

Exposé einer Projektarbeit zum Thema:

Sentimentanalyse von Politischer Kommunikation im Deutschen Bundestag mit Python

Fragestellung:

In den vergangenen Jahren hat die Sentimentanalyse und das *Opinion Mining* von Textdaten zunehmend an Popularität gewonnen und kommt in verschiedensten Bereichen von Marktforschung bis Wahlkampfmonitoring zum Einsatz. Charakteristisch für die Analyse ist oft, dass vor allem große Mengen kurzer Texte, z.B. Tweets auf Twitter, untersucht werden, um Tendenzen aus den Daten extrahieren zu können.

Diese Projektarbeit soll eine etwas andere Herangehensweise zeigen, in der zwar große Mengen an Textdaten verarbeitet werden, die Analyse aber hoffentlich auch spezifische Aussagen über einzelne, längere, Texte fördert. Dies stellt – wie auch das betreffende Feld der politischen Kommunikation – durchaus domänenspezifische Anforderungen, wie schon das Sichten der Forschungsliteratur in diesem Bereich andeutet.

Schlussendlich soll so also wiederum beantwortet werden, welche Muster sich mit einer Sentimentanalyse von Bundestagsreden entdecken lassen, indem man diese Auswertung mit Metadaten kombiniert, mit denen sich z.B. Redner in Fraktionen einteilen lassen und Reden in Themenkomplexe eingeordnet werden.

Ziel der Arbeit:

Ziel ist das Entwickeln eines Tools zur Sentimentanalyse von Reden aus der aktuellen und eventuell auch zukünftigen Wahlperioden des Deutschen Bundestags. Zuvorderst soll ein Baseline-System auf Basis von bestehenden Ressourcen wie Sentiment-Lexika entworfen werden, mit dem eine grundsätzliche Analyse möglich ist. Darauf aufbauen ließe sich sowohl in Form eines interaktiven Tools, z.B. einer Web-App, und weiteren methodologischen Überlegungen und der Herangehensweise an das Problem mit state-of-the-art Ansätzen des maschinellen Lernens.

Von der resultierenden Analyse erhoffe ich mir vor allem Erkenntnisse in zwei Richtungen. Einerseits lässt sich durch das Aggregieren von Reden einzelner Sprecher oder Fraktionen deren Haltung zu bestimmten Themenbereichen oder Akteuren skizzieren (allerdings wäre wohl für die verlässliche Auswertung von Relationen noch die Extraktion anderer Informationen nötig). Zweitens könnte auch die Natur der Bundestagsreden untersucht werden, ob also bestimmte Themen besonders polarisieren und ob sich Strukturen innerhalb des Verlaufes der Reden entdecken lassen (z.B. dass stets mit besonders polarisierenden Segmenten geschlossen wird).

Zusätzlich soll natürlich auch ein Überblick über das Forschungsfeld der Sentimentanalyse gegeben werden, um die Arbeit am Projekt besser organisierbar zu machen und voraussichtlich auch, um eine etablierte Methodik als Leitlinie festzulegen.

Voraussetzungen

Als Datenbasis für die Analyse sollen voraussichtlich Protokolle der ersten 150 Sitzungen (die 150. ist noch ausstehend) der 19. Wahlperiode des Deutschen Bundestags dienen. Das Tool soll jedoch mit Blick auf kontinuierliches Hinzufügen von neuen Daten in den Bestand umgesetzt werden.

Die Bundestagsreden sind in den Plenarprotokollen enthalten, die zu jeder Sitzung des Bundestags erstellt werden. Die Plenarprotokolle sind im XML-Format verfügbar, das zugehörige DTD-Schema

ebenfalls (dieses ist auch ausführlich dokumentiert¹). Ein XML-Dokument enthält jeweils ein Transkript einer Sitzung. Verschiedene Metadaten, z.B. zu Datum, Tagesordnung, aktuellem Redner, usw. sind in den Dokumenten enthalten und können z.B. extrahiert werden, um die Äußerungen eines Politikers über viele Sitzungen hinweg zu aggregieren und zu verfolgen.

Vorläufige Gliederung:

1. Einleitung
 - 1.1. Zielsetzung
 - 1.2. Endprodukt der Projektarbeit
 - 1.3. Gliederung der Dokumentation
2. Theoretischer Hintergrund
 - 2.1. Sentimentanalyse und *Opinion Mining*
 - 2.1.1. Ursprung der Sentimentanalyse
 - 2.1.2. Anwendungsgebiete
 - 2.2. Methodologien der Sentimentanalyse (historisch und state-of-the-art)
 - 2.3. Domänenspezifische Anforderungen
 - 2.4. Analyse von politischen Dokumenten mithilfe von automatischer Textanalyse
3. Dokumentation der Projektarbeit: Tool zur Sentimentanalyse von Bundestagsreden
 - 3.1. Methodologische Basis
 - 3.1.1. Genutzte Frameworks, usw.
 - 3.1.2. Eigenschaften der genutzten Daten
 - 3.1.3. Projektpipeline
 - 3.2. Entwicklung des Tools
 - 3.3. Funktionsweise des Tools
- (4. Evaluation)
5. Fazit und Ausblick (evtl. Erweiterungsmöglichkeit in Richtung *Opinion Mining*, usw.)

Literatur

- Bird, Steven, Ewan Klein und Edward Loper (2009). Natural Language Processing with Python. O'Reilly.
- Breck, Eric und Claire Cardie (2017). „Opinion Mining and Sentiment Analysis“. In: The Oxford Handbook of Computational Linguistics. Hrsg. von Ruslan Mitkov. Oxford University Press. doi: 10.1093/oxfordhb/9780199573691.013.43.
- Busch, Andreas und Madeline Kaupert (2018). „Die Regierungserklärungen deutscher Bundeskanzler von 1949 bis 2018 im Spiegel automatisierter Textanalyse“. In: Zeitschrift für Politikwissenschaft 28, S. 359–370.
- Cambria, Erik u. a., Hrsg. (2017). A Practical Guide to Sentiment Analysis. Socio-Affective Computing 5. Springer.
- Goodfellow, Ian, Yoshua Bengio und Aaron Courville (2016). Deep Learning. MIT Press.
- Grimmer, Justin und Brandon M. Stewart (2013). „Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts“. In: Political Analysis 21.3, S. 267–297.
- Haselmayer, Martin und Marcelo Jenny (2017). „Sentiment analysis of political communication: combining a dictionary approach with crowdcoding“. In: Quality and Quantity 51, S. 2623–2646.

1 https://www.bundestag.de/resource/blob/577234/f9159cee3e045cbc37dcd6de6322fcdd/dbtplenarprotokoll_kommentiert-data.pdf

- Kan, Hanhoon, Seong Joon Yoo und Dongil Han (2012). „Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews“. In: Expert Systems with Applications 39.5, S. 6000–6010.
- Liu, Bing (2015). Sentiment Analysis. Mining Opinions, Sentiments, and Emotions. Cambridge University Press.
- Mullen, Tony und Robert Malouf (2006). „A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse“. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, S. 159–162.
- Nasukawa, Tetsuya und Jeonghee Yi (2003). „Sentiment Analysis: Capturing Favorability Using Natural Language Processing“. In: Proceedings of the K-CAP-03, 2nd International Conference on Knowledge Capture.
- Pang, Bo und Lillian Lee (2008). „Opinion Mining and Sentiment Analysis“. In: Foundations and Trends in Information Retrieval 2.1-2, S. 1–135.
- Pang, Bo, Lillian Lee und Shivakumar Vaithyanathan (2002). „Thumbs Up? Sentiment Classification Using Machine Learning Techniques“. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2002).
- Rauh, Christian (2018). „Validating a sentiment dictionary for German political language - a workbench note.“ In: Journal of Information Technology & Politics 15.4, S. 319–343.
- Stone, Philip (1968). „The General Inquirer: A Computer Approach to Content Analysis.“ In: Journal of Regional Science 8.1, S. 113–116.
- Tan, Songbo u. a. (2009). „Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis“. In: ECIR 2009, LNCS 5478.
- Turney, Peter D. (2002). „Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews“. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). S. 417–424.
- Wiebe, Janyce u. a. (2004). „Learning Subjective Language“. In: Computational Linguistics 30.3, S. 277–308.