Fakultät Informatik

Topic Modelling on Large Multilingual Twitter Data Set on Ukraine War

Expose Bachelorarbeit im Studiengang Informatik

vorgelegt von

Robin Feldmann

Matrikelnummer 3538270

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 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]. 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 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]. Sazzed 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]. Sufi uses Latent Dirichlet Allocation on a dataset of 40,000 tweets in 54 languages, pre-processing the data with translations to ensure a cohesive analytical approach [14]. In her Bachelor thesis, Becker applies Latent Dirichlet Allocation to a Kaggle dataset identical to mine, comprising approximately 70 million tweets, employing a focused approach by segmenting the data into small time intervals and language-specific subsets [15]. To the best of my knowledge, there have been no studies yet that utilize multilingual topic modeling techniques on extensive multilingual social media datasets to analyze the Ukraine-Russian war.

Technology

This section briefly outlines the essential algorithms and technology employed in the thesis, with a detailed exposition reserved for the main body. Latent Dirichlet Allocation (LDA) [16] has emerged as a leading probabilistic topic modeling approach in recent years, particularly noted for its efficacy in monolingual text corpora. Due to its

reliance on word embeddings, Latent Dirichlet Allocation (LDA) is unsuitable for direct application on multilingual text corpora. In this thesis, LDA will be utilized solely for comparison purposes on translated texts.

The primary topic modeling approach examined in this thesis is based on the BertTopic Framework, a versatile and modular framework consisting of four distinct, interchangeable steps. The initial step involves a Text Embedding model, which embeds the text into a high-dimensional vector space. This is followed by a dimensionality reduction algorithm that simplifies these vector representations. The third step clusters the reduced vectors to form distinct topics. Finally, the fourth step involves generating representations of these topics, characterizing each topic within the dataset [17]. For the embedding step the thesis will utilize and evaluate two models for the text embedding algorithm, both grounded in Sentence-Bert (SBert) Sentence Embeddings using Siamese BERT-Networks [18]. The first model is based on the RoBERTa improvement, known for its enhanced training techniques and optimized performance [19]. The second model employs Language-agnostic BERT Sentence Embedding (LABSE) [20]. Both of these models are freely accessible on the Hugging Face website [21]. In the BertTopic Framework, dimensionality reduction is defaultly achieved using the UMAP algorithm, and clustering is conducted through the HDBSCAN algorithm. While exploring alternative algorithms would be intriguing, it falls outside the scope of this thesis, thus these default methods will be retained [22, 23].

Data Set

The dataset utilized in this study is publicly accessible on Kaggle, a widely recognized platform for data sharing and analysis. ”Kaggle is an online community for data science and machine learning (ML) enthusiasts. It is a top learning tool for novices and pros, with realistic practice problems to sharpen your data science skills. Owned by Google, it is currently the world’s largest crowdsourced web platform for data scientists and ML practitioners. Thus, Kaggle gives you access to several professionals in your field that you can brainstorm, compete, and solve real-life problems with” [24]. The dataset, compiled by the user BwandoWando, for whom limited information is available [25], was specifically gathered to monitor the ongoing Ukraine-Russia conflict through filtering for Twitter hashtags. It encompasses a total of around 70 million tweets, spanning from the first date of February 24, 2022, to the last date of June 14, 2023, covering 476

unique dates. The dataset spans 66 languages, with the most prevalent being English, German, French, Italian, Spanish, Ukrainian, and Russian, collectively accounting for approximately 86% of all tweets. A comprehensive explorative analysis of this dataset will be a part of the thesis, delving into detailed aspects and patterns within the data.

Preprocessing

In the preprocessing section of the thesis, the dataset will undergo a comprehensive exploration, cleaning, and preprocessing process. This will be conducted in accordance with the methods outlined in "Blueprints for Text Analytics Using Python". Additionally, the principles and techniques from Professor Albrecht's lecture on Natural Language Processing will be integrated into these processes [26]. To address duplicates in the dataset, particularly retweets and similar entries within specific timespans, a Min-Hash Algorithm will be employed for noisy deduplication [27]. For the purpose of testing, a sample consisting of 100 tweets per language, per week, from non-English languages will be translated into English. This translation will be accomplished using the free test version of the Google Cloud Translate API. This systematic approach ensures a sample for each language and time period for testing later in the research process [28].

Methodology

The topic modeling component of this research will be conducted in three distinct steps, each focusing on a specific subset of the data. For the initial case study, I will manually select 100 tweets, distributed evenly across four self-chosen topics, amounting to 25 tweets per topic. This curated dataset will serve as the basis for comparing results from various topic modeling approaches. The first method involves applying Latent Dirichlet Allocation (LDA) [16] to the English translations of these tweets. Additionally, the BertTopic [17] algorithm will be employed, incorporating different embedding models to analyze the data. The evaluation of this case study will be methodically conducted in three stages to ensure a thorough analysis of the topic modeling outcomes. The first

stage involves a subjective assessment, where the relevance and appeal of the identified topics will be evaluated from a human perspective. In the second stage, the results of the topic modeling will be compared against the pre-selected topics. This comparison aims to measure the alignment between the machine-generated topics and the manually chosen ones. The third and final stage of evaluation employs established metrics for topic models: Coherence, measured using Normalized Pointwise Mutual Information (NPMI) [29] and Diversity, determined by calculating the percentage of unique words per topic [30]. These metrics provide an objective and quantitative analysis of the model's performance, ensuring a comprehensive evaluation of its effectiveness in topic identification and differentiation.

In the subsequent case study, the scope of analysis will significantly expand to encompass all translated tweets, amounting to approximately 49,000 tweets. On those both Latent Dirichlet Allocation (LDA) and BertTopic will be utilized to cluster the data. The topics identified through those will be evaluated using subjective assessment, coherence, and diversity metrics. Additionally, a direct comparison between the topics from each method will be conducted to assess their respective clustering performances on the large dataset.

The outcomes of these two case studies will be instrumental in guiding key decisions and optimizing hyperparameters for the project. Insights gained from the results will inform the best practices for data cleaning and preprocessing, as well as the selection of suitable sentence embedding models and other hyperparameters. Armed with this refined understanding, the final approach will exclusively employ the BertTopic algorithm, applying it to the entire dataset. The topics derived from this will be evaluated based on subjective assessments, coherence, and diversity.

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] Clara Maathuis, Iddo Kerkhof: The first two months in the war in Ukraine through topioc 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 Internation Conference on Big Data* (2022).

[13] Pradham Nayak, Lakshmi JVN, Vanda V. Bhagat: Topic Modelling of ongoing conflict between Russia and Ukraine. In: *International Conference on Trends in Quantum Computing and Emerging Business Technologies* (2022).

[14] Fahim Sufi: Social Media Analytics on Russia-Ukraine Cyber War with Natural Language Processing: Perspectives and Challenges. In: *MDPI* (2023).

[15] Ivanna Becker: Analyse der öffentlichen Meinung während des russisch-ukrainischen Krieges: Eine Twitter-Daten-basierte Studie. (2023).

[16] David M. Blei, Andrew Y. Ng, Michael I. Jordan: Latent Dirichlet Allocation. In: *Journal of Machine Learning Research* (2003).

[17] Maarten Grootendorst: BERTopic: Neural topic modeling with a class-based TF-IDF procedure. (2022).

[18] Nils Reimers, Iryna Gurevych: Sentece-BERT: Sentece Embeddings using Siamese BERT-Networks. (2019).

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov: RoBERTa: A Robustyle Optimized BERT Pretraining Approach. (2019).

[20] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, Wei Wang: Language-agnostic BERT Sentence Embedding. (2022).

[21] ; (Hrsg.): *HuggingFace*. [Online]. https://huggingface.co/sentence-transformers/LaBSE. – Zugriff am: 2024-01-17.

[22] Leland McInnes, John Healy, James Melville: UMAP: Uniform Manifold Approxiamtion and Projection for Dimension Reduction. (2020).

[23] Ricardo J.G.B Campello, Davoud Moulavi, Joerg Sander: Densitiy-Based Clustering Based on Hierarchical Density Estimates. (2013).

[24] Joshua Adegoke: *A Beginner's Guide to Kaggle for Data Science.* [Online]. https://www.makeuseof.com/beginners-guide-to-kaggle/. Version: 2023. – Zugriff am: 2023-06-10.

[25] BwandoWando: *Bwando Wando.* [Online]. https://www.kaggle.com/bwandowando. Version: 2023. – Zugriff am: 2023-06-10.

[26] Jens Albrecht, Sidharth Ramachandran, Christian Winkler: Blueprints for Text Analytics Using Python. (2020).

[27] Emily Silcock, Luca D'Amico-Wong, Jinglin Yang, Melissa Dell: Noise-Robust De-Duplication at Scale. (2022).

[28] Google: *Cloud Translate API.* [Online]. https://cloud.google.com/translate/docs/reference/rest. Version: 2024. – Zugriff am: 2024-01-15.

[29] Gerlof Bouma: Normalized (Pointwise) Mutual Information in Collocation Extraction. (2009).

[30] Adji B. Dieng, Francisco J.R. Ruiz, David M. Blei: Topic Modeling in Embedding Spaces. In: *Transactions of the Association for Computational Linguistics* (2022).