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

Inhaltsverzeichnis

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 the context of the Russian-Ukraine conflict, 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 ideal for extracting structured information, opinions,

and facts from large text corpora that are beyond human readability due to their size [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]. Multilangual 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

Die Arbeit wird im Zuge eines Praktikums beim Institut für Arbeitsmarkt- und Berufsforschung unter der Betreuung von Fr. Sabrina Mühlbauer verfasst. Diese Arbeit soll auf der Veröffentlichung *Machine Learning for Labour Market Matching* aufbauen [10]. In dieser Arbeit wird ein Ansatz beschrieben anhand von statistischen Methoden und Methoden des maschinellen Lernens Arbeitssuchende bei der Auswahl des zukünftigen

Berufsfelds zu unterstützen. Hierfür wird ein Datensatz über Berufsbiographien verwendet um anhand der Berufshistorie einer Person Aussagen über das zukünftige Berufsfeld treffen zu können. Hierbei wird die Effektivität von statistischen Methoden mit dem Random Forest und dem K-Nearest-Neighbours Algorithmus verglichen. Fr. Mühlbauer kommt zu dem Ergebnis, dass der Random Forest Algorithmus hierfür am besten geeignet ist [10]. In anderen Veröffentlichungen werden auch sogenannte *Learning to Rank* Ansätze untersucht, um bei Online Jobbörsen Vorschläge für Arbeitssuchende zu bewerten [11]. Es soll für einen Arbeitssuchenden ermittelt werden, in wie fern ein spezifisches Jobangebot zu diesem passt um so Arbeitsvermittler bei der Auswahl von Jobvorschlägen zu unterstützen. Außerdem soll mit dem XGBoost Algorithmus eine Verbesserung des Random Forest getestet und evaluiert werden.

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 [12]. Ein weiterer Datensatz, zu dem zum mir 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 spezifisches 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 Jobsuchenden 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 Labeln bezeichnet man als Regressionsproblem [13]. 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 Recherechephase 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 [14].
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 der Durchschnitt der einzelnen Entscheidungen der Entscheidungsbäume ermittelt [15].
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 [16].
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 [17]. 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 Informationen 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 Nachtteile 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 Urkaine 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 Tiwtter) users worldwide from 2019 to 2024.* [Online]. https://www.statista.com/statistics/303681/twitter-users-worldwide/. – Zugriff am: 2024-01-08.

[4] Miha Grcar, 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 Pandeimc on Twitter: Sentiment Analysis and Topic Modeling Study. In: *JMIR PUBLIC HEALT AND SURVEILLANCE* (2021).

[7] Weiwu Zhang, Thomas J. Johnson, Trent Seltzer, Shannon L. Bichard: The Revoultion will be Networked - The Influce 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] S. Mühlbuer, E. Weber: Machine Learning for Labour Market Matching. In: *Articles on labour market issues* (2022), Nr. 3, S. 5.

[11] Marios Kokkodis, Panagiotis Papadimitriou: Hiring Behavior Models for Online Labor Markets. (2015).

[12] 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-iebs. – Zugriff am: 2023-05-25.

[13] M. Mohri, A. Rostamizadeh, A. Talwalkar: *Foundations of Machine Learning*. London, England : MIT Press, 2018.

[14] A. Mammone, M. Turchi, N. Christianini: Support vector machines. In: *Journal of statistical software* 1 (2009).

[15] Leo Breiman: Random Forests. In: *Machine Learning* 45 (2001), S. 5–35.

[16] T. Chen, C. Guestrin: XGBoost: A Scalable Tree Boosting System. (2016).

[17] A. Botchkarev: Performance Metrics in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology. (2018), S. 4.