PoliAna: Nachrichten und Politikanalyse mittels Natural Language Processing

Leopold Fuchs, Felix Hoffmann

2023-01-09

Inhaltsverzeichnis

1 Kurzbeschreibung der Arbeit

Ereignisse wie die Wahl von Donald Trump zum US-Präsidenten, das Brexit-Referendum, als auch der Krieg in der Ukraine werfen die Frage auf, welchen Einfluss die Neuen Medien auf die politische Meinungsbildung haben [2,8]. Plattformen wie Twitter, Facebook und Instagram haben in den letzten Jahren die Art und Weise, wie Nachrichten erstellt und verbreitet werden, stark verändert. Unter anderem die Möglichkeit eines Nutzers, seine Meinung zu Themen schnell und direkt zu äußern, rückt den konkreten sprachlichen Ausdruck sowie polarisierende Meinungen von Politikern deutlich stärker in den Fokus.

Bisherige Arbeiten beschäftigen sich meist mit englischsprachigen Daten aus den USA oder Großbritannien. Außerdem umfassen Arbeiten wie die von Sältzer und Stier über die Bundestagswahl 2021 lediglich Tweets von Twitter [12]. Dennoch wird in der Arbeit von Sältzer und Stier gezeigt, dass es möglich ist, Trends zu analysieren und die Parteizugehörigkeit eines Politikers anhand seiner Tweets zu klassifizieren.

Ziel dieser Arbeit ist es, mittels Natural Language Processing und Machine Learning die Meinung der Bevölkerung zu einem einzelnen Politiker oder Parteien in Deutschland zu analysieren und ein besseres Verständnis in die Reaktionen und Meinungen der Bevölkerung zu bekommen. Dafür sollen im Vergleich zu bisherigen Arbeiten Daten aus mehreren Quellen einbezogen werden.

Mögliche Fragestellungen sind für diese Arbeit sind:

- Lassen sich Trends in der Semantik einzelner Politiker feststellen?
- Ist es möglich, auf Basis der Semantik festzustellen, ob ein Politiker zum Flügel einer Partei gehört?
- Können Politiker anhand ihrer Nachrichten und Daten klassifiziert werden?
- Wie stark ist die Übereinstimmung der Klassifikation mit der wahrhaftigen Parteizugehörigkeit?
- Unterstützen Nutzer ebenfalls die Politiker, welche am nächsten an ihrer Meinung sind?
- Lässt sich die Parteizugehörigkeit aufgrund einzelner Worte oder Phrasen bestimmen?

2 Gliederung und Zeitplan

Als erste Gliederung der Arbeit kann die folgende Struktur angenommen werden:

- Einleitung
  - Problemstellung
  - Zielsetzung
  - Methodik
  - Struktur der Arbeit
- Grundlagen
  - Politikapparat und Parteienlandschaft
  - Erörterung, welche Medienplattformen und Nachrichtenquellen verwendet werden sollen
  - Auswahl von Quellen und Sammeln von Daten
  - NLP-Pipeline
  - Machine Learning / Clustering
- Datenbeschaffung und -analyse nach CRISP-DM
  - *Business Understanding*
    * Festlegung auf spezifische Ereignisse, die näher untersucht werden sollen
  - *Data Understanding*
    * Identifizieren von geeigneten Nachrichtenquellen
    * Sammeln der Trainingsdaten aus verschiedenen Quellen
  - *Data Preparation*
  - *Modeling* (Unterteilung in verschiedene Thesen und Analysen)
    1. Trainieren eines NLP-Modells, das Texte nach Parteien klassifiziert
    2. Analysieren von Nachrichten und Tweets zu ausgewählten Ereignissen
  - *Evaluation*
    * Vergleich und Evaluation der trainierten Modelle

        * Mathematische Betrachtung der Ergebnisse zur Trenddetektion
    - *Deployment*
        * Bereitstellung des besten Modells für die weiteren Analysen
        * (grafische) Darstellung der Analyseergebnisse

- Fazit
    - Zusammenfassung
    - Ausblick

Der Hauptteil der Arbeit wird nach dem CRISP-DM Prozessmodell aufgebaut. Die einzelnen Schritte stellen dabei die wesentlichen Meilensteine dar, die während des Projekts erreicht werden sollen.

In den kommenden drei Monaten soll die praktische Implementierung des Projekts abgeschlossen werden. Dabei sollen auch parallel die wesentlichen Inhalte, die praktisch umgesetzt werden, in der schriftlichen Arbeit dokumentiert werden. Nach Abschluss der praktischen Arbeit soll bis zum Ende der Gesamt-Projektlaufzeit die Vollendung und Optimierung der schriftlichen Arbeit erfolgen.

3 Grundlegende Literatur

Eine gute Übersicht über die notwendigen politikwissenschaftlichen Hintergründe bieten Bukow et al., die die Organisation und Funktion der deutschen Parteien beschreiben [3].

Kalyanam et al. untersuchen die Auswirkungen von Events in der realen Wert auf die Social Media Aktivität [6]. Ebenso untersuchen Tsytsarau et al. diesen Zusammenhang, mit einem Fokus auf die Verbindung der medialen Berichterstattung und dem Sentiment, der durch die Social Media Aktivitäten dargestellt wird [13]. Gimpel et al. nähern sich der Thematik der Social Media Nutzung durch eine Cluster-Analyse zu verschiedenen Rollen in Twitter-Diskussionen [5]. Zudem untersucht Sältzer die Positionen von Bundestagskandidaten auf Twitter und betrachtet diese einerseits im Vergleich innerhalb der Parteien sowie andererseits auf einem allgemeinen politischen Koordinatensystem [11,12].

Li et al. sowie Kowsari et al. bieten je einen umfassenden Überblick sowie Vor- und Nachteile verschiedener Arten von Text-Klassifikation und gehen dabei sowohl auf traditionelle als auch auf Deep-Learning-Ansätze ein [7,9]. Minaee et al. untersuchen und vergleichen die Verwendung von Deep-Learning-Modellen für die Aufgabe der Text-Klassifikation [10].

Wong et al. bestimmen die politische Ausrichtung aufgrund des Verhaltens von Personen auf Twitter durch die Betrachtung, welche Accounts ähnliche andere Accounts retweeten [14]. Zudem vergleichen Doan et al. verschiedene Sprachmodelle zur Klassifizierung von Reden nach Parteien und führen dies für verschiedene Länder bzw. Sprachen durch [4]. Auch Biessmann et al. klassifizieren Reden nach Parteien und nutzen zum Trainieren Parlaments-Debatten des Bundestages. Zudem wenden sie den Klassifikator auf andere Arten von Texten wie Social Media Posts an [1].

Die Darstellung der Literatur zeigt, dass sich bereits einige Arbeiten mit der Analyse von Social Media Aktivitäten im Zusammenhang mit realen Events sowie mit dem Problem, Texte nach Parteizugehörigkeit zu klassifizieren, beschäftigen. In unserer Studienarbeit wollen wir neue Klassifikationsverfahren nutzen und vergleichen, um eine höhere Performance als vergangene Arbeiten zu erreichen. Zudem wollen wir nicht nur die Texte von Reden nutzen, sondern auch Social Media Posts und Parteiprogramme als Trainingsdaten für den Klassifikator einbeziehen.

Für die darauf aufbauenden Analysen wollen wir zusätzlich zu den Social Media Aktivitäten, die mit bestimmten Ereignissen in Verbindung stehen, auch die jeweilige politische Einstellung der Nutzer, bestimmt durch den trainierten Klassifikator, nutzen.

4 Literaturverzeichnis

[1] Felix Biessmann, Pola Lehmann, Daniel Kirsch, und Sebastian Schelter. 2016. Predicting political party affiliation from text. (2016).

[2] John Brandon. 2022. Russia Has Radically Redefined The Term ‚Fake News' During The Ukraine War. Forbes. Abgerufen 7. Januar 2023 von https://www.forbes.com/sites/johnbbrandon/2022/07/31/russia-has-radically-redefined-the-term-fake-news-during-the-ukraine-war/

[3] Sebastian Bukow und Thomas Poguntke. 2013. Innerparteiliche Organisation und Willensbildung. In *Handbuch Parteienforschung*, Oskar Niedermayer (Hrsg.). Springer Fachmedien, Wiesbaden, 179–209. DOI:https://doi.org/10.1007/978-3-531-18932-1_6

[4] Tu My Doan, Benjamin Kille, und Jon Atle Gulla. 2022. Using Language Models for Classifying the Party Affiliation of Political Texts. In *Natural Language Processing and Information Systems* (Lecture Notes in Computer Science), Springer International Publishing, Cham, 382–393. DOI:https://doi.org/10.1007/978-3-031-08473-7_35

[5] Henner Gimpel, Florian Haamann, Manfred Schoch, und Marcel Wittich. 2018. USER ROLES IN ONLINE POLITICAL DISCUSSIONS: A TYPOLOGY BASED ON TWITTER DATA FROM THE GERMAN FEDERAL ELECTION 2017. (2018).

[6] Janani Kalyanam, Mauricio Quezada, Barbara Poblete, und Gert Lanckriet. 2016. Prediction and Characterization of High-Activity Events in Social Media Triggered by Real-World News. *PLOS ONE* 11, 12 (Dezember 2016). DOI:https://doi.org/10.1371/journal.pone.0166694

[7] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, und Donald Brown. 2019. Text Classification Algorithms: A Survey. *Information* 10, 4 (April 2019), 150. DOI:https://doi.org/10.3390/info10040150

[8] Jenna Marina Lee. 2020. How Fake News Affects U.S. Elections. University of Central Florida. Abgerufen 7. Januar 2023 von https://www.ucf.edu/news/how-fake-news-affects-u-s-elections/

[9] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, und Lifang He. 2021. A Survey on Text Classification: From Shallow to Deep Learning. Abgerufen 3. Januar 2023 von http://arxiv.org/abs/2008.00364

[10] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, und Jianfeng Gao. 2022. Deep Learning–based Text Classification: A Comprehensive Review. *ACM Computing Surveys* 54, 3 (April 2022), 1–40. DOI:https://doi.org/10.1145/3439726

[11] Marius Sältzer. 2022. Finding the bird's wings: Dimensions of factional conflict on Twitter. *Party Politics* 28, 1 (Januar 2022), 61–70. DOI:https://doi.org/10.1177/1354068820957960

[12] Marius Sältzer und Sebastian Stier. 2022. Die Bundestagswahl 2021 auf Twitter. *easy social sciences No. 67 (2022): Die Bundestagswahl 2021: Perspektiven und Daten aus der deutschen Wahlstudie* (2022). DOI:https://doi.org/10.15464/EASY.2022.05

[13] Mikalai Tsytsarau, Themis Palpanas, und Malu Castellanos. 2014. Dynamics of news events and social media reaction. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '14), Association for Computing Machinery, New York, NY, USA, 901–910. DOI:https://doi.org/10.1145/2623330.2623670

[14] Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, und Mung Chiang. 2016. Quantifying Political Leaning from Tweets, Retweets, and Retweeters. *IEEE Transactions on Knowledge and Data Engineering* 28, 8 (August 2016), 2158–2172. DOI:https://doi.org/10.1109/TKDE.2016.2553667