

Fakultät Informatik

Topic Modelling on Large Multilingual Twitter Data Set on Ukraine War

Expose Bachelorarbeit im Studiengang Informatik

vorgelegt von

Robin Feldmann

Matrikelnummer 3538270 - Ca. 9000 Zeichen

© 2024

Dieses Werk einschließlich seiner Teile ist **urheberrechtlich geschützt**. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Autors unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen sowie die Einspeicherung und Verarbeitung in elektronischen Systemen.

Inhaltsverzeichnis

Introduction	2
Objective	3
State of Research	3
Methodik	4
Vorgehensweise	5
Literaturverzeichnis	7

Introduction

The ongoing conflict between Ukraine and Russia represents a paramount challenge in the realm of external politics for both Germany and Europe. This confrontation not only tests diplomatic relations but also significantly influences the geopolitical landscape of the region, underscoring its critical importance [1]. In the context of the Russian-Ukraine War, social media serves a dual role: it acts as a crucial medium for disseminating political opinions and information, and simultaneously, it emerges as a vital source of information relevant to military strategies and operations. [2]. Twitter holds a position of exceptional significance in this context, underscored by its expansive user base exceeding 350 million in 2022 and 2023 [3]. This platform has previously demonstrated its influential role in shaping discourse around major global events, notably during the Brexit referendum and the COVID-19 pandemic. Its pervasive reach and real-time information dissemination capabilities make it a critical tool for political engagement and public awareness in such high-stake scenarios [4–7]. Natural Language Processing methods are particularly effective for analyzing large datasets of tweets, where the sheer

volume of data renders manual reading unfeasible. These techniques enable the extraction of structured information, opinions, and factual insights from these extensive text corpora, transforming vast digital conversations into comprehensible data. [8]. Especially Topic Modelling is a method that aims to find latent Topics and Structures inside text corpora and to group them in a semantic logical way [9]. Multilingual Topic Modelling expands this technique to multi lingual text corpora.

Objective

This thesis aims to critically evaluate the effectiveness of contemporary multilingual topic modeling techniques in extracting meaningful insights from extensive multilingual Twitter datasets, specifically focusing on the discourse surrounding the Russian-Ukraine war. It seeks to determine whether these advanced computational methods are adept at navigating the complexities and nuances inherent in large-scale, multilingual social media data, thereby offering a robust tool for understanding the dynamics of digital communication in the context of international conflicts.

To achieve this objective, the thesis will first process and analyze a substantial multilingual Twitter dataset obtained from Kaggle, focusing on content related to the Russian-Ukraine conflict. Subsequently, it will provide a comprehensive overview of the current developments in multilingual topic modeling, highlighting the advancements and challenges in this field. Finally, the study will apply suitable multilingual topic modeling techniques to the dataset, aiming to effectively discern and interpret the patterns and themes that emerge from this complex digital discourse.

State of Research

Maathuis and Karkhof's study utilizes Latent Dirichlet Allocation and Non-Negative Matrix Factorization, coupled with Kullback-Leibler Divergence, to analyze a Telegram-sourced dataset on the Ukraine war, revealing key themes and patterns in digital communication. In their initial approach, Maathuis and Karkhof analyzed the first two months

of the Ukraine war using fewer than 10,000 messages, later expanding their study to cover the first six months with an extended dataset of fewer than 50,000 messages [10, 11]. Sazed employs Latent Dirichlet Allocation and BERTopic to analyze approximately 80,000 tweets, subsequently integrating these findings with sentiment analysis for a comprehensive understanding of public discourse [12]. Nayak, JVN, and Bhagat apply Latent Dirichlet Allocation to scrutinize about 1,200 articles from news websites, focusing on the textual analysis of the Ukraine-Russian war [13].

Methodik

Datengrundlage für die Arbeit ist der Datensatz über Integrierte Erwerbsbiografien (IEB) des Instituts für Arbeitsmarkt und Berufsforschung. Dieser Datensatz enthält Erwerbsverläufe aller Personen die in Deutschland seit 1975 sozialversicherungspflichtig beschäftigt war. Er enthält Informationen über Geburtsjahr und Ausbildung, Leistungsbezug und Beschäftigung [14]. Ein weiterer Datensatz, zu dem zum Zeitpunkt dieses Expose noch keine genaueren Informationen erhältlich sind wird historische Daten über offene Stellenangebote beinhalten. In diesem Datensatz wird zu den Stellenangeboten Informationen über deren Erstellungsdatum und Aufslaufdatum, über den Betrieb und in welche Kategorie die Arbeitsstelle fällt, beinhalten. In einem ersten Schritt sollen die Erwerbsbiografien des IEB anhand der Betriebe, der Zeitpunkte und der Art der Beschäftigung mit historischen Stellenangeboten identifiziert werden. Es werden also paare von Datenpunkte erstellt, bei denen sichergestellt ist, dass eine spezifische Arbeitsbiografie eine spezifische Tätigkeit beinhaltet, die vorher als Jobangebot ausgeschrieben war. Diese Kombination aus zusammengehörigen Erwerbsbiografien und Stellenangeboten wird die Grundlage für das Training eines Algorithmus des maschinellen Lernens bilden.

Da als Vorbereitung bereits passende Verknüpfungen zwischen Informationen über Job-suchenden aus dem Datensatz über Erwerbsbiographien und Informationen aus dem Datensatz über Stellenangebote hergestellt wurden, handelt es sich um *überwachtes Lernen*. Das Schätzen von Wahrscheinlichkeiten also von reellwertigen Labels bezeichnet man als Regressionsproblem [15]. Im Laufe der Arbeit sollen verschiedene Algorithmen des überwachten Lernens zum lösen von Regressionsproblemen verglichen und anhand von vorher definierten Metriken bewertet werden. Zum jetzigen Zeitpunkt ist geplant drei solcher Algorithmen zu vergleichen: Support Vector Machine (SVM), Random Forest

(RF) und XGBoost (XGB). Diese Liste kann und soll aber im Laufe der Arbeit noch erweitert werden. In der Recherchephase der Bachelorarbeit sollen weitere mögliche Algorithmen anhand des aktuellen Forschungsstands ermittelt werden.

Beim Support Vector Machine Algorithmus werden Datenpunkte in einem Vektorraum dargestellt. Es wird versucht möglichst große Hyperebenen zu finden, welche Klassen von Datenpunkten separieren. Beim Klassifizieren neuer Datenpunkte wird dann überprüft auf welcher Seite der Hyperebene dieser Datenpunkt liegt [16].

Der Random Forest ist ein Algorithmus bei dem der Datensatz zufällig in mehrer Samples zerteilt wird. Auf jedem dieser Samples wird ein Entscheidungsbaum erstellt und bei der Klassifikation wird der Durchschnitt der einzelnen Entscheidungen der Entscheidungsbäume ermittelt [17].

Der XGBoost Algorithmus ist ein *gradient boosting* verfahren. Hierbei werden Modelle iterativ erweitert um ihre jeweiligen Schwächen auszubauen. Da hierbei auch mit Entscheidungsbäumen gearbeitet wird, ist er eine Erweiterung des Random Forest Algorithmus [18].

Zur Evaluation der Modelle wird der Datensatz in einen Trainingsdatensatz und einen Testdatensatz aufgeteilt. Hier sind Aufteilungen von 80% als Trainingsdaten und 20% als Testdaten üblich. Zur Bewertung werden Datenpunkte des Testdatensatzes ausgewertet und die so erhaltenen Werte mit den tatsächlichen Labels verglichen. Diese werden Anhand gebräuchlicher Metriken bewertet. Hierfür üblich sind Mean Squared Error und der R-Score. Mean Squared Error meint hier den durchschnittlichen aufsummierten Quadratischen Fehler. Der R-Score berechnet sich als eins minus der Mean Squared Error geteilt durch die quadratische Standardabweichung [19]. Weitere Metriken zur Bewertung der Modelle sollen im Laufe der Recherchephase ermittelt werden.

Vorgehensweise

Der Prozess eine Bachelorarbeit zu verfassen ist aufwendig. Um dieses große Aufgabe sinnvoll anzugehen wird die Arbeit in kleiner Abschnitte unterteilt. Die Bearbeitung der Bachelorarbeit soll in einer groben Einteilung fünf Schritte umfassen.

In der Recherchephase soll der aktuelle Forschungsstand ermittelt werden. Außerdem sollen Informationen zur aktuellen Methodik gefunden werden. Für diese Phase ist in etwa einen Monat Zeit eingeplant.

In der zweiten Phase werden die Datensätze begutachtet. Zuerst werden die Datensätze explorativ analysiert. Hier soll die Qualität der Datensätze überprüft werden. Auch Infor-

mationen über die einzelnen Felder der Daten müssen gesammelt werden. Dann müssen die beiden Datensätze miteinander identifiziert werden. Hierbei soll herausgefunden werden welche Stellenausschreibung mit welcher Erwerbsbiografie in Zusammenhang steht. Da die Identifikation nur aufgrund der Betriebe und Zeitpunkte passiert, ist hier besondere Vorsicht geboten. Auch hierbei muss überprüft werden wie gut diese Identifikation funktioniert. Für diese Phase ist auch etwa ein Monat Zeit eingeplant.

Anhand im ersten Schritt recherchierten Methodik sollen nun verschiedene Modelle des maschinellen Lernens anhand der im zweiten Schritt identifizierten Datensätze trainiert werden. Auch müssen verschiedene Parameter der Algorithmen angepasst und optimiert werden. Dafür sind bis zu zwei Monate an Zeit vorgesehen.

Bei der Evaluierung sollen die verschiedenen im dritten Schritt trainierten Modelle verglichen und bewertet werden. Dies geschieht anhand von Metriken, die bei der Recherche der Methodik im ersten Schritt gefunden wurden. Nun sollen Vor- und Nachteile der Modelle anhand dieser Metriken gegenübergestellt werden. Dies sollte ungefähr zwei Wochen Zeit in Anspruch nehmen.

In der letzten Phase sollen die in den vorherigen Schritten gewonnenen Erkenntnisse niedergeschrieben werden. Für das Verfassen der Bachelorarbeit sind in etwa ein Monat an Zeit vorgesehen.

Verläuft alles nach Plan sollte die Bachelorarbeit also in etwa fünf Monaten zu schaffen sein.

Literaturverzeichnis

- [1] OLAF SCHOLZ: *Zeitenwende - Regierungserklärung in der Sondersitzung zum Krieg gegen die Ukraine vor dem Deutschen Bundestag.* [Online]. <https://www.bundesregierung.de/resource/blob/992814/2131062/78d39dda6647d7f835bbe76713d30c31/bundeskanzler-olaf-scholz-reden-zur-zeitenwende-download-bpa-data.pdf>. Version: 2022. – Zugriff am: 2024-01-08.
- [2] PAUL STROBEL ; KONRAD ADENAUER STIFTUNG (Hrsg.): *Die sozialen Medien im Ukraine Krieg.* [Online]. <https://www.kas.de/de/web/die-politische-meinung/blog/detail/-/content/die-sozialen-medien-im-ukrainekrieg>. – Zugriff am: 2024-01-08.
- [3] STATISTA: *Number of X (formerly Twitter) users worldwide from 2019 to 2024.* [Online]. <https://www.statista.com/statistics/303681/twitter-users-worldwide/>. – Zugriff am: 2024-01-08.
- [4] MIHA GRČAR, DARKO CHEREPNALKOSIK, IGOR MOZETIC, PETRA KRAJ NOVAK: Stance and influence of Twitter users regarding the Brexit referendum. In: *Comput Soc Netw* (2017).
- [5] AMIRARSALAN RAJABI, ALEXANDER V. MANTZARIS, KULDIP SINGH ATWAL, IVAN GARIBAY: Exploring the disparity of influence between users in the discussion of Brexit on Twitter. In: *Journal of Computational Social Science* (2021).
- [6] SAKUN BOON-ITT, YUKOLPAT SKUNKAN: Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. In: *JMIR PUBLIC HEALTH AND SURVEILLANCE* (2021).
- [7] WEIWU ZHANG, THOMAS J. JOHNSON, TRENT SELTZER, SHANNON L. BICHARD: The Revolution will be Networked - The Influence of Social Networking Sites on Political Attitudes and Behavior. In: *Social Science Computer Review* (2010).
- [8] SONJA GRIGOLEIT: Natural Language Processing. In: *Europäische Sicherheit und Technik* (2009).

- [9] DAVID M. BLEI, JOHN D. LAFFERTY: Topic Models. In: *Text Mining* (2009).
- [10] CLARA MAATHUIS, IDDO KERKHOF: The first two months in the war in Ukraine through topic modeling and sentiment analysis. In: *Reg. Sci Policy Pract* (2023).
- [11] CLARA MAATHUIS, IDDO KERKHOF: First Six Months of War from Ukrainian Topic and Sentiment Analysis. In: *Reg. Sci Policy Pract* (2023).
- [12] SALIM SAZZED: The Dynamics of Ukraine-Russian Conflict through the Lens of Demographically Diverse Twitter Data. In: *IEEE International Conference on Big Data* (2022).
- [13] FAHIM SUFI: Social Media Analytics on Russia-Ukraine Cyber War with Natural Language Processing: Perspectives and Challenges. In: *MDPI* (2023).
- [14] BUNDESMINISTERIUM FÜR ARBEIT UND SOZIALES: *Stichprobe der Integrierten Erwerbsbiografien des IAB*. [Online]. <https://kooperationen.zew.de/dfgflex/links/datensaetze-deutschland/stichprobe-der-integrierten-erwerbsbiografien-des-iab-ies>. – Zugriff am: 2023-05-25.
- [15] M. MOHRI, A. ROSTAMIZADEH, A. TALWALKAR: *Foundations of Machine Learning*. London, England : MIT Press, 2018.
- [16] A. MAMMONE, M. TURCHI, N. CHRISTIANINI: Support vector machines. In: *Journal of statistical software* 1 (2009).
- [17] LEO BREIMAN: Random Forests. In: *Machine Learning* 45 (2001), S. 5–35.
- [18] T. CHEN, C. GUESTRIN: XGBoost: A Scalable Tree Boosting System. (2016).
- [19] A. BOTCHKAREV: Performance Metrics in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology. (2018), S. 4.