to researchers

  → Provide teaching material on both approaches, composed as a coherent online course

2 GENERAL THEORETICAL CONTEXT

2 GENERAL THEORETICAL CONTEXT

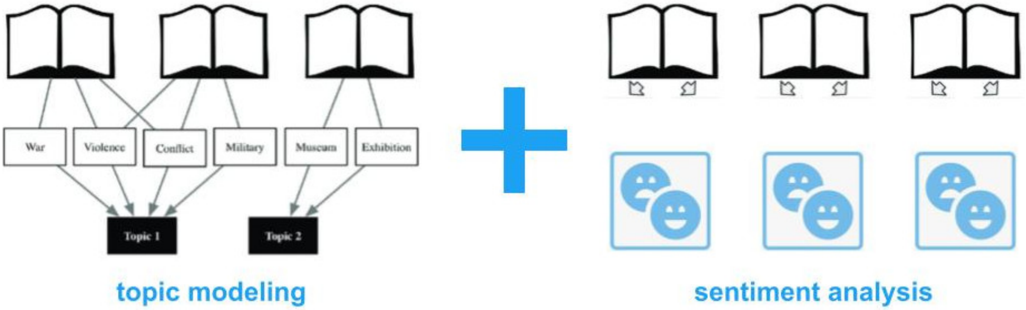

Figure: Adopted and modified from Min and Park (2016)

→ Topic-specific sentiment analysis

2 TOPIC MODELING: IDEA

- **Goal**: discover latent semantic structures in a corpus & group documents into topical clusters with characteristic topic-word distributions
  - Exploratory tool $\rightarrow$ unsupervised learning task
  - Means of dimensionality reduction
- For each document $d \in \{1, 2, \dots, D\}$, assign a topic label $k \in \{1, 2, \dots, K\}$
  - K: key **hyperparameter**
  - Interpretability up to human practitioner

2 TOPIC MODELING: TAXONOMY

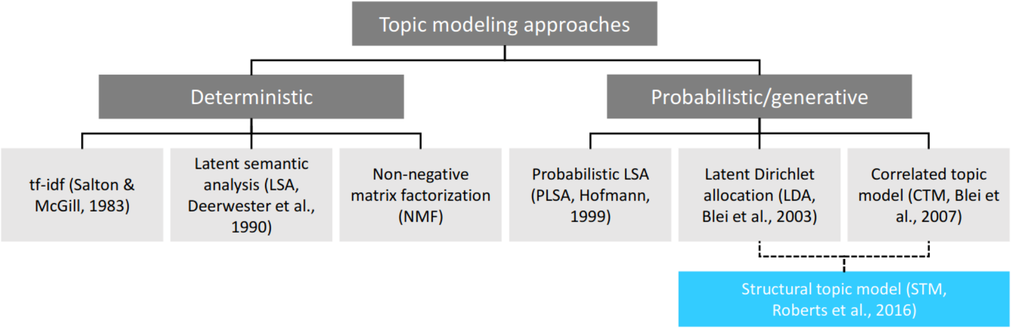

Figure: Source: own representation, published on https://lisa-wm.github.io/nlp-twitter-r-bert/

2 TOPIC MODELING: GENERATIVE APPROACHES

Idea: reverse-engineer the imaginative process of document generation with hierarchical Bayesian mixture models

1. For each document $d \in \{1, 2, \dots, D\}$, draw a vector of topic proportions from some assumed distribution
2. For each word position $n \in \{1, 2, \dots, N_d\}$, $N_d \in \mathbb{N}$,
    1. draw a topic assignment from the distribution associated with the document-specific topic proportions
    2. draw a word from the distribution associated with the topic

2 TOPIC MODELING: GENERATIVE APPROACHES

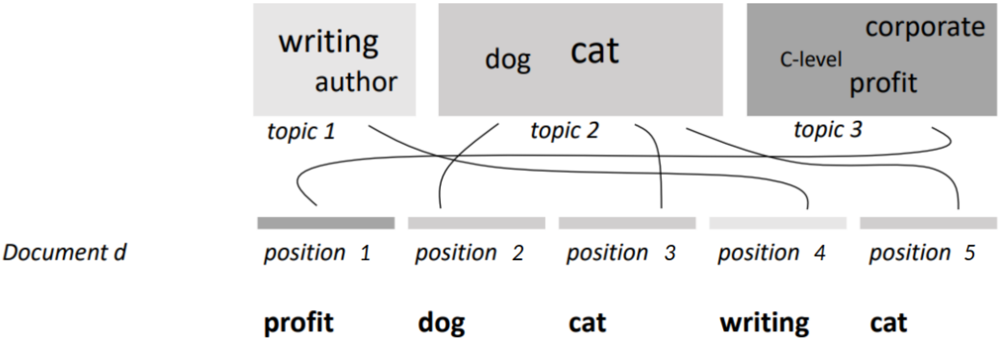

Figure: Source: own representation, published on https://lisa-wm.github.io/nlp-twitter-r-bert/

3 ANALYSIS

4 KNOWLEDGE TRANSFER

5 DISCUSSION

REFERENCES

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

Benoit, K. and Matsuo, A. (2020). *spacyr: Wrapper to the 'spaCy' 'NLP' Library*. R package version 1.2.1.
URL: *https://CRAN.R-project.org/package=spacyr*

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., Matsuo, A., Lowe, W. and Müller, C. (2021). *quanteda: Quantitative Analysis of Textual Data*. R package version 3.0.0.
URL: *https://CRAN.R-project.org/package=quanteda*

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.

Breiman, L., Friedman, J. H., Olshen, R. J. and Stone, C. J. (1984). *Classification and Regression Trees*, Chapman & Hall/CRC.

Devlin, J., Chang, M., Lee, K. and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding, *CoRR* **abs/1810.04805**.
URL: *http://arxiv.org/abs/1810.04805*

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Feurer, M. and Hutter, F. (2019). Hyperparameter optimization, *in* F. Hutter, L. Kotthoff and J. Vanschoren (eds), *Automated Machine Learning. Methods, Systems, Challenges*, Springer, pp. 3–34.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*, MIT Press. http://www.deeplearningbook.org.

Hastie, T., Qian, J. and Tay, K. (2021). An introduction to glmnet.

Japkowicz, N. and Shah, M. (2011). *Evaluating Learning Algorithms. A Classification Perspective*, Cambridge University Press.

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L. and Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R, *Journal of Open Source Software* .
URL: *https://joss.theoj.org/papers/10.21105/joss.01903*

Lindsey, J. K. (1997). *Applying Generalized Linear Models*, Springer.

Louppe, G. (2014). *Understanding Random Forests. From Theory to Practice*, PhD thesis, University of Liege.

Min, S. and Park, J. (2016). Mapping out narrative structures and dynamics using networks and textual information.

Murphy, K. P. (2021). *Probabilistic Machine Learning: An Introduction*, MIT Press.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning, *IEEE Transactions on Knowledge and Data Engineering* **22**(10): 1345–1359.

Pavlopoulos, I. (2014). *Aspect-Based Sentiment Analysis*, PhD thesis, Athens University of Economics and Business.

Pennington, J., Socher, R. and Manning, C. (2014). GloVe: Global vectors for word representation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543.
URL: *https://www.aclweb.org/anthology/D14-1162*

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

Richardson, L. (2007). *Beautiful Soup Documentation*.

Roberts, M. E., Stewart, B. M. and Airoldi, E. M. (2016). A model of text for experimentation in the social sciences, *Journal of the American Statistical Association* **111**(515): 988–1003.

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.

Roberts, M., Stewart, B., Tingley, D. and Benoit, K. (2020). *stm: Estimation of the Structural Topic Model*. R package version 1.3.6.
URL: *https://CRAN.R-project.org/package=stm*

Roesslein, J. (2020). *Tweepy: Twitter for Python!*

Ruder, S. (2019). *Neural Transfer Learning for Natural Language Processing*, PhD thesis, National University of Ireland, Galway.

Schulze, P. and Wiegrebe, S. (2020). Twitter in the parliament - a text-based analysis of german political entities, *Technical report*, Ludwig-Maximilians-Universität, Munich.

Selivanov, D., Bickel, M. and Wang, Q. (2020). *text2vec: Modern Text Mining Framework for R*. R package version 0.6.
URL: *https://CRAN.R-project.org/package=text2vec*

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017). Attention is all you need, *CoRR* **abs/1706.03762**.
URL: *http://arxiv.org/abs/1706.03762*

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

Xu, H., Liu, B., Shu, L. and Yu, P. S. (2019). Post-training for review reading comprehension and aspect-based sentiment analysis, *Proceedings of NAACL-HLT*, Minneapolis, USA, p. 2324–2335.

Zhang, A., Lipton, Z. C., Li, M. and Smola, A. J. (2020). *Dive into Deep Learning*. https://d2l.ai.