Multilingual Retrieval of Crisis-Relevant Information from Social Media

Fedor Vitiugin

TESI DOCTORAL UPF / year 2023

THESIS SUPERVISOR
Dr. Carlos Castillo
Department of Information and Communication Technologies

Dedicated to the brave people providing relief and support during emergencies

Acknowledgments

Special thanks to my advisor, Carlor Castillo, for their guidance, support, and patience. They introduced me to such an important topic and gave me so much knowledge, experience, and wisdom during the past 5 years. I'm very appreciated of lessons about keeping a work-life balance, which made my pre-doctoral path much less stressful.

I would like to thank Hemant Purohit for his patience and support during our collaborations and for being an incredible host during my research stay at George Mason University. Furthermore, I want to thank colleagues from the Humanitarian Informatics Lab, where I had a privilege to learn more about the topic of my research.

Thanks to ISI Foundation and specially to Yelena Mejova for the opportunity to make research on the personally significant topic and for her contributions to this thesis.

I'm grateful to DTIC Department at UPF and all my colleagues. In particular, I want to thank Manuel Portela, who's become my older brother-in-science, for guiding and answering to all my questions about academic life in Spain. Moreover, thanks to both "Queens and the Boomers" and "GPTeam". It was a pleasure to share this journey with them.

An exceptional thanks to Lena, who unconditionally supported me since the first day, when my mind generates the first thought to start this journey. I will always be grateful for her patience, valuable advice and a bit of craziness.

Finally, thanks to my parents, Elena and Mikhail. The way I worked and how I ended this journey is mainly because of them. I thank them for showing me that I could do what I wanted through dedication and hard work.

This work has been partially supported by:

- the project “Capturing the «Big Picture» from Social Media to Enhance Flood Awareness at Global Scale (project number: Facebook PR10818)”;
- the Ministry of Science and Innovation of Spain with project “COMCRISIS”, reference code PID2019-109064GB-I00;
- the EU-funded “SoBigData++” project, under Grant Agreement 871042.

Abstract

Social media is a valuable platform for sharing real-time perspectives and insights, particularly during evolving events. Extracting relevant information from social media during emergencies can be challenging, especially when dealing with multiple languages. To address this issue, we have studied multilingual approaches that retrieves and summarizes crisis-related information from social media, providing a comprehensive solution. Our method involves experts and volunteers to address various information needs of emergency managers, allowing them to access distilled and aggregated data for validation and event monitoring. For crisis responders, targeted social media messages tailored to their operational requirements facilitate efficient work. Additionally, we propose a method for detecting and prioritizing critical help-seeking requests in multiple languages during disasters, enabling prompt and effective responses. By leveraging social media, our approach enhances emergency management by facilitating information assimilation, monitoring, and response coordination.

Resumen

Las redes sociales son una plataforma valiosa para compartir perspectivas e ideas en tiempo real, especialmente durante eventos en evolución. Extraer información relevante de las redes sociales durante emergencias puede ser un desafío, especialmente al lidiar con varios idiomas. Para abordar este problema, hemos desarrollado un enfoque multilingüe que recupera y resume información relacionada con crisis de las redes sociales, ofreciendo una solución integral. Nuestro método involucra a expertos y voluntarios para satisfacer las diversas necesidades de información de los administradores de emergencias, permitiéndoles acceder a datos destilados y agregados para la validación y el monitoreo de eventos. Para los respondientes a crisis, mensajes específicos de las redes sociales adaptados a sus requisitos operativos facilitan un trabajo eficiente. Además, proponemos un método para detectar y priorizar las solicitudes críticas de ayuda en múltiples idiomas durante desastres, lo que permite respuestas rápidas y efectivas. Al aprovechar las redes sociales, nuestro enfoque mejora la gestión de emergencias al facilitar la asimilación de información, el monitoreo y la coordinación de respuestas.

Contents

List of figures **XIII**

List of tables **XVI**

1 INTRODUCTION **1**
1.1 Goals 3
1.2 Challenges 5
1.3 Contributions 6
1.4 Additional Output 8

2 RELATED WORK **13**
2.1 Social Media during Disasters 13
2.2 Multilingual Natural Language Processing 16
2.3 Transfer Learning 17

3 ANALYSIS OF MULTILINGUAL NARRATIVES AROUND RUSSO-UKRAINIAN WAR EPISODES **21**
3.1 Introduction 22
3.2 Related Work 24
3.3 Methodology 26
3.3.1 Data collection and description 26
3.3.2 Network construction 28
3.3.3 Community detection 29
3.3.4 Location Extraction 30
3.3.5 Bot-or-not annotation 31

3.3.6 Content analysis 31
3.4 Results 32
3.4.1 Community structure 32
3.4.2 Community behavior 33
3.5 Narrative Analysis 38
3.5.1 Prominent Actors 38
3.5.2 Prominent Entities 44
3.5.3 Narrative Evolution 47
3.6 Discussion and Conclusions 48
3.6.1 Limitations and Privacy 50
3.6.2 Disclaimer of Positioning 51
3.6.3 Reproducibility 51

4 MULTILINGUAL INTERACTIVE ATTENTION NETWORK 53
4.1 Introduction 54
4.2 Related Work 57
4.2.1 Hate Speech Definitions 57
4.2.2 Hate Speech Detection Models 58
4.2.3 Human-Machine Collaboration for Hate Speech Detection 59
4.3 Methodology: Multilingual Interactive Attention Network (MLIAN) model 60
4.3.1 Interactive Attention Networks 61
4.3.2 Model Architecture 62
4.3.3 Transformer-Based Multilingual Embeddings . . 62
4.3.4 Human Feedback Guided by Frame Semantics Theory 64
4.4 Experiment setup 66
4.5 Result Analysis and Discussion 68
4.5.1 Multilingual Interactive Attention Network (MLIAN) Performance 68
4.5.2 Analysis of Frame Semantics Theory-based Human Feedback 69

4.5.3 Analysis of the Impact of Human Feedback . . . 70
4.5.4 Analysis of Cross-Lingual and Cross-Target Classification . . . 72
4.6 Conclusion . . . 74
4.6.1 Reproducibility . . . 75

5 CROSS-LINGUAL INFORMATION EXTRACTION AND SUMMARIZATION 77
5.1 Introduction . . . 78
5.2 Related Work . . . 79
5.2.1 Mining Social Media for During Crises . . . 79
5.2.2 Classification of Crisis-Related Messages . . . 80
5.2.3 Crisis-Related Information Summarization . . . 81
5.3 Method overview . . . 82
5.3.1 Classification Model . . . 82
5.3.2 Cross-lingual Ranking Model . . . 84
5.3.3 Summarization Model . . . 85
5.4 Experimental Setup and Evaluation . . . 86
5.4.1 Multilingual Data Collection . . . 87
5.4.2 Queries . . . 89
5.4.3 Message Classification Schemes . . . 90
5.4.4 Summarization Methods . . . 91
5.4.5 Evaluation Metrics . . . 92
5.5 Results . . . 93
5.5.1 Cross-lingual Classification . . . 93
5.5.2 Recall of Factual Claims . . . 95
5.5.3 BERTScore: Similarity with Official Reports . . . 95
5.5.4 Readability evaluation . . . 97
5.5.5 Expert Evaluation . . . 98
5.6 Conclusions and Limitations . . . 100
5.6.1 Reproducibility. . . . 100

6 MULTILINGUAL SERVICEABILITY MODEL FOR DETECTING AND RANKING HELP REQUESTS ON SOCIAL MEDIA DURING DISASTERS 101
6.1 Introduction 102
6.2 Related Work 104
6.2.1 Social Media Requests 104
6.2.2 Knowledge Distillation and Teacher-Student Model 105
6.3 Method 106
6.3.1 *MulTMR*: Multiple Teachers Model for Ranking 106
6.3.2 Behavioral Fine-Tuning of Pre-Trained Models . 108
6.4 Experiment Setup 109
6.4.1 Data 110
6.4.2 Schemes 112
6.4.3 Model Implementation 113
6.5 Result Analysis and Discussion 114
6.5.1 Multiple Teachers Model for Ranking (MulTMR) Performance 114
6.5.2 Impact of Behavior-Guided Models 115
6.5.3 Analysis of Behavior-Guided Modeling 117
6.5.4 Cross-Lingual Performance 118
6.5.5 Learning to Rank Serviceable Help Requests . . 119
6.6 Conclusions 120
6.6.1 Reproducibility 121

7 CONCLUSIONS AND FUTURE WORK 123
7.1 Limitations 125
7.2 Future directions 125
7.2.1 Multilingual Large Language Models 125
7.2.2 Crisis Information Extraction 128

List of Figures

1.1 Growth of internet penetration by regions 2

3.1 Number of tweets per day (all languages). 27
3.2 User communities for each language and event 34
3.3 Account creation date (year) 35
3.4 Jaccard similarity measuring the overlap of communities 36
3.5 Results of BotOrNot analysis over the events 37
3.6 Number of users posting in each event in Russian 38

4.1 Attention weights distribution 55
4.2 The overall architecture of MLIAN model. 63
4.3 High Level Architecture of Frame Extraction Process. . . 70
4.4 Attention weight maps of texts in English and Spanish . 71

5.1 Overview of our cross-lingual summarization framework 82
5.2 Combining the embeddings with features using deep MLP 86

6.1 The overall architecture of MulTMR. 107
6.2 Attention weight maps of the texts in 3 languages 117

List of Tables

3.1 Statistics of 10-day datasets 28
3.2 Statistics of GCC networks 30
3.3 Top hashtags by odds ratio in English 39
3.4 Top hashtags by odds ratio in Russian 40
3.5 Top hashtags by odds ratio in Ukrainian 41

4.1 Example of messages with hate speech 55
4.2 Train and test data. 66
4.3 Results of binary classification 68
4.4 Minimal human impact 72
4.5 Results of cross-lingual classification 73
4.6 Results of cross-target classification 74

5.1 Values of hyperparameters. 84
5.2 Number of annotated messages for each event 87
5.3 Categories for multilingual information extraction 89
5.4 Example query 90
5.5 Results of "leave-one-language-out" classification 94
5.6 Results of "leave-one-event-out" classification 94
5.7 Recall of factual claims 96
5.8 BERTScore of cross-lingual summaries 96
5.9 BERTScore of English-only summaries 97
5.10 Readability evaluation of summaries 98
5.11 Expert evaluation results 99

6.1 Examples of messages with serviceability characteristics 102

6.2 Error examples 110
6.3 Summary of datasets. 111
6.4 Results of binary classification 114
6.5 Behavior-specific models impact 116
6.6 Results of “leave-one-language-out” classification 118
6.7 Comparison of the average $nDCG$ 120

Acronyms

ACC Accuracy

ACL the Association for Computational Linguistics

AI Artificial Intelligence

AUC Area Under the Receiver Operating Characteristic Curve

BART Bidirectional and Auto-Regressive Transformers

BERT Bidirectional Encoder Representations from Transformers

BiLSTM Bidirectional Long Short-Term Memory

BPE Byte-Pair Encoding

CLiQC-CM Cross-LIngual Query-based Classification of Crisis Messages

CLiQS-CM Cross-LIngual Query-based Summarization of Crisis Messages

CLiQS-D-CM Diversified Cross-LIngual Query-based Summarization of Crisis Messages

CNN Convolutional Neural Networks

CTG Controllable Text Generation

DistilmBERT Distilled version of multilingual BERT

ERCC Emergency Response Coordination Centre

GBDT Gradient-Boosted Decision Trees

GCC Giant Connected Component

GLOVE Global Vectors for Word Representation

IAN Interactive Attention Network

LASER Language-Agnostic SEntence Representations

LLM Large Language Model

LSTM Long Short Term Memory

MLIAN Multilingual Interactive Attention Network

MLP Multi-Layer Perceptron

MulTMR Multiple Teachers Model for Ranking

nDCG normalized Discounted Cumulative Gain

NER Named Entity Recognition

NLG Natural Language Generation

NLP Natural Language Processing

PEGASUS Pre-training with Extracted Gap-sentences for Abstractive Summarization

POS Part Of Speech

RoBERTa Robustly Optimized BERT approach

SVM Support Vector Machines

T5 Text to Text Transfer Transformer

UN United Nations

VOST Virtual Operations Support Teams

Chapter 1

INTRODUCTION

In 2015, United Nations (UN) member states adopted a global blueprint to guide disaster risk reduction efforts over a 15-year period – The Sendai Framework for Disaster Risk Reduction 2015-2030 [36]. It aims to reduce the impact of disasters by focusing on preventing and mitigating risks, as well as enhancing resilience at all levels. Social media was recognized in the framework as one of the useful tools to help enhance early warning systems, preparedness, response coordination, and community engagement [85]. Their role lies in facilitating timely and widespread communication, two-way interaction, situational analysis, and fostering a resilient and connected community during disaster risk reduction and response efforts.

Social media is becoming ubiquitous and has become one of the main sources for users in numerous domains, including public health, economics, and politics. Discussions about emergencies and the damage caused by disasters [7] have become an object of research. This helps in analyzing the users' perspective, and in the case of disasters, among other things, it helps to understand the communication between the public and authorities tasked with emergency management [165].

This work started in 2020 when there were about 3.8 active social media users in the world [1]. At the time when this work was finished, in 2023, the number of active users have grown to 4.76 [2] billions. The

total growth in almost one billion of active users in 4 years. Based on both reports we could declare "the next billion users" [16] are widely spread in the whole world, but the largest growth ($\geq 10\ p.p.$) was observed by 5 out of 19 regions: Central Asia,Northern Africa, Western Africa, Eastern Asia, and Southern Africa. Figure 1.1 illustrates the distribution of difference in internet penetration all over the world.

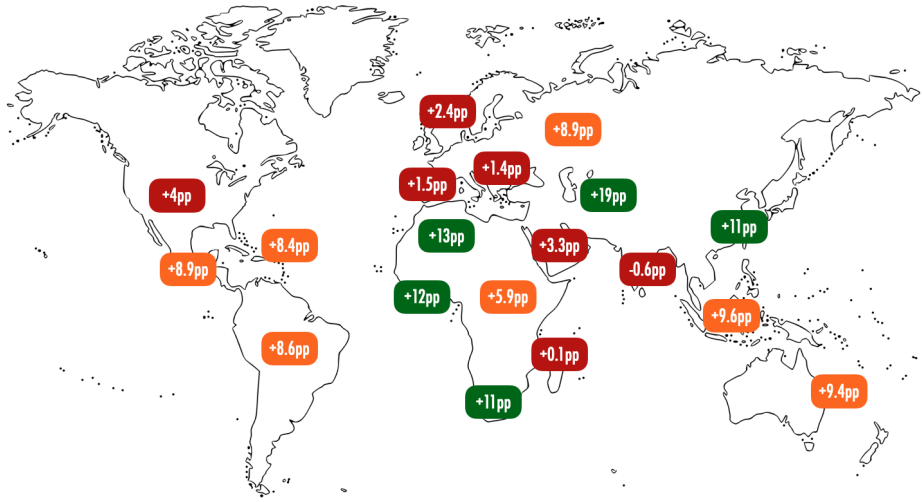


Figure 1.1: Growth of internet penetration by regions: difference in the share of active users against the total population in 2020 and 2023 years.

Despite the relentless growth of internet adoption in countries of the "Global South," numerous languages spoken by tens of millions of people in Africa, Asia, and the Americas are significantly underserved in terms of research and technology. This disparity is evident in the research community's emphasis on English. For instance, 63% of the papers published at the annual meeting of the Association for Computational Linguistics (ACL) 2008 were focused exclusively on English, and this portion increased to 70% at ACL 2021 [201]. This multilingualism problem becomes even more crucial when we consider the regions' high susceptibility to natural disasters [75]. These regions are disproportionately affected

by natural disasters, making it imperative to address the language gap in research and technology. By providing resources and support in local languages, we can bridge the communication barriers and enhance disaster management efforts in these vulnerable areas.

1.1 Goals

Social media is a relevant data source for gathering information about disasters, offering timely and invaluable insights into the spatial [194, 229] and temporal [209] development of a crisis. Additionally, it aids in the identification of key disaster-related sub-event, urgent needs, and actors already operating on the ground [124]. However, the existing research has predominantly focused on examining messages in a single language [203].

The primary objective of this study is to investigate the significance of collecting disaster-related information provided in diverse languages and propose effective strategies for processing this information.

The research work has been planned for answering the following research questions:

- *RQ1. What are the main actors and their respective roles in social media during crisis across various languages?*

  The response to this question aims to showcase the contrasting information provided in different languages, encompassing variations in actors, narratives, and tones. To address this question, in Chapter 3 we study three separate events in three different languages. Additionally, an analysis of accounts and their interconnections is presented alongside the content evaluation.

- *RQ2. What potential enhancements and improvements could bring a combination of context features and a human-in-the-loop approach in the classification of multilingual social media?*

  To address this question, in Chapter 4 we developed an experiment based on including minimal human guidance in the training

process of a hate speech classification model for achieving higher performance. Human feedback helps to detect hate subtleties and phrases for extracting features during model training, where the human feedback is guided by common element of the frames to express a hate speech, i.e., hate targets. The experimental dataset includes social media posts from two different topics in two languages.

- *RQ3. How does the use of a transfer learning approach contribute to enhancing the performance of detecting, classifying, and summarizing multilingual information from social media during disasters?*

  To answer this question, in Chapter 5, a flexible cross-lingual method, based on queries, was developed to collect relevant social media posts in multiple languages pertaining to specific information categories. These detected posts were then utilized as input for a transformer-based language model to generate crisis summaries. The empirical validation of this approach involved the assessment of a dataset comprising five events in a total of ten languages, conducted by crowdsourced annotators and emergency management experts.

- *RQ4. Could the implementation of behavior-guided models improve detecting and ranking of multilingual help-seeking requests on social media during disasters?*

  To address this question in Chaprer 6, the presented framework relies on the knowledge distillation process for designing a computational framework called Multiple Teachers Model for Ranking (MulTMR), which detects and prioritizes multilingual serviceable help requests on social media during disasters. This process aims to transfer knowledge from one or more complex models (teacher) to a simpler model (student) for a task, to train it to mimic the teacher models. It creatively leverages behavior-guided teacher models in the knowledge distillation process for achieving higher performance on a task. The approach was tested on data collected during ten

events in three languages.

1.2 Challenges

Despite the recognized value of social media analysis in providing timely data and analytical methods for studying disasters, we encountered several challenges during our research, which we endeavored to address at various stages of our work.

As previously mentioned, the majority of relevant research has been presented in a single language. As our research is data-driven, we encountered difficulties in finding data in multiple languages collected during the same event. Consequently, we prepared two multilingual datasets comprising data collected from eight events for our study. Furthermore, we included an additional dataset in this work, consisting of data collected from various events but within a single language. In total, we collect and release crisis data in 13 languages, all of which are freely available for research purposes.

In continuation of the preceding challenge, we confronted cross-cultural differences [71] in users' responses during disasters. These disparities posed a significant challenge in developing models capable of accurately processing such diverse data. To address this challenge, we employed two approaches: the use of behavior-related features and a human-in-the-loop approach. Both approaches enhanced the performance of our models in multilingual settings. However, it is important to note that the study of cross-cultural differences goes beyond the scope of this work.

Finally, in the real-world scenarios involving the utilization of disaster-related social media, the involvement of experts in the field is crucial [149]. In Chapter 4 and Chapter 5, we propose methods for integrating domain experts as assessors for few-shot learning tasks and as providers of domain knowledge. Such integration ensures a high level of expertise and enhances the practical applications of our research.

1.3 Contributions

In this thesis, we study multilingual social media during disasters and propose the approaches to process and mitigate these data.

Chapter 2 describes the state-of-the art on disaster-related social media, multilingual Natural Language Processing (NLP) and particularly applying of transfer learning for NLP tasks.

In **Chapter 3** we delve into the analysis of user discussions related to three episodes of the Russian-Ukraine war in English, Russian, and Ukrainian languages. We find a substantial showing of pro-Russia communities in English and Russian language, despite Twitter suspending pro-Russia accounts at a higher rate than others. On the other hand, we find a variety of groups supporting Ukraine in all three languages, including the Russian opposition, Eastern-European and international media, and numerous politicians. Notably, the pro-West/Ukraine and pro-Russia communities vilify different actors and build their narrative around these different centers. The multilingual approach adopted in this study allows us to capture the diverse actors and their roles in international events, revealing the complexity and nuances that would have otherwise been missed. This chapter is also a paper:

Fedor Vitiugin, Yelena Mejova [Under review]. Multilingual Battle of Narratives around Russo-Ukrainian War Episodes.

In **Chapter 4** we research online hate speech because it was become a critical issue on social media platforms, exacerbated during conflicts. Recent research has primarily focused on detecting hate speech in one language, but hate speech transcends language and geographical boundaries in the interconnected world. This necessitates further investigation into multilingual hate speech detection methods, with a strong emphasis on model interpretability for understanding contextual errors.

In the study, we propose the Multilingual Interactive Attention Net-

work (MLIAN) model for detecting hate speech in multilingual social media text. Our model utilizes attention networks for interpretability and incorporates a human-in-the-loop paradigm for adaptability. By dynamically learning attention towards relevant words and leveraging labels provided by simulated human feedback, our model achieves superior performance. We evaluated the model on datasets in English and Spanish, and our results reveal that involvement of humans in the pipeline enhances model performance by establishing meaningful connections between attention weights and semantic frames across languages. This chapter was published as:

Fedor Vitiugin, Yasas Senarath, Hemant Purohit (2021). Efficient Detection of Multilingual Hate Speech by Using Interactive Attention Network with Minimal Human Feedback. 13th ACM Web Science Conference, Virtual Event.

Chapter 5 introduces a cross-lingual method that retrieves and summarizes crisis-related information from social media posts. Our approach utilizes structured queries are made by human experts to express various information needs in a uniform manner, allowing for the retrieval of relevant sentences in different languages. To generate abstractive summaries, we employ multilingual transformer embeddings.

The evaluation of our method involved crowdsourcing evaluators and emergency management experts, who assessed collections of Twitter data from five large-scale disasters in ten languages. The results demonstrate the flexibility of our approach. The generated summaries are considered more focused, structured, and coherent compared to existing state-of-the-art methods. Experts also favorably compare our summaries to those created using current techniques. This research was presented in the following paper:

Fedor Vitiugin, Carlos Castillo (2022). Cross-Lingual Query-Based Summarization of Crisis-Related Social Media: An Abstractive Approach Using Transformers. HyperText'22: Proceedings of the 33rd ACM Con-

ference on Hypertext and Social Media, Barcelona, Spain. **The paper was awarded as "Ted Nelson Newcomer Best Paper".**

Chapter 6 presents MulTMR, a knowledge distillation framework that combines the strengths of task-related and behavior-guided models as diverse teachers. This framework trains a student model to efficiently detect serviceable request posts across languages and regions on social media during natural disasters. Our method demonstrates improved performance, across various multilingual test scenarios. We validate our results using real-world data collected in three languages from ten disasters across seven countries. The inclusion of behavior-guided teacher models in MulTMR enhances attention to relevant indicators of serviceability characteristics.

The adoption of the MulTMR framework could alleviate the cognitive load on emergency services personnel during disaster events. Additionally, it has the potential for deployment in different languages and regions worldwide, enhancing its applicability on a global scale. This research is also available in the following paper:

Fedor Vitiugin, Hemant Purohit [Under review]. Multilingual Serviceability Model for Detecting and Ranking Help Requests on Social Media during Disasters.

Finally, we draw conclusions and discuss future directions of research in **Chapter 7**.

1.4 Additional Output

As part of my studies, I had the chance to collaborate on additional papers that are not included in this manuscript.

An analysis of social media discourse on COVID-19 pandemics by governments and public health agencies across multiple countries.

During the COVID-19 pandemic, information is being rapidly shared by public health experts and researchers through social media platforms. Whilst government policies were disseminated and discussed, fake news and misinformation simultaneously created a corresponding wave of "infodemics." This study analyzed the discourse on Twitter in several languages, investigating the reactions to government and public health agency social media accounts that share policy decisions and official messages. The study collected messages from 21 official Twitter accounts of governments and public health authorities in the UK, US, Mexico, Canada, Brazil, Spain, and Nigeria, from 15 March to 29 May 2020. Over 2 million tweets in various languages were analyzed using a mixed-methods approach to understand the messages both quantitatively and qualitatively. Using automatic, text-based clustering, five topics were identified for each account and then categorized into 10 emerging themes. Identified themes include political, socio-economic, and population-protection issues, encompassing global, national, and individual levels. A comparison was performed amongst the seven countries analyzed and the United Kingdom (Scotland, Northern Ireland, and England) to find similarities and differences between countries and government agencies. Despite the difference in language, country of origin, epidemiological contexts within the countries, significant similarities emerged. Our results suggest that other than general announcement and reportage messages, the most-discussed topic is evidence-based leadership and policymaking, followed by how to manage socio-economic consequences.

During work on this project, I was responsible for data collection and processing, as well as consulting of colleagues during analysis.

Li, L., Aldosery, A., Vitiugin, F., Nathan, N., Novillo-Ortiz, D., Castillo, C., Kostkova, P. (2021). The response of governments and public health agencies to COVID-19 pandemics on social media: a multi-country analysis of twitter discourse. Frontiers in Public Health, 9, 716333.

Emotion Detection for Spanish by Combining LASER Embed-

dings, Topic Information, and Offense Features. The paper describes the system submitted by the WSSC Team to the EmoEvalEs@IberLEF 2021 emotions detection competition. We propose a novel model for Emotion Detection that combines transformer embeddings with topic information and offense features. The system classifies social media text emotions leveraging its context representations. Our results show that, for this kind of task, our model outperforms baselines and state-of-the-art text classification methods.

Vitiugin, F., Barnabo, G. (2021). Emotion Detection for Spanish by Combining LASER Embeddings, Topic Information, and Offense Features.

Use of LLMs for effective summarization of DBpedia abstracts. This study addresses the limitations of existing short abstracts of DBpedia entities, which often lack a comprehensive overview due to their creating method (i.e., selecting the first two-three sentences from the full DBpedia abstracts). We leverage pre-trained language models to generate abstractive summaries of DBpedia abstracts in six languages (English, French, German, Italian, Spanish, and Dutch). In particular, we performed several experiments to compare the quality of generated abstracts by different language models. We evaluated the generated summaries using human judgments and automated metrics such as Self-ROUGE and BERTScore. Finally, we studied the correlation between human and automated metrics used for evaluating the generated summaries under different facets: informativeness, coherence, conciseness, and fluency.

Pre-trained language models outperform the existing short abstracts of DBpedia abstracts, especially for long ones, by producing more informative and concise summaries. BART-based models effectively overcome the current limitations of DBpedia short abstracts. Further, we find that BERTScore and ROUGE-1 are reliable measures for evaluating the informativeness and coherence of generated summaries with respect to full DBpedia abstracts. We also observe a negative correlation between

conciseness and human ratings. Additionally, the evaluation of fluency remains challenging without human assessment.

During work on this research, I was responsible for experiment design and evaluation, as well as result analysis.

Zahera, H., Vitiugin, F., Sherif, M., Castillo, C., Ngomo, A. C. N. (2023). Using Pre-trained Language Models for Abstractive DBpedia Summarization: A Comparative Study. In Proc. SEMANTiCS 2023.

Chapter 2

RELATED WORK

While performing the research for this thesis, we collected relevant studies on social media during disasters, multilingual natural language processing, and transfer learning. Our work primarily focuses on language-related research, supplemented by the utilization of network science and ranking methods.

2.1 Social Media during Disasters

Social media plays a crucial role in emergency response and enhancing situational awareness during and after natural disasters [106, 114, 225]. However, there are several challenges involved in gathering and extracting relevant information from social media platforms. These challenges include dealing with the large volume of data, unstructured data sources, the signal-to-noise ratio, ungrammatical and multilingual content, and the identification and removal of fraudulent messages [214, 176]. Due to the vast amount and diversity of data generated on social media, the information extracted varies in quality and usefulness. For instance, a tweet with geographical information (geo-tagging) attached, indicating a roadblock on a hilly road, provides more valuable contextual details compared to a similar tweet lacking geo-tagging. Likewise, a tweet accompanied by images has the potential to enhance situational awareness significantly.

For example, if someone shares photos of a roadblock on a hilly road, it can help drivers in the vicinity understand the current situation and opt for an alternate route to bypass the blocked area [181]. In this section, our focus is on examining the methods used for mining, classifying, and summarizing social media messages relevant to crises.

During various types of crises, including both natural and man-made disasters, social media serves as a crucial communication channel. The integration of computational methods from diverse disciplines facilitates the development of mining and retrieval systems, offering valuable assistance to emergency managers [35]. Crisis-related social media comprehends a wide range of information categories, including timely messages conveying urgent needs of affected populations, as well as updates on the condition of damaged infrastructure such as bridges or roads. Together, this information plays a significant role in emergency response, recovery management, and the assessment of the costs of damages [117].

For classification of crisis-relevant social media, a range of approaches have been proposed, encompassing both "traditional" supervised learning methods like Naive Bayes and Support Vector Machines (SVM), as well as neural-network-based methods [214]. SVM, in particular, has consistently demonstrated high performance when combined with semantic features derived from external knowledge bases [122], while deep learning methods utilizing architectures like Convolutional Neural Networks (CNN) with word embeddings have also proven effective in detecting crisis-relevant messages [172, 147]. Moreover, the inclusion of event-specific information, such as hydrological data for floods, has been shown to further enhance classification performance [53].

Crisis-related social media messages often consist of pieced content, resulting in fragmented information. Therefore, the consolidation and summarization of this information are crucial [204, 205]. A well-crafted summary plays an essential role in providing stakeholders with situational awareness and enabling effective resource management [247].

Two primary approaches are employed for text summarization:

- *Extractive* methods involve selecting informative phrases or even complete sentences from the source text to construct the summary

[65].

- *Abstractive* summarization techniques, in contrast, generate summaries by capturing the underlying semantics of the given text. Consequently, abstractive summaries may include words or sentences that do not explicitly appear in the source document(s).

Abstractive summarization often employs a generative approach, which has gained significant popularity in recent years, particularly with the advancement of pre-trained large language models [142, 134].

Despite the advantages of abstractive approaches, extractive methods are still considered the state-of-the-art for summarization, owing to their simplicity and high performance levels [196, 113]. However, extractive approaches often fall short in including crucial elements required for comprehensive reports, such as answers to "what," "who," "where," "when," and "how" questions. These elements hold significant importance in the domain of disaster and crisis management and need to be succinctly incorporated into summaries [130]. Query-based approaches have been proposed as an effective means of integrating this information to enhance the quality of reports [197]. In general, abstractive methods have the potential to generate more informative summaries that are not limited to sentences directly extracted from the source text, thereby expanding the scope of information synthesis [171].

Furthermore, social media has emerged as a popular platform for individuals to seek help from emergency services during natural disasters [174, 247, 41, 60]. Whether it is for rescue operations, essential supplies, or critical information, social media often serves as the primary point of contact for those in need. In contrast to other online contexts, posts made during disasters need immediate attention and should be directed towards the appropriate recipients, such as rescue teams, to ensure prompt offline responses. Consequently, specialized strategies have been devised to prioritize and address actionable posts requesting help [221, 187, 108].

In the context of disaster management, the gathering of relevant and timely information from social media platforms is essential for effective response and mitigation efforts [188]. However, the challenge intensifies

when dealing with posts written in different languages, adding complexity to the information extraction process [238, 149].

Our work focuses on addressing the complexity introduced by language variety in social media posts during the information extraction process.

2.2 Multilingual Natural Language Processing

Presently, various platforms, including social media, offer people from diverse backgrounds and languages the opportunity to connect and share information. It has become common to come across comments in different languages on posts made by international celebrities or data providers. Consequently, in the field of NLP, there is a growing interest in comprehending cross-lingual content and addressing the challenges of multilingualism. The ability to handle multilingual content in NLP has emerged as a significant research area in this era. Many endeavors have been made to harness existing NLP technologies and tackle the complexities associated with understanding and processing content in multiple languages.

The majority of existing methods for mining social media during disasters, as described in the extensive literature, are predominantly monolingual in nature [248, 181, 149]. This limitation restricts their applicability in countries where languages other than English are spoken, including English-speaking countries with increasingly diverse multilingual urban populations [149]. Notably, a recent analysis revealed that media coverage of crises varies across different countries, highlighting the importance of considering multilingual perspectives [96]. Therefore, the ability to leverage social data mining methods on user-generated content across multiple languages holds the potential to significantly enhance disaster response efforts [238]. Cross-lingual and multilingual classification and summarization methods offer an opportunity to gather complementary information from various languages spoken in affected regions.

Research into multilingual social media analysis is still in its nascent stages. While machine translation is a straightforward yet less effective

approach for processing social media data [217], recent advancements have demonstrated promising results in multilingual hate speech detection using SVM and Bidirectional Long Short-Term Memory (BiLSTM) models on three datasets comprising English, Italian, and German languages [47]. Additionally, large-scale analyses of deep learning models have been conducted to develop classifiers for multilingual classification tasks, encompassing 16 datasets from nine languages. The findings indicate that for low-resource languages, Language-Agnostic SEntence Representations (LASER) embedding combined with logistic regression achieves superior performance, while Bidirectional Encoder Representations from Transformers (BERT) based models outperform others in high-resource settings [14].

Some existing approaches propose the utilization of external knowledge sources, such as specific lexicons, to enhance detection systems by leveraging multilingual, fine-grained resources [56]. Although this approach can be effective, it necessitates labor-intensive efforts to develop and maintain these knowledge sources, especially in a multilingual context, where maintaining up-to-date lexicons can be challenging.

To address these challenges, recent studies have introduced innovative approaches for retrieving and classifying information from multilingual social media. These methods harness the power of deep learning models with multilingual embeddings [147, 119] and pre-trained Large Language Models (LLMs). The latter approach involves fine-tuning LLMs for specific tasks and domains, resulting in enhanced performance in crisis-related analysis [145, 211, 72].

Our work effectively leveraged multilingual embeddings and large language models to tackle multilingual challenges in crisis-related tasks.

2.3 Transfer Learning

The field of NLP has witnessed remarkable progress recently, characterized by the emergence of influential models like ELMo [105], BERT [59], and GPT-3 [30], attracting significant attention and recognition in the re-

search community. This advancement in NLP has been primarily driven by the development of transfer learning techniques, enabling the effective transfer of knowledge acquired in one context to different scenarios, languages, or tasks [18]. Transfer learning is a machine learning approach that utilizes knowledge gained from pre-training a model on general tasks to enhance efficiency and expedite fine-tuning in related tasks [202]. This technique was first introduced with the successful large CNN model, ImageNet, in 2010 [58]. ImageNet involved fine-tuning deep neural networks using over 14 million images across more than 20,000 categories. In the realm of NLP applications, transfer learning has been extensively employed, leading to state-of-the-art results in various studies, including sentiment analysis [146], hate speech detection [166], question answering [164], etc.

The utilization of transformer-based models, known for their superior performance and robustness, has become prevalent in crisis classification tasks, surpassing conventional linear and deep learning models [145, 50, 173]. Furthermore, the availability of pretrained language models, coupled with well-established procedures for fine-tuning them to specific tasks or domains, has significantly alleviated the computational and training time burdens, making these techniques more accessible to practitioners with limited resources.

In the context of disaster-related social media analysis, various approaches have been proposed to leverage the capabilities of LLMs. These include employing Teacher-Student training methods for cross-lingual transfer scenarios [129], integrating graph neural networks and transformer-based LLMs through cross-attention mechanisms [86], and utilizing multimodal neural networks for combined textual and visual data analysis [127]. These techniques enable more effective and comprehensive analysis of disaster-related social media content.

The Teacher-Student model, a knowledge distillation approach [98], has proven valuable in transferring knowledge from complex models to simpler ones, and it has been applied in various tasks and domains [216, 100, 39]. Recently, a multilingual knowledge distillation approach has been proposed, leveraging the Teacher-Student framework to transfer knowl-

edge from high-performance monolingual models to a multilingual model, resulting in improved performance [246]. Moreover, task-specific models can also serve as teacher models, enabling the transfer of specific knowledge to the student model [125].

In our work, we employed the state-of-the-art in transfer learning and knowledge distillation techniques to enhance the efficiency and effectiveness of large language models in crisis-related tasks.

Chapter 3

ANALYSIS OF MULTILINGUAL NARRATIVES AROUND RUSSO-UKRAINIAN WAR EPISODES

During crisis events, the information shared in social media across different languages can vary significantly. In this chapter, we aim to examine the disparities in information provided during the same crisis events across English, Russian, and Ukrainian languages. By analyzing a vast collection of approximately 58 million tweets, we investigate the stances and affiliations of users discussing three war-related episodes. Our focus is on uncovering major communities in the conflict through community detection in retweet networks.

Our findings reveal that language-related communities not only support and vilify different actors and narrative, but also contribute to the diversity of actors and narratives themselves. This highlights the significance of understanding the role of language in shaping public opinion during crisis events. Additionally, we discuss the importance of narra-

tives in maintaining public support and the role of social media platforms as influential arbiters in this process. This chapter sets the stage for the subsequent chapters, providing background and context for our research.

3.1 Introduction

On February 24, 2022, the 8-year ongoing conflict between Russia and Ukraine has commenced a major escalation, with Russia sending troops across the North and East borders of Ukraine [38]. As military situation developed on the ground, a similar escalation took place in the war to control the narrative around the invasion in the mainstream and social media [44]. This control is especially important, given that the billions of euros in support sent to Ukraine by the Western countries are often authorized by elected legislative bodies [8]. Indeed, the willingness to support Ukraine ranges widely across the West: in May, 2022 survey, about half of respondents from Italy and Germany opted to end the war, even if the lands captured by Russia are not returned to Ukraine, and only 16% of those from Poland choose this route [128]. However, a difference in opinion on the war also exists in Russia, as far as it can be ascertained via surveys [219] and public anti-war protests [92], although the country's block of Facebook and Twitter shortly after the invasion made it necessary to use VPN services for the dissenting voices to be heard in the Western-controlled social media [163].

Twitter has been ascribed an important role in Ukraine in the past decade. Unlike the original Orange Revolution protests in 2004, which were largely televised, the 2014 Euromaidan protests were also transmitted on fledgling social media websites [161]. Meanwhile, Russia stepped up its efforts to stir social media narratives in its favor, including using Russian Internet Research Agency (IRA) to run propaganda and influence operations online [63]. In response to increased activity and attention, in March 2022 Twitter announced the creation of special curation efforts around war-related content, including flagging links to Russian state-affiliated media websites [160]. At the same time, conflict-related

rising energy prices and post-pandemic supply chain disruptions have precipitated increased inflation in daily cost of living and potential recessions in the West, challenging the will of these governments to continue their support of Ukraine [208]. Thus, it is crucial to understand the battle over establishing a dominant narrative around the military events in Ukraine, as it may lead to further developments in the conflict.

In this work, we examine the Twitter discussions around three events: (i) Mariupol theatre airstrike[1] (17/03/2022), (ii) Bucha massacre[2] discovery (03/04/2022), and (iii) Kremenchuk shopping mall attack.[3] (27/06/2022) The first and third events are bombing incidents of civilian targets, while second is a discovery by Ukrainian forces of possible killings of civilians after the Russian forces withdrew from the city of Bucha. Spanning 3 months of the conflict, these events figured prominently in Western media, and produced a response in the Russian media [227, 228, 22]. In particular, in this study we focus on the discussion around these events in Ukrainian, Russian, and English languages. As of 2001, 67% of Ukrainians spoke Ukrainian, and 30% Russian [232], while Education First gave Ukraine "moderate" English proficiency score in 2022 [66]. Thus, we aim to capture both the multitude of voices within Ukraine, but also from Russia, and the West. Using this data, we aim to address the following research questions:

- **RQ 1.1**: How do communities, detected based on user retweets, self-identify in terms of support (pro-Russia vs. pro-West/Ukraine) and state/media affiliation?

- **RQ 1.2**: What are the properties of the detected communities, in terms of (i) account creation date, (ii) continuity across events, (iii) use of bots and subsequent suspension by Twitter?

- **RQ 1.3**: What narratives in terms of main actors and their roles are built by the above communities across these events?

[1] https://en.wikipedia.org/wiki/Mariupol_theatre_airstrike
[2] https://en.wikipedia.org/wiki/Bucha_massacre
[3] https://en.wikipedia.org/wiki/Kremenchuk_shopping_mall_attack

This study illustrates the importance of using multilingual approach to analyzing international conflict events by capturing the variety of actors and angles engaged in shaping of the discussion around it.

3.2 Related Work

Since the 2022 invasion, as well as in the time since the 2014 annexation of Crimea [239], social media studies have revealed a plethora of actors (some automated) competing to establish a specific narrative vilifying or favoring one or the other side. Botometer has been used on tweets with affiliation hashtags such as #IStandWithPutin or #IStandWithUkraine by Smart et al. [218], who found that, although bot-like activity was detected on both sides of ideological divide, human-controlled accounts, or accounts which appear less bot-like, have more influence in the captured social network. They speculate that this could be potentially due to "their behaviour or perception". Focusing on Ukraine-related disinformation and its debunking, Singh et al. [217] find that tweets spreading disinformation were shared and retweeted significantly more, compared to those containing debunks. The authors recommend that machine translation is used to tackle the multilingual nature of the data, and to match debunking content with the original disinformation. An earlier study on the accounts from the IRA found that such accounts "sought to keep the audience distracted" with a diverse set of conspiracy theories, using the principle "if nothing is true, then anything is possible" [63]. The spread of information and misinformation is also likely to be affected by the partisan and ethnolinguistic affinities of social media users. A 2019 survey study of Ukrainians found that those with closer partisan ties to Russia were more likely to believe pro-Kremlin disinformation across different topics [67]. These topics, and the overarching narratives, are further likely to be different depending on the writer's country of origin. A recent analysis of the war coverage by Chinese, Russian, and Western press revealed that "Western press outlets have focused on the military and humanitarian aspects of the war, Russian media have focused on the purported justifica-

tions for the "special military operation" such as the presence in Ukraine of "bio-weapons" and "neo-nazis", and Chinese news media have concentrated on the conflict's diplomatic and economic consequences" [96]. A heterogeneity also existed (at least at the start of the war) among the Russian media sources, as exemplified by an analysis of a dataset that combined Twitter and VKontakte posts from a variety of Russian media outlets [180]. Whereas independent outlets discussed the legality of the invasion, state-affiliated ones focused on capability and policy. In fact, even those on the West opposing the war may have a range of opinions on the appropriate response to the conflict. A survey by the European Council on Foreign Relations conducted in mid-May 2022 showed that residents of different European countries disagreed on the long-term goals: a "Peace" camp wanted the war to end as soon as possible, and a "Justice" camp believed the more pressing goal is to punish Russia [128]. The survey found that in all countries, apart from Poland, the "Peace" camp was larger than the "Justice" camp. However, such surveys provide a snapshot of public opinion, which is likely to develop as military events on the ground unfold, economic shocks of the war reverberate through the global economy, and attention fatigue shifts media's attention to other topics. In this study, we examine snapshots of three events over three months, and examine the development of the main discursive actors and their narratives around major developments in the war.

To find the most prominent actors and sides of the debates around these three events, we employ community detection [212], which has been applied widely to social [242] and political [37] controversies around the world. Louvain, an iterative greedy community detection algorithm, has been used to identify different sides of debates around the 2019 European [182] and 2017 French [84] elections, the Munich July 2016 Attack [23], the 2016 UK Brexit vote [90], and the 2019 Brazil oil field auction [199], among many others. It has been shown that such communities closely approximate real-world relationships, such as political affiliation [42]. Such communities computed over successive time periods can then be used to build a community evolution graph [68] which shows the persistently influential users, as well as changes in community member-

ship and structure around events. Recently, similar methodology was employed by Evkoski et al. [69] who tracked the change of ex-Yugoslavian Twitter from before to after Russia invaded Ukraine. They find that some communities show a clear pro-Ukrainian (Croatian, Bosnian and Montenegrin) and others pro-Russian (Bosnian Serbs) stance. In this work, we perform a mixed-methods examination of communities around three events in order to discover (dis)continuity in the responses and framing of the conflict from the perspective of the Ukrainians and Russians, as well as the Western world (taking English as the *lingua franca*).

Once discovered, we use content and network analysis techniques to characterize the discovered communities. Centrality measures are often used to identify most central, or influential, actors within each community [182, 199] whose rhetoric is then examined at length. In this study, we use hashtags to describe the communities, as hashtags have been used as units of meaning for the description of various topics [87, 170]. We then take an approach from de Saint Laurent et al. [55] who treat memes as "partial stories" that add up to a narrative, which can be described using actors and their relationships. We apply their characterization of actors in a narrative (grounded in previous literature in rhetoric) as either Persecutor, Victim, Hero, and Fool, to the hashtags in our dataset. Further, we contextualize the extracted hashtags in the contemporaneous news and discussions to build cohesive narratives for each community involved in the discussion.

3.3 Methodology

3.3.1 Data collection and description

We used the public Twitter Streaming API to collect tweets about the Russo-Ukrainian war since 26 February 2022, resulting in a multilingual dataset capturing the conversations about the war worldwide. The keywords[4] used for this collection include 61 versions of the word "Ukraine"

[4]Available at https://tinyurl.com/ukraine22keywords

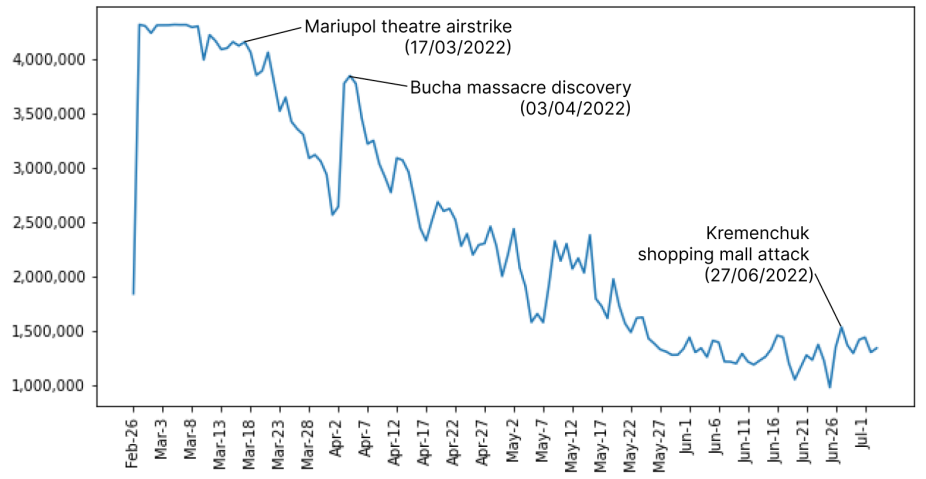


Figure 3.1: Number of tweets per day (all languages).

in different languages, as well as location-related keywords that spanned important geographic locations such as “Kyiv”, and “Donbas”, war-related keywords like “UkraineRussia”, politically-charged statements like “RussianAggression”, and “FreeUkraine”, and political figures such as “Zelensky” and “Putin”. The query was run without any location or language limitations. The data stream provides only public tweets, i.e. the collection does not contain tweets posted by private accounts or sent as direct messages. The data collection was still ongoing when this work was finished.

Figure 3.1 demonstrates the number of collected tweets (in all languages) per day up to July 4, 2022, when the dataset consisted of 312 182 245 tweets. The volume is highest at the start of the conflict, and decreases over time. Several peaks appear in March, April, and June. Because of the fast-pace development of the war, we chose three of the significant war-related episodes as follows:

- Mariupol theatre airstrike (17/03/2022) – Total number of tweets: 4 159 948;
- Bucha massacre discovery (03/04/2022) – Total number of tweets:

Table 3.1: Statistics of 10-day datasets for each event and language.

	Number of tweets	**Number of users**
Mariupol theatre airstrike		
English	25 834 917	3 357 113
Russian	872 487	134 528
Ukrainian	664 097	94 402
Bucha massacre discovery		
English	20 301 318	2 538 975
Russian	957 395	130 763
Ukrainian	633 417	83 334
Kremenchuk shopping mall attack		
English	7 732 416	1 438 390
Russian	530 422	80 198
Ukrainian	381 850	59 722

3 778 515;

- Kremenchuk shopping mall attack (27/06/2022) – Total number of tweets: 1 530 994

We used a 10-day window for each event: starting three days before and finishing seven days after. We limited the scope of the analysis of these event-related tweets to three languages: English, Russian and Ukrainian. We used the information about the language used in tweets provided by Twitter. We leave the analysis of other languages captured in the data for future work. Table 3.1 provides the volume and user statistics of tweets in the language/event datasets.

3.3.2 Network construction

Next, we use the collections of tweets described in the previous step for constructing retweet networks, which are known in the literature to capture the agreement with the opinion expressed in the retweeted mes-

sage [83, 48]. A retweet network is a directed weighted graph, where nodes represent Twitter users and edges represent the retweet relation. The direction of an edge corresponds to the direction of information spreading or influence; the weight of the edge is the number of times one user retweets the other [42]. We found that communication of users with the same language is stronger than across languages, resulting in language-based clusters if we build one multilingual retweet network per event. Thus, to study the opinion communities formed around war-related debates, we built nine retweet networks, one for each language, per each event.

Following previous literature [83], we exclude edges with a weight equal to one to reduce the noise in the data. Next, we extract the Giant Connected Component (GCC) for each network. Table 3.2 presents the sizes and modularity of the GCCs. The modularity of the networks in Russian and Ukrainian remain largely stable, but those in English increase substantially, suggesting that the communities in these networks became more dense and less connected among each other from one event to the next. In total, we analyzed $5\,251\,099$ unique users and $57\,908\,319$ tweets.

3.3.3 Community detection

We use the Louvain Community Detection algorithm [28] to identify groups of users who have similar opinions on the ongoing events. Louvain is a hierarchical clustering algorithm that maximizes the modularity score for each partitioning, determining the best number of communities. To test the stability of the communities identified by Louvain (which is not deterministic), for each network, we applied Louvain algorithm 10 times and calculated Rand index with the first community assignment described in 3.4.1. Average Rand indexes range in [0.94, 0.98] with standard deviation in [0.0040, 0.0141], indicating that the communities are stable. For visualizing the generated networks, we used the ForceAtlas2 algorithm [112] in Gephi. We then examined the communities having at least 1000 nodes for Ukrainian and Russian, and $10\,000$ for English.

We annotate each community with a topic/stance manually by exam-

Table 3.2: Statistics of GCC networks for each event and language: number of nodes, edges, and modularity.

	Number of nodes	**Number of edges**	**Modularity**
Mariupol theater bombing			
English	508 429	904 620	0.386
Russian	19 831	71 082	0.509
Ukrainian	19 303	73 416	0.426
Bucha massacre discovery			
English	387 039	1 674 804	0.478
Russian	20 417	72 893	0.512
Ukrainian	15 074	60 187	0.425
Kremenchuk shopping mall attack			
English	175 439	550 302	0.587
Russian	12 934	43 420	0.539
Ukrainian	9645	32 646	0.470

ining the top 20 retweeted users and top 20 retweeted tweets. The annotation was done by authors, who are familiar with the political situation in Russia and the West, and who are proficient both in Russian and English. Annotation of Ukrainian content was done with the help of machine translation, and was verified by a Ukrainian citizen familiar with the Ukrainian political scene.

3.3.4 Location Extraction

To assign a location to the posts, we employ geo-localization via users' Location profile field. The location information was matched to the GeoNames database[5] which includes locations around the globe, including in their native language spelling and with several variations [162]. Out of the selected language/event data, 42.5% were geolocated at a country level.

[5]https://www.geonames.org/

Locations for 1000 most popular matches in each language were manually checked and corrected where needed.

3.3.5 Bot-or-not annotation

We annotate the activity of Twitter accounts with Botometer (formerly BotOrNot) service. It gives each account a score based on information from follow-back groups, managing a high volume of political content, fake followers, spam, etc [245], with higher scores meaning more bot-like activity. During the preliminary evaluation of results, we found that Botometer gives a higher score to media and bloggers, which mostly have verified accounts. Therefore we exclude verified accounts from our analysis and collect scores of "regular" users only. For Russian and Ukrainian, we get scores for all users in the networks, but for English language users, we use stratified sampling using follower information to select 10% of all users in each event network.

3.3.6 Content analysis

Next, we perform content analysis using the tweets posted by users from different communities extracted above. Following previous literature [33], we use the odds ratio metric for detecting hashtags in tweets that distinguish a community from others during the event. The odds ratio is based on the frequency of hashtags that appeared in tweets of the target community compared to the same hashtag in other communities. We avoid low-frequency hashtags using the threshold of one standard deviation.

After detecting hashtags for each community, we find specific narratives they built during the analyzed events, in each language. We apply the narrative roles' framework [55] to actors consisting of persons, organizations, and countries. There are four roles: Persecutor, Victim, Hero, and Fool. These help us uncover scenarios that appear in tweets and compose the larger narrative structure.

3.4 Results

3.4.1 Community structure

We begin with the discussion of the overall structure of the networks across languages and events (resulting in GCCs of nine retweet networks), as seen in Figure 3.2. Although we did not pre-define the number of communities during the community detection, for each language-event combination our methodology resulted in 5-6 communities, which are identified in the Figure 3.2 with a different color, a bounding box, and a name given by the authors based on an examination of the contents. The size of the community in terms of users is also shown under the name. Note here that the names of the communities reflect the top retweeted accounts, not all accounts within the community, and thus signify a general interest of those included.

As can be seen from the structure of the networks, the communities in Russian networks are more disconnected than those in English and Ukrainian, likely due to the use of Russian language by those involved on both sides of the conflict. In all of the Russian networks, we find the Russian-affiliated and state media community to have few connections to the rest of the network, and to consistently encompass around 19% of the network's users. The rest of the network includes Russian opposition media and bloggers, as well as the Ukrainian bloggers and eastern-European media. This larger group also includes non-political Russian accounts and alternative media in the first two events (in orange), but not in the last, potentially due to Russia's blocking of VPN services in early June [109]. Note that the small gray cluster connected to Eastern-European media is a language misclassification of Mongolian content. The international botnet detected in the Kremenchuk network contains accounts that have been suspended at the time of writing, or accounts that retweet content in a variety of languages (some pro-Ukrainian) but who stopped posting in the summer of 2022.

The English-language networks also display a somewhat removed component that includes pro-Russia and anti-Ukraine communities (although

during Bucha the anti-Ukraine community is not distinguished by our method). Throughout these events, the pro-Russia/anti-Ukraine clusters account for 32%, 27% and 38%, respectively, seemingly not diminishing in popularity by the third event. Unlike in Russian language networks, in English we find the east-European media clustered topically: about politics or about war. Close to these, we further find pro-Ukraine media, as well as international media. The latter four communities are quite distinct during the first event (Mariupol), but are more integrated in the last (Kremenchuk), especially the pro-Ukrainian media, suggesting a cross-pollination of attention across time.

Ukrainian language networks are the most integrated and stable, composing of Ukrainian politicians, officials and defense organizations, news, as well as activists and military bloggers. The news agencies are often in the center of the networks, being retweeted by the rest.

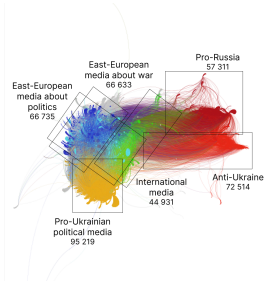
Thus, we conclude that the most prominent viewpoints represented in the discussion are largely stable over time in each language. We also find a polarization of opinion that is reflected in the networks in English and Russian, but not in Ukrainian, pointing to a unity of interpretation of the different events on the Ukrainian side.

3.4.2 Community behavior

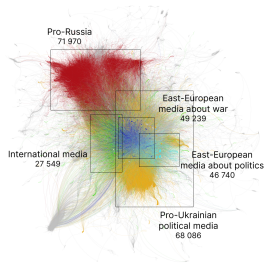
We continue by addressing specific research questions related to the behavior of the communities discovered in the previous section.

RQ 1.2.i. When were the accounts posting about the war created?

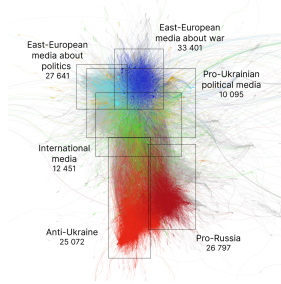
Figure 3.3 demonstrates the dates when accounts were created for each language, aggregated over all events. We find many accounts were created in 2022, suggesting that some users joined the platform specifically to engage in this topic. This spike is especially evident in Russian language for Russian opposition bloggers, as well as accounts around Eastern-European media and Ukrainian bloggers. In Ukrainian, many news agencies and politicians joined the network in 2022 (interestingly, not the officials, who were already on the network). Interestingly, a large spike in account creation appears in English in the pro-Russian commu-



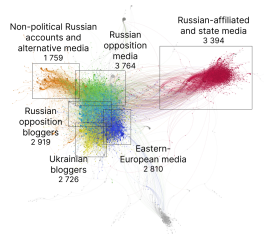
(a) English – Mariupol theatre airstrike



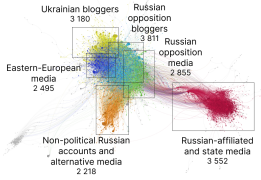
(b) English – Bucha massacre



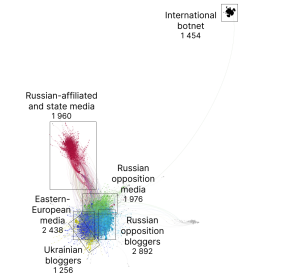
(c) English – Kremenchuk shopping mall attack



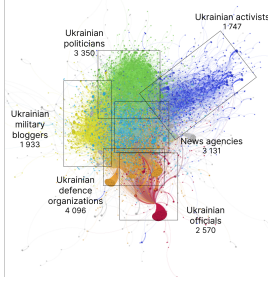
(d) Russian – Mariupol theatre airstrike



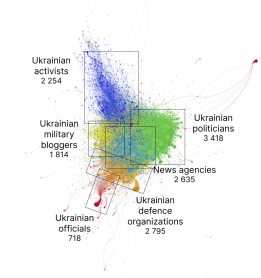
(e) Russian – Bucha massacre



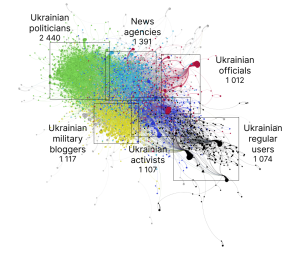
(f) Russian – Kremenchuk shopping mall attack



(g) Ukrainian – Mariupol theatre airstrike

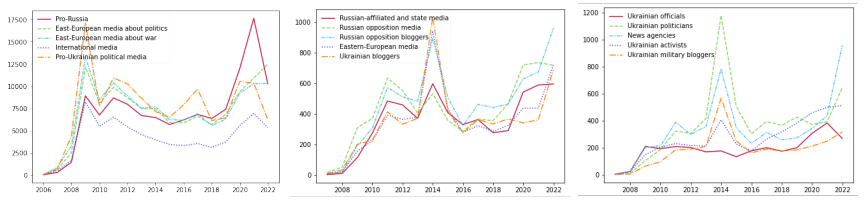


(h) Ukrainian – Bucha massacre



(i) Ukrainian – Kremenchuk shopping mall attack

Figure 3.2: User communities for each language and event, colored using Louvain algorithm, layout using the ForceAtlas2. Number of users in each community indicated under the description of each.



(a) English (b) Russian (c) Ukrainian

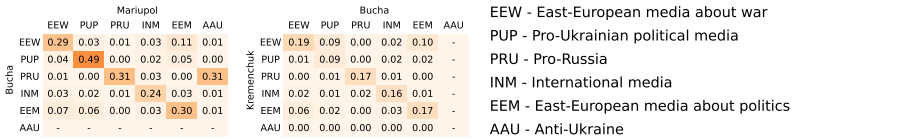
Figure 3.3: Account creation date (year) for each language, per detected community.

nity in 2021. A second large spike can be seen around 2014 for both Russian and Ukrainian during the annexation of Crimea, which is often considered as the start of the conflict. Such spike does not appear in English, and instead 2009 is a year when many accounts were created in all communities.

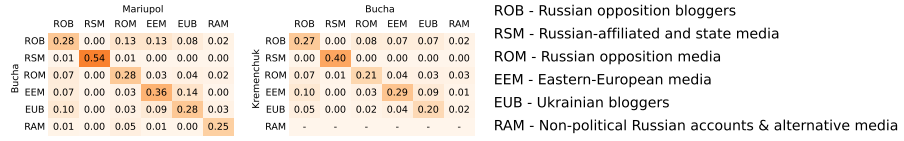
RQ 1.2.ii. How continuous is the discussion over the three events?

Although in Figure 3.2 the communities we found in each language/event instance had similar topical character, their membership was only somewhat overlapping between the three events. Figure 3.4 shows the Jaccard similarity in the membership overlap between the consecutive events (Mariupol/Bucha, Bucha/Kremenchuk), for each language. Note that for an account to "join" a community, all it takes is to retweet a certain central account in that community, and not others.

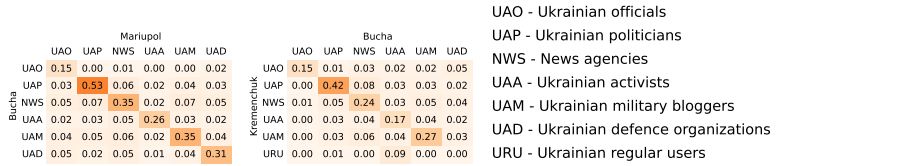
We can find that the overlap is highest between the communities having similar topics (on the diagonal), however we see several instances of communities sharing members across others in a subsequent event. For instance, in Ukrainian, the users in community we dub as Ukrainian activists during Bucha appear both in the eponymous community during Kremenchuk, but also among the Ukrainian regular users community. In English, many accounts in Pro-Ukrainian political media during Bucha then also appear in the cluster we dub as East-European media about war during Kremenchuk. Such "cross-overs" are more likely to happen among the groups sharing the same outlook on the conflict, such as accounts in



(a) English



(b) Russian



(c) Ukrainian

Figure 3.4: Jaccard similarity measuring the overlap in users of communities in consecutive events.

anti-Ukraine community during Mariupol who appear in pro-Russia community during Bucha in English. This supports our intuition that the users in these communities share a coherent opinion or stance on the conflict.

RQ 1.2.iii. How extensive was the use of bots by different sides, and how does this relate to suspensions by Twitter?

Figure 3.5 shows the median values of BorOrNot scores, percentage of accounts suspended by Twitter, and deleted by the user, aggregated over the three events. For English and Russian, we distinguish between communities supporting Russia (pro-Russia), and the rest (pro-Western). We find that in terms of the bot score, Russian pro-Russia, English pro-Russia, and Ukrainian users have about the same median score at 0.25. The score is slightly lower for English pro-Western and Russian pro-Western accounts. Comparing pro-Russia vs pro-West in English, we find p-value of 0.885, while comparing pro-Russia vs. pro-West in Russian re-

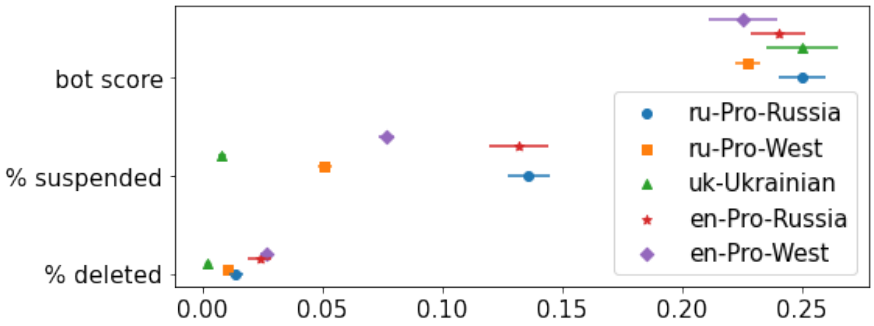
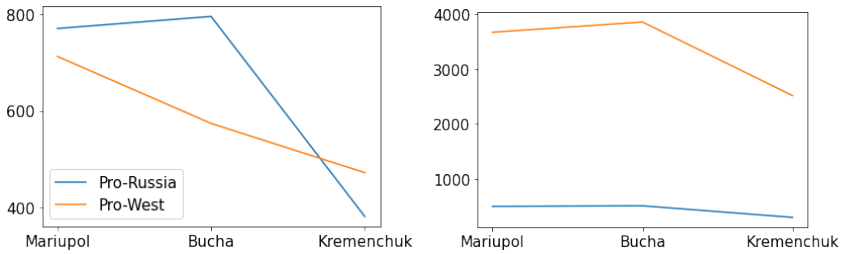


Figure 3.5: The medians (points) and standard errors (whiskers) of BotOrNot scores, percentage of accounts suspended by Twitter, and deleted by the user, aggregated over the events.

sults in $p = 0.048$ (using Mann-Whitney U Test). The account suspensions tell a clearer story: both Russian and English-language pro-Russia accounts are suspended the most (around 12%), whereas pro-Western less (around 6%), and Ukrainian accounts are suspended the least (about 1%). The difference between Pro-Russia and Pro-West in English is significant at $p = 0.004$ and between Pro-Russia and Pro-West in Russian is significant at $p = 0.011$. The deletions are very small for all kinds of communities at around 1-2%, but the Ukrainian ones are again the smallest. These results show that, despite having a similar bot score, the Twitter platform has suspended more pro-Russia accounts than pro-Western ones (though some of those were also suspended).

Next, we focus specifically on the Russian-language accounts. Figure 3.6(a) shows that pro-Russia communities inside Russia had disproportionately fewer engaged users around the third event, compared to both to pro-Russia communities outside of Russia, and pro-Western communities anywhere. Whereas in the first two events, just under 800 users were in pro-Russia communities, in the third, less than 400 engaged. The levels of engagement by pro-Russia communities outside Russia remained at around 500 at all events. We also find a decrease in engagement of pro-Western accounts inside Russia, but not to the same extent as the drop



(a) from Russia

(b) out of Russia

Figure 3.6: Number of users posting in each event in Russian, separately in communities having pro-Russia and pro-West leaning.

in pro-Russia ones. The results suggest that either censorship of Russian pro-Russia accounts by Twitter has been successful, or that interest in the event fell dramatically inside Russia (likely both have contributed, including other confounders).

3.5 Narrative Analysis

3.5.1 Prominent Actors

We explore the narratives of these communities guided by the top hashtags obtained using odds ratio, which compares how much more likely a hashtag is being used by a community versus all others (within a network). Top 10 such hashtags for each network can be seen in Tables 3.3 – 3.5. First, we annotated hashtags referring to actors (persons) using the narrative role framework [55], we then explored the context of these roles in the tweet content and the outside resources to which it links.

Only one actor was present during all events and in all three languages — the Russian president Vladimir Putin. In pro-Western communities, he is mostly portrayed in a role of persecutor (16 times in all languages/events), such as #PutinHitler #PutinIsaWarCriminal, although one was used in a sarcastic manner (role of a "fool") where #ZaПутина

Table 3.3: Top hashtags by odds ratio for each community in English, for communities that appear in all three events.

Community	Mariupol	Bucha	Kremenchuk
Pro-Russia	NaziUkraine, ZelenskyWarCriminal, NoWarWithRussia, UkraineNazis, BiolabsinUkraine, biolabs, WEF, AzovBattalion, NeoNazis, NeoNazi	DPR, ZelenskyWarCriminal, ScottRitter, NaziUkraine, WEF, AZOVNAZIS, NWO, AzovBatallion, SBU, AzovBattalion	NaziUkraine, Seversk, DPR, AlinaLipp, Aidar, LPR, ZelenskyWarCriminal, CIA, Washington, Lugansk
East-EU media - politics	GenocideOfUkrainians, CloseTheSky, ClosetheSkyoverUkraine, StandwithUkraine, StopRussia, HelpUkraine, StopPutinNOW, Nestle, StopRussianAggression, SlavaUkraïni	IStandWithUkraine, dogs, PutinHitler, BoycottRussia, StopWar, StopRussia, RussianAggression, SlavaUkraine, SlavaUkraini, StopPutinNOW	WalterReport, StopRussiaNow, RussiaWarCrimes, OpRussia, Anonymous, NAFO, OSINT, StopRussianAggression, RussiaIsATerroristState, ArmUkraineNow
East-EU media - war	Brovary, Hostomel, OSINT, BayraktarTB2, Bucha, Makariv, Gostomel, Latvia, Chernihiv, Kramatorsk	WALTERREPORT, LIVECHAT, Motyzhyn, OSINT, Hostomel, Sumy, russianinvasion, Kazakhstan, Zhytomyr, Estonia	Traitor, GOP, Brexit, SCOTUS, WarCriminal, Trump, Kazakhstan, PutinIsaWarCriminal, PutinsWar, SlavaUkraïni
International media	inflation, PMModi, crypto, VolodymyrZelensky, oil, food, Oil, gas, ElonMusk, energy	WATCH, UNGA, journalism, PopeFrancis, VolodymyrZelensky, Erdogan, Ireland, Refugees, auspol, G7	OOTT, oott, natgas, OPEC, ONGT, Egypt, ClimateCrisis, Türkiye, food, Indonesian
Pro-Ukraine political media	GOPBetrayedAmerica, MoscowMitch, FauxNews, PutinsGOP, TraitorTrump, ResistanceUnited, VoteBlueIn2022, GOPLiesAboutEverything, MadisonCawthorn, PartyOfTreason	DemVoice1, wtpBLUE, BlueVoices, FreedomForUkraine, VoteBlueIn2022, FreshResists, Fresh, VoteBlueToSaveDemocracy, ONEV1, LongLiveUkraine	FoodSecurity, TTE, EU2022CZ, URC2022, BackBoris, Lugano, StrategicConcept, Starmer, foodsecurity, humanrights

is accompanied with statements about using people as “cannon fodder”:

> Russia is gathering “cannon fodder”. Conscripts get ready. #WARINUKRAINE #war #RussiaUkraineWar #ForPutin #russia #conscripts #Ukraine #UkraineRussianWar #stopwar #Kramatorsk (Original: Россия набирает "пушечное мясо". Срочникам приготовиться #WARINUKRAINE #war #RussiaUkraineWar #ZaПутина #россия #срочники #Ukraine #UkraineRussianWar #stopwar #Kramatorsk)

However, upon closer examination, we also find tweets supporting the Russian president in a broader political context. In English language, we find mentions of larger geopolitical alignments (here, according to the eponymous Twitter user, TWGRP stands for Trump World Groups, and #pureblood is associated with those refusing vaccination):

> #Putin is cleaning house, taking out #Cabal #KHAZARIANMAFIA, etc and he is destroying #NWO plans. They ALMOST had us, were planning to wipe out 95% of the 6.8

Table 3.4: Top hashtags by odds ratio for each community in Russian, for communities that appear in all three events.

Community	Mariupol	Bucha	Kremenchuk
Russian-affiliated & state media	КрымНаш (Crimea is ours), USA, крым (Crimea), Крым (Crimea), Польша(Poland), crimea, Чернигов (Chernihiv), Donbass, Львов (Lviv), правда (truth)	Европа (Europe), Сербия (Serbia), бандерня (Banderites), Родина (homeland), укрофашизм (Ukraine fascism), USSR, укрофашисты (Ukraine fascists), Мариуполь (Mariupol), утилизация (utilization), Genocide	СВО (SMO), памятник (monument), история (history), ностальгия (nostalgia), usa, политика (politics), красота (beauty), Крым (Crimea), набережная-Салгира (shore of Salgira), интервью (interview)
Russian opposition media	санкции (sanctions), свободуНавальному (freedom to Navalny), NoWarInUkraine, свободуполитзаключённым (freedom to political prisoners), stopwar, StandWithUkraine, StopPutinNOW, PutinWarCriminal, кадыров (Kadyrov), FckPutin	СвободуНавальному (freedom to Navalny), Чернобыльской (Chernobyl), Днепр (Dnipro), Белгородскаяобласть (Belgorod region), РСФСР (RSFSR), Mariupole, фашистами (by fashists), noWAR, Нетвойне (no to war), хабаровск (Khabarovsk)	нетвойне (no to war), ПАСЕ (PACE), MH17, войнасукраиной (war with Ukraine), санкции (sanctions), СБУ (SSU), Германия (Germany), Кабаева (Kabaeva), РФ (RF), Австралия (Australia)
Russian opposition bloggers	русскиепапуасы (Russian-Papuans), таквижу (how I see), ооорф (ooorf), UKRAINE, ZaРоссию (for Russia), Бойкот (boycott), Война_Россия (War Russia), нефть (oil), колония (colony), Чернобаевка (Chornobaivka)	белгород (Belgorod), Azerbaijan, славаукраїні (Slava Ukraine), Рубцовск (Rubtsovsk), мародер (marauder), V, АйдосСарым (Aydros Sarim), ЗимняяВишня (winter cherries), НамНеСтыдно (we're not ashamed), БучанскаяРезня (Bucha massacre)	русские (russians), убейпутина (kill Putin), идлиб (Idlib), плохиеновости (bad news), русским (russians), лукашенко (Lukashenko), путина (Putin's), ПрямаяЛиния (Direct Line), PinkFloyd, уёбки (assholes)
East-European media	ЗСУ (AFU), UkraineUnderAttack, BayraktarTB2, NurembergTribunal, PutinsWarCrimes, Ucraina, Харків (Kharkiv), mariupol, Toronto, Маріуполь (Mariupol)	Казахстан (Kazakhstan), War, Литва (Lithuania), Байден (Biden), Bloomberg, RussianUkrainianWar, Рубежное (Rubizhne), Китай (China), NoWar, Арестович (Arestovich)	RussiaWarCrimes, UkraineWar, NATO, россиясегодня (RussiaToday), Kyiv, russiansoldiers, Оккупация (occupation), нато (NATO), новостионлайн (news online), War
Ukrainian bloggers	OSINT, новини (news), Краматорск (Kramatorsk), Потерьнет (no losses), RussiaInvadedUkraine, Кадыровцы (Kadirov's), Британия (Britain), Грецией (Greece), Дмитриев (Dmitriyev), потєрьнєт (no losses)	ZaПутина (for Putin), Дагестан (Dagestan), PutinHitler, Краснодар (Krasnodar), срочники (unprofessional soliders), РоссияСмотри (Russia watch), RussianWar, Соловьев (Solovyev), Гиркин (Girkin), Калмыкия (Kalmykia)	PoZZия (RoZZia), Яковенко (Yakovlenko), Гиркин (Girkin), ruZсолдат (RuZZian solider), Порошенко (Poroshenko), kremenchuk, kyiv, weaponforukraine, warukraine, WarInUkraine

billion world pop. And yet, so many ppl are #Brainwashed into hating #Putin. #TWGRP #pureblood

Indeed, English-language data has a strong presence of US politicians and pundits, often associated with the political right, posting pro-Russia content. For instance, among the top accounts in the English-language network's pro-Russia/anti-Ukraine communities during Kremenchuk shopping mall attack are U.S. representatives Marjorie Taylor Greene (@RepMTG), Lauren Boebert (@laurenboebert), Jim Jordan (@Jim_Jordan), and Tulsi Gabbard (@TulsiGabbard), and "GOP Committeewoman" Lavern Spicer (@lavern_spicer).

Table 3.5: Top hashtags by odds ratio for each community in Ukrainian, for communities that appear in all three events.

Community	Mariupol	Bucha	Kremenchuk
Ukrainian officials	російської (russian), свободуНавальному (free Navalny), Україні (Ukraine), WeStandWithUkraine, stopruSSiZm, Криму (Crimea), Німеччині (Germany), freeNavalny, ПАРЄ (PACE)	грузZ00 (cargoZ00), Грузія (Georgia), джавеліни (javelins), вч_57367, Грузії (Georgia), бурят (buryats), мрія (Mriya), війни (war), БМП (IFV), Бурятия (Buryatia)	Formula1, МАГАТЕ (IAEA), SlavaUkraine, Києві (Kyiv), Київській (Kyiv's), Avalanche, Lisichansk, F1, freeNavalny, свободуНавальному (free Navalny)
Ukrainian politicians	рф (RF), Білорусь (Belorussia), мова (language), UkraineUnderAttak, блокада (blockade), історія (history), МАРІУПОЛЬ (Mariupol), МаріупольУкраїна (Mariupol Ukraine), Японія (Japan), Крим (Crimea)	Крим (Crimea), США (USA), UkraineUnderAttack, Ізюм (Izyum), война (war), россия (Russia), RussianFascism, путін (Putin), RussiaWarCrimes, украина (Ukraine)	НАТО (NATO), RussiaWarCrimes, russiannazis, OTAN, ukrainewarvideos, RussiaInvadedUkraine, nazis, ВСУ (AFU), USA, Миколаїв (Mykolaiv)
News agencies	PutinHitler, SaveUkraine, kiev, NoFlyZoneOverUkraine, херсон (Kherson), HelpUkraine, Кім (Kim), Україну (Ukraine), Киев (Kiev), новости (news)	Война (war), санкции (sanctions), OlafScholz, новини (news), закон (law), УкрТВІ (UkrTWI), monobank, розкіш (luxury), коней (horses), Russianterror	Росії (Russia), HelpUkraine, PutinsWar, PrayForUkraine, FreeUkraine, новости (news), UkraineUnderAttak, StopWar, Russia, нюдсопятница (nude friday)
Ukrainian activists	StopBelarusianAggression, Сумщіна (Sumshchyna), СТОЇМО (We are standing), дякуюДСНС (thank you DSNS), RomanGaydakov, kievnow, паралімпіада (paralympics), Russians, tobiizu, дякую_захисникам_україни (thank you protectors of Ukraine)	SaveUkraine, PutinWarCriminal, EuropeanUnion, UkraineUnderAttak, украрт (ukrainian art), StandWithUkraine, PutinHitler, укртві (ukrainian twitter), Одеса (Odesa), RussiaUkraineWar	Макіавеллі (Machiavelli), Парадигма (paradigm), волонтери (volunteers), media, Херсон (Kherson), харьков (Kharkov), КАПІТУЛЯЦІЯ (capitulation), business, goodwilltodie, vacancy
Ukrainian military bloggers	FB, украина (Ukraine), news, світ (world), Dnipro, UkraineRussiaWar, Ізюм (Izyum), Чернігів (Chernihiv), StrongTogether, Азов (Azov)	Краматорськ (Kramatorsk), Бородянка (Borodianka), українці (ukrainians), War, Дніпро (Dnipro), putin, standwithukraine, UkraineRussianWar, StopRussianAgression, ClosetheSkyoverUkraine	PutinWarCriminal, Кременчук (Kremenchuk), путін (Putin), Допомога (help), Europe, Кременчуг (Kremenchuk), Haagen, WarCrimes, Zmiiny, Полонені (captives)

In Russian language, in pro-Russian communities Putin is often mentioned in the contexts of victories, often in tweets containing many hashtags, such as this one commemorating the annexation of Crimea:

> #Crimea #Feodosiya #Koktebel #Crimea #CrimeanSpring #CrimeaForPutin #WeDon'tLeaveOursBehind #CrimeanSpring #TogetherForever #CrimeaIsOurs Happy Day of Crimea's Reunification with Russia! And the cats agree with us too! [link] (Original: #Крым #Феодосия #Коктебель #Crimea #КрымскаяВесна #КрымZaПутина #СвоихНебросаем #КрымскаяВеснаВместенавсегда #КрымНаш С Днём воссоединения Крыма с Россией! И котики тоже с нами солидарны! [link]

On the other hand, pro-Russia communities prefer to write about more recent "heroes", for example, pro-Russia German Russian blogger Alina

Lipp [222] and former UN weapon inspector Scott Ritter who criticized NATO.[6] However, these personages serve as an introduction to content on larger actors, including the U.S., such as in the following tweet asserting the U.S. involvement in the training of Ukraine soldiers, with an emphasis on the controversial Azov battalion (see below):

> "We Trained #Nazis" - Former #US #Marine Corp Intelligence Officer, #ScottRitter "The first troops to be trained by $US and #British soldiers were the neo-Nazi #Azov Battalion" #RussiaUkraineConflict #Russia #Ukraine #News #istandwithrussia [link]

Similarly, the criticism extends to other Western countries, such as Germany:

> Germany is trying to shut her up from telling any further truth. #AlinaLipp #Donbass She revealed the fact that the Ukrainian military killed people by tying their hands in a hospital in #Mariupol. #Lyssytschansk #Lysichansk #Sloviansk #Severodonesk NO #FreeSpeech in Germany

A similar narrative structure is mirrored on the pro-Western side, wherein the Russian opposition leader Alexey Navalny appears in constant calls for his freedom in hashtags such as #freeNavalny (#свободуНавальному) during all three events. In that sense, the tweets use the ongoing Ukrainian conflict to remind the users of Twitter platform about this political prisoner:

> 423 days. We are not forgetting and every day reminding others! #FreedomToNavalny #FreedomToPoliticalPrisoners #UkraineRussiaWar #FreeNavalny #freeNavalnyNow #FreeUkraine #FreeUkraineResists [link] (Original: 423 дней. Не забываем сами и ежедневно напоминаем другим! #свободуНавальному #свободуполитзаключённым #UkraineRussiaWar #FreeNavalny #freeNavalnyNow #FreeUkraine #FreeUkraineResists [link]

[6] https://en.wikipedia.org/wiki/Scott_Ritter

The next most popular actor is the Ukrainian president Volodymyr Zelensky, who was framed both as a persecutor and a hero. In the pro-Russia community, he is the principal persecutor (#ZelenskyWarCriminal), while international media gave him a hero role.

During the three events, the pro-Russian communities associated Zelensky with manipulation of the narrative around the unfolding events, for instance, asserting his involvement in the coverage of the Bucha incident:

> #Ukraine #bandernya #ukronazism #ukrofascism Zelensky banned the removal of corpses from the streets in Bucha to show them to Western "spectators" This was stated by the former deputy of the Verkhovna Rada Ilya Kiva. He explained that Ukrainians need corpses.. [link] (Original: Readovka #Украина #бандерня #укронацизм #укрофашизм Зеленский запретил убирать трупы с улиц в Буче, чтобы показывать их западным «зрителям» Об этом заявил бывший депутат Верховной Рады Илья Кива. Он пояснил, что трупы нужны украин.. [link]

As well as more mild comments on Zelensky's speeches:

> #Zelenskyy quoting "I Have A Dream" is peak disrespectful to all the oppressed ppl around this world. #NaziUkraine

Interestingly, the users posting criticism of Ukrainian side showed awareness of the platform's policies, and even informed the followers that "Twitterban" was coming (indeed, the tweet has been removed at the time of writing):

> #ikvertrek #vonderLeyen #wef #Ukraine #ZelenskyWarCriminal #marathonrotterdam #biden Ukranian military beating ukranian people. Thats what zelensky dont want you to see. (Twitterban incoming) [link]

Mentions of Zelensky on the pro-Western side are much milder, associating him with news coverage with other leaders, and using his name as a hashtag alongside other encouraging messages:

> @ZelenskyyUa @JustinTrudeau Hold on, dear! To OUR heroes, glory #ukraine #kiev #russia #stoprussia #kyiv #stopputin #lviv #ua #StayWithUkraine #Ukraine #SaveUkraine (Original: @ZelenskyyUa @JustinTrudeau Тримайся, рідна! НАШИМ героям Слава #ukraine #kiev #russia #stoprussia #kyiv #stopputin #lviv #ua #StayWithUkraine #Ukraine #SaveUkraine

3.5.2 Prominent Entities

Beyond individuals, we also find groups of people and organizations mentioned as actors. We find both of the two primary opponents in the war as prominent actors. Russia is seen a persecutor 18 times (#RussiaIsATerroristState, #RussiaWarCrimes), and in some cases we find dark humor where Russia is implied to be a business (#ооорф) or an "uncivilized tribe" (#русскиепапуасы (#RussianPapuans)), or where the hashtag is used sarcastically (#ZaРоссию (#forRussia)).

Some tweets address Russian people, for instance in this tweet, which links to a video message from the Ukrainian professional boxer Wladimir Klitschko:

> And where is the one in the bunker, from OOORF? #RussianPapuans #ooorf #nowar All Russians watch! You #Russians laughed at #Klitschko's phrases. Look here for everyone: The guy learned #Russian for conversion! On the video: a patriot of the Motherland and a man - the guardian of the Fatherland! [link] (Original: А где бункерный, из ОООРФ? #русскиепапуасы #ооорф #нетвойне Всем россиянам смотреть! Вы, #россияне смеялись над фразами #Кличко. Сюда всем смотреть: Парень выучил #русский язык для обращения! На видео: патриот Родины и мужчина - охранитель Отечества! [link]

In addition to critique from both sides (similar to that involving individual actors described above), the data includes many instances of on-the-ground coverage of the ongoing conflict. For instance, the destruction of Antonov An-225 Mriya, a large cargo aircraft damaged in the early days of fighting (link leads to an Instagram post):

> This is how the airport Gostomel looks like, where there is the wreckage of the airplane An-225 “Mriya” #Mriya #AirplaneMriya #News #Ukraine @ Ukraine [link] (Original: Так выглядит аэропорт Гостомеля, где находятся обломки самолёта Ан-225 «Мрія». #мрія #самолетмрія #новости #україна @ Ukraine [link]

Others come directly from the ongoing operations, such as a TikTok video of a soldier aiming a video at a target (note that here the use of country-wide hashtags in both Ukrainian and Russian suggests the intended audience to be in both countries):

> Minus the racist “Buratino” #video #war #Ukraine #Ukraine #war #russia #russia #Stuhna [link] (Original: Мінус рашистська "буратіна"#Відео #війна #украина #україна #война #россия #росія #стугна [link]

Note that the posting of such sensitive information while the conflict is ongoing may reveal important information about the location, capabilities, and other situational details. The use of cellphones in general has been linked to breaches of security, which have cost lives on the battlefield [120], however due to extremely sensitive nature of this topic, we leave such analysis to security researchers.

An entity that is a popular target on both sides of the conflict is the Ukraine’s Azov regimen, which is a part of the National Guard of Ukraine, and which have been associated with far-right [88]. The narratives around the battalion are starkly divided by the side, and are often written in the language of the targeted audience. For instance, a tweet accusing Azov of using civilians as human shields is in English:

> Translated video: 3 residents of #Mariupol. #UkraineNazis from Azov set up artillery in residential areas, used civilians as #humanshields. #Ukraine troops prevented people from leaving & #DNR & Russian troops rescued them. Wake up #sleepyjoe! @miniakissa (link to deleted tweet)

Whereas a reaction about the Kremenchuk shopping mall bombing is in Russian:

> #kremenchuk #RussiaIsATerroristState In broad daylight! Shopping mall!!! Again, Russia will say military facilities?! Or another NATO base with a food basket?! Or combat t-shirts with pants?! Oh of course! Another Azov base?! Inhumans!!! (Original: #kremenchuk #RussiaIsATerroristState Среди белого дня! Торговый центр!!! Опять Россия скажет военные объекты?! Или очередная база нато с продуктовой корзиной?! Или боевые футболки со штанами?! Ах конечно! Очередная база азова?! Нелюди!!!

Previous research has shown that language-specific targeting can be important in a political sphere [24]. In the present case, despite being annotated by Twitter as Ukrainian, users from communities in Ukrainian dataset actively used hashtags in all three languages. Out of the top hashtags by odds ratio in Ukrainian dataset, 39% are in English, whereas only 34% are in Ukrainian and 17% are in Russian (the last two languages have very similar alphabet and many common words, the last 10% of hashtags could be in Ukrainian or Russian). Whether such linguistic targeting achieves a wider circulation in the desired audiences during an international crisis is an interesting research direction.

Outside of the battlefield, other groups used Twitter to rally support for their cause. For instance, women's groups used the platform to organize a march to fund-raise:

> #Ukraine #Bucha #IStandWithUkraine With each day, we receive more heartbreaking evidences of sexual violence toward women in Ukraine Support organisations who are assisting women with medications, food and shelter National "Women's March" [link]

3.5.3 Narrative Evolution

As a case study, we focus on the Russian-language, pro-Russia community, which shows an emotional arc throughout the three events: the first focuses on places involved in the war effort (in Ukraine, and also globally), the second besmirches the opposition (calling Ukrainians #ukrainofascist), and the last brings nostalgic nationalist feelings (around #history, #monuments, #beauty and #nostalgia). During the Mariupol bombing, for instance, a music video featuring Chechen fighters was produced by the Minister of Culture of the Chechen Republic and circulated in Russian (but with English hashtags):

> Aishat #Kadyrova dedicated the song to the events in #Ukraine and the Chechen fighters #Mariupol #Ukriane #Donetsk #Kiev #Donbass #Russia [link] (Original: Айшат #Кадырова посвятила песню событиями на #Украине и чеченским бойцам #Мариуполь #Ukriane #Donetsk #Kiev #Donbass #Russia [link]

Bucha, on the other hand, is marked by the information war, wherein links to materials contradicting Ukrainian narrative are shared. For instance, the following tweet links to a VKontakte (Russian website similar to Facebook) post with a video (the video in question has been debunked [57] by Western media).

> Fragments of a video appeared in the media, in which the Ukrainian military #ukronazism #ukrofascism #ukraine lay out corpses in Bucha. According to the director's idea, all the corpses should be in the frame. It does not matter that the corpses will lie on the road in the direction of travel.. [link] (Original: В СМИ оказались фрагменты видео, на котором украинские военные укронацизм укрофашизм Украина раскладывают трупы в Буче. Согласно задумке режиссера — все трупы должны попасть в кадр. Не важно, что трупы будут лежать на дороге по ходу движе.. [link]

An exhaustive search of all debunked content is outside of the scope of this study, but we will make the data available to researchers interested in this field, in accordance with Twitter's Terms of Service (see below).

Finally, during the third event we find several campaigns around #history, #monuments, etc. We admit that the topic of these tweets is not necessarily about the Ukraine war, but as they matched on of the query terms (Russia in several languages), they provide a context of self-reflection co-occurring with the war events. For instance, one tweet listed prices of products during the rule of Leonid Brezhnev:

> With nostalgia for the USSR: Prices in Brezhnev's Time. • ° #Brezhnev #USSR #history #products #Russia #nostalgia #prices #profits (Original: С ностальгией об СССР: цены во времена Брежнева . • ° #Брежнев #СССР #история #продукты #Россия #ностальгия #цены #прибыль [link]

Thus, even within a stance and a language, we find a variety of narrative constructions, encompassing encouragement of friends and criticism of enemies, propagation of content from other sources supporting one's narrative, and building a larger historical context. Whether these narratives reach similar audiences within Russia, and how they may reinforce or complement each other would be important in understanding the impact of the cultural delineation of the Russian (as well as Ukrainian) side, alongside the purely informational warfare.

3.6 Discussion and Conclusions

Forbes dubbed the ongoing Russo-Ukrainian war as the "first social media war", which follows the Gulf War as "first Internet war" and the Vietnam as "first television war" [223]. Unlike during the "Arab Spring" revolutions when social media was widely acknowledged to provide a supportive role to the protesters, in this case the Ukrainian government is encouraging the use of social media, with some of the first-hand content potentially being used for war crime investigations [4]. However,

on the Russian side, state control over internet is increasingly creating a new Iron Curtain (or "splinternet" [45]), with the 2019 Sovereign Internet Law and the blocking of VPN services adding hardware to the many speech restrictions already on the law books [104]. Twitter platform itself is another important player in the possible propaganda war. In April 2022 it has announced that it will update its content moderation rules to limit the propagation of posts by Russian government accounts (though not remove them) and ban images of prisoners of war [241]. On the other hand, Twitter has made efforts to be accessible in Russia by launching a privacy-protected version on Tor to bypass surveillance and censorship by the Russian state [179]. Despite the efforts by several sides to curate the conversation, we find both pro-Russia and pro-Western sides represented in English (around 35%) and Russian (19%) networks (but not Ukrainian ones). The pro-Russian communities (in Russian) were rather stable in membership across the events, however the audiences around Ukrainian politicians (in Ukrainian) were similarly stable, compared to most other kinds of communities. Interestingly, the Russian-language pro-Russia communities inside Russia have dropped precipitously by the third event, suggesting either a drastic lack of interest in the ongoing conflict, or a potential success of the curation from the Western side. Ongoing monitoring of the engagement by these communities, especially state-sponsored or potentially automated accounts, is an exciting continued direction of this research. In this work, we have found the Botometer score to be uninformative in terms of distinguishing the different communities, especially since by manual examination we found the score to be high for popular accounts.

The battle over the narrative around the war is especially important for Ukraine, as it depends on the support of its Western allies to resupply its army, house its refugees and economic sanctions, as stated by the Ukrainian president himself [132]. The Ukrainian government and numerous volunteer teams strive to create war-related content, that tells the international audience that "Ukraine is actually capable of winning" [5]. Following the Russian army's retreat from Kyiv region after initial foray for the capital [61] and later from Kherson [111], Russia is also careful to

manage the image of military setbacks and direct blame (away from Putin [152]). The role of language should not be underestimated, as it serves both as a way to bridge nations and a way to establish one's own national identity. We find that the hashtags used by in the ostensibly Ukrainian dataset were actually quite multilingual, potentially with the aim of reaching international audiences. As language use has been an important part of nation-building [184], the use of Ukrainian vs. Russian may become a contentious point in Ukraine. A poll in March 2022 has found that 12% believe that there may be issues between Ukrainian-speaking and Russian-speaking citizens of Ukraine, and that it's a threat to domestic security [193], and situation may change as the conflict unfolds. The development of identifying markers, national branding, and framing of the conflict will continue to be central to both parties in the conflict.

3.6.1 Limitations and Privacy

There are some notable limitations to this study, some of which have already been outlined above. Twitter is by far not the only deliberative space on internet, with Facebook and WhatsApp groups playing an important role in the information sharing in the ongoing conflict [235, 78]. As researchers, we are also not in the purview of the policy and algorithmic changes the platform took to direct the discussion, not mentioning the throttling and blocking by nation states. Further, methodologically we acknowledge that the set of keywords used in this study may not cover all of the conversations around the war, especially since it was not changed throughout the collection for consistency. Similarly, the output of Louvain algorithm is guided by the retweet information in the network, not by the ideological stances of its members, possibly resulting in heterogeneity of opinions within the clusters (however upon manual inspection, we found the sampled users and content to be largely cohesive). Also, the case studies presented here are snapshots of particular instances of the many events taking place in the war, and continued analysis is needed to detect the further evolution of the conversation from the many parties interested in this topic.

As we mentioned above, the data used in this study was collected using the Twitter Streaming API, and it thus includes posts which were public at the time of the collection. However, by the time of the analysis it also contains tweets which were later deleted or whose posters were suspended by Twitter. For this reason, as well as due to the limitations by the Twitter Terms of Service,[7] we do not make public the original tweets used in this study. Further, it is possible that, if the geo-location of the users captured in this data is accurate and precise, that we have captured potentially sensitive data coming from the war zone itself. We urge the research community to be mindful of the potentially vulnerable populations captured in social media data, and to take steps to limit the exposure of sensitive data.

3.6.2 Disclaimer of Positioning

Due to the sensitive nature of the subject, and strong polarization on the topic, the authors would like to disclose the authors admit a pro-Western bias in the evaluation of the conflict. This has likely colored the positioning, methodology, and interpretation of the results.

3.6.3 Reproducibility

The tweet IDs for the examined events, including community labels, are made available according to the Terms of Service of Twitter at public repository.[8]

[7] https://developer.twitter.com/en/developer-terms/agreement-and-policy
[8] https://github.com/vitiugin/war_narrative

Chapter 4

MULTILINGUAL INTERACTIVE ATTENTION NETWORK

In the previous chapter, we explored the variations in data shared on a social media platform during three crisis events across different languages. While performing that research, we discovered a significant prevalence of offensive content posted by users during these events. Building upon this observation, the focus of this chapter is to present a novel model designed for detecting hate speech in multilingual social media.

Our approach combines attention networks for interpretability and adopts a human-in-the-loop paradigm to enhance the adaptability of the model. By incorporating attention mechanisms, our model can dynamically learn to prioritize relevant contextual words, enabling a more accurate identification of hate speech. Additionally, we leverage the labels for hate target mentions obtained through simulated human feedback, creating an interactive learning process that strengthens the model's detection capabilities.

This chapter delves into the development and implementation of our proposed model, highlighting its unique features and the underlying methodologies. By harnessing the power of attention networks and human-in-

the-loop adaptability, we aim to address the challenges of identifying hate speech across multiple languages in social media platforms.

4.1 Introduction

Hate speech has become a general phenomenon in modern society. Particularly, the prevalence of hate speech is pronounced on social media platforms and other means of online communication. Users often anonymously and freely express themselves in online communication forums, including social media. The ability to express freely oneself is an important human right, but inducing and spreading hate towards another group is an abuse of this liberty. While the research in hate speech detection has been growing rapidly, multilingual hate speech detection is still a challenging task. Most of the existing research studies [19, 157, 153] have focused on one language only (mainly English), and their methods often depend on external knowledge sources, such as a hate speech lexicon [91, 240, 51]. These sources are resource-intensive and time-consuming to create in every language. A key challenge of these methods is the obsolescence of the data source, given the developing language of user-generated content on social media. The language changes quickly, especially under the pressure of moderation, which brings us to the next important challenge. Current multilingual hate speech detection models cannot effectively deal with the local derogatory slangs in specific language (e.g., 'sudaca' is xenophobic term to call people from South America by Central Americans and North Americans who speak Spanish language) and local context of implicit hate speech (e.g., 'building a wall' could implicitly refer to hate against immigrants). Another enormous challenge is a significant drop in the performance when the existing models are tested on datasets different from training data [91] in terms of the target of hate speech such as *immigrants* versus *women*. Examples of such tweets are presented in Table 4.1.

Last, the state-of-the-art hate speech detection models have used deep learning techniques, however, the decisions made by the deep learning

Figure 4.1: Attention weights distribution in English and Spanish texts. The darker color of the word means the higher weight.

Table 4.1: Example of messages with different targets of hate speech.

Message	Target	
The U.S must stop importing the Worlds Poor if they cant take care of themselves #sendthemback Stop allowing Foreigners to live off U.S Taxpayers #Trump #MAGA	Immigrants	Group
@USER @USER You won the “life time recipient for Hysterical Woman” a long time ago	Woman	Individual

models can be opaque and difficult for humans to interpret why the decision was made and analyze the model errors. While human-in-the-loop

paradigm has been shown to assist such techniques, there is still a challenge for the ability of humans to provide effective feedback to the model to improve it. To address this challenge, we hypothesize that a theoretical approach of *frame semantics* from cognitive linguistics [74] can help better explain and rectify the model reasoning provided through attention weight map (c.f. Figure 4.1). Frame semantics suggests that word meanings are defined relative to frames in a given text and thus, if the model can learn to give attention to the elements of the correct frame as per human interpretation, the model performance could improve. For example, in Table 1, in order to correctly interpret the posts, a model will need a good understanding of the targets of the hate that could be easily understood by evoking a specific frame for interpretation by a human when looking over the attention maps.

This study investigates the following research questions:

- RQ 2.1. Can a hybrid method of Interactive Attention Network (IAN) with human-in-the-loop approach improve the detection of hate speech in multilingual data with local slangs and implicit context for hate?
- RQ 2.2. Is there an effect of framing in the human feedback to IAN that helps toward faster convergence for the hate detection model?
- RQ 2.3. How much human feedback is required for significant improvement of hate speech detection results?

To address these questions, our method relies on including minimal human guidance in the training process of IAN classification model for achieving higher performance. Human feedback helps to detect hate subtleties and phrases for extracting features during model training, where the human feedback is guided by common element of the frames to express a hate speech, i.e. hate targets. Explanation of decisions of the IAN model with the help of human feedback is analyzed by comparing the distribution of attention weight maps and semantic frames in the textual posts. The experimental dataset includes social media posts from two different topics in two languages. The proposed method allows the design

of a novel multilingual hate speech detection system with the help of humans that shows high level of performance and can explain decisions by demonstrating an attention weight map (c.f. Figure 4.1) of the analyzed texts.

The main contribution of this study is a Multilingual Interactive Attention Network (MLIAN) model for detecting hate speech in text, regardless of language. We not only show improved model performance compared to baselines in two languages but also show a principled way of integrating frame semantics for analyzing the interpretability of the model reasoning. A comparison of distributed attention weight map with semantic frames shows that our model accurately captures implicit frame elements in the text that help to detect hate speech. We achieved this with simulated human feedback and identified the minimal level of feedback required to improve the model performance compared with the baselines.

The remainder of this chapter is organized as follows. We first describe the related work, followed by our MLIAN methodology, experimental setup, and then, result analyses.

4.2 Related Work

Spreading hate towards distinct groups is an abuse of human rights to express themselves freely. Many online forums such as Facebook and Twitter have policies to remove hate speech content [70], albeit detecting hate speech is challenging. We summarize the definitions, existing detection techniques, and the role of human feedback to improve them.

4.2.1 Hate Speech Definitions

There are many definitions of hate speech that make the task specification of the detection of hate speech difficult. Here are some examples of such definitions: (1) "Hate speech is speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity." [133] (2) "We

define hate speech as a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We also provide some protections for immigration status. We define attack as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation." [70] (3) "Language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group." [51] (4) "Hate speech is a language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used." [76] Some definitions above consider hate towards a group while others consider attacks on an individual. A general observation among the definitions is that some aspect of the group's or a person's identity becomes a base for the offense, however, in the given text it may not be explicitly stated, as shown in Table 4.1. While in one definition the specific identity aspect is ignored, other definitions provide specific identity characteristics. These ambiguities can challenge the task specification and what conventional text-based classification approaches could capture, especially when data is multilingual. Thus, human feedback could be valuable to support model training process.

4.2.2 Hate Speech Detection Models

The earlier efforts to build hate speech classifiers used both simple methods using dictionary lookup [93] and bag-of-words features [32] as well as deep learning techniques. SVM classifier is widely used for hate speech detection. Training includes diverse features such as character n-grams, word n-grams, word skip-grams, and knowledge-base features [154, 215]. Also, the list of features can include Brown cluster features along with approaches including ensemble classifiers and meta-classifiers [155]. During the analysis of the TRAC-1 workshop results [131], we found that

authors use both deep learning (e.g., Long Short Term Memory (LSTM), BiLSTM, CNN) and traditional machine learning classifiers (e.g., SVM, Random Forest, Naive Bayes).

The HASOC 1 workshop organized at FIRE 2019 [156] notes that the most widely used approach for hate speech detection in Indo-European Languages was LSTM networks coupled with word embeddings. The participants used a wide variety of models such as BERT, SVM, and LSTM. Furthermore, a unified deep learning architecture based on LSTM networks reached high performance without change of the architecture but only training a model for each task (i.e., different abusive behavior types) [77]. Results of the most recent shared tasks in aggression identification and misogynistic aggression identification show that the superior performance of the SVM classifier was achieved mainly because of its better prediction of the majority class. BERT-based classifiers were found to predict the minority classes better [20].

Research into the multilingual aspect of hate speech is relatively new. Using Twitter hate speech corpus from five languages annotated with demographic information, the authors of [103] studied the demographic bias in hate speech classification. Hate speech detection models based on SVM and BiLSTM show outstanding performance on three datasets from three languages (English, Italian, and German) [47]. Moreover, large-scale analysis of deep learning models to develop classifiers for multilingual hate speech classification (16 datasets from 9 languages) shows that for low resource languages, LASER embedding with logistic regression performs the best, while in a high resource setting BERT-based models perform better [14]. One limitation of these methods is the lack of interpretability of the reasoning for the models' decisions.

4.2.3 Human-Machine Collaboration for Hate Speech Detection

Complex behaviors such as hate speech require efficient automated models for detection of such communication. Previous work in this direction informs the requirement for continuous model transformation techniques

[234] due to the complexity of the task. The findings of human-machine collaboration for content regulation (based on the Reddit case) suggest a need for tools to help tune the performance of automated modeling mechanisms, a repository for sharing tools, and improving the division of labor between human and machine decision making [115]. Some existing approaches advocate the use of external knowledge sources, such as a hate speech lexicon, where detection systems could leverage multilingual, fine-grained Profanity and Offensive Word (POW) lexicons (an NLP resource for toxic language) [56]. This type of approach can be effective but it requires developing these knowledge sources that is labor-intensive, especially for multilingual setting, and furthermore, such sources need to be up to date, which is not always possible. Thus, an effective alternative can be a human-machine collaboration to fine-tune an automated hate speech detection model during the training/re-training process, with targeted human feedback to improve the model's understanding for the hate speech context.

4.3 Methodology: MLIAN model

Recent approaches for hate speech detection propose solutions that use deep learning techniques for text classification of hate speech. While these solutions make decisions automatically, they make errors due to the biases in learning patterns and the reasoning behind those decisions can be difficult to interpret and unclear for humans. The resulting systems that automatically censor social media posts would end up needing a human's attention for majority of the appealed cases. The multilingual content can make such systems even more human resource-demanding. In this section we describe our proposed MLIAN model that can enable efficient human-in-the-loop paradigm along with interpretability of multilingual hate speech classification decisions, by employing a meaningful human feedback guided through frame semantics theory in the deep learning architecture of interactive attention networks.

4.3.1 Interactive Attention Networks

Our method builds upon the Interactive Attention Network (IAN) architecture. Deep learning models are widely used for hate speech detection tasks [19], but many of such models make automated decisions hard for understanding, or can be explainable only by using special techniques [153]. In recent years, models with attention mechanism have not only shown good performance, but also can be used as a tool for interpreting the behavior of neural network architectures [81, 54, 80]. In the processing of natural language, the tokens composing the source text are characterized by having each a different relevance to the task at hand. The attention mechanism constructs the context vectors of the tokens that are required by the decoder to generate the output sequence in the encoder-decoder neural architecture. The IAN model was proposed for interactive learning of attentions in the context vector and special tokens (i.e. hate targets in our study), and generate the representations for the special tokens and contexts separately. The IAN model has shown high performance results in many tasks such as aspect-level sentiment classification [151], adverse drug reactions [12], pedestrian detection [250], and other classification tasks.

Unlike the existing methods for hate speech detection task that mainly work for monolingual data or require special linguistic resources created with labor-intensive efforts, we propose to adapt this IAN model. It can facilitate an approach for multilingual hate speech detection with integration of the human-in-the-loop paradigm to improve both the model performance and interpretability. The role of human agents is to provide more contextual information about hate speech that could be informed by appropriately invoking the correct frame. Specifically, we detect whether a given text message has personal or group hate target as an additional parameter for interactive model training. We choose this feature because it can be extracted in real cases creatively – users of the social media platforms can provide the appropriate frame of interpretation and identify the hate target to themselves and label the posts at scale. We evaluate our proposed model against state-of-the-art baseline methods in Section 4.4.

4.3.2 Model Architecture

The overall architecture of MLIAN model is shown in Figure 4.2. MLIAN model contains two parts that interactively model the targets (left part in the Figure 4.2) and context (right part in the Figure 4.2).

Our specific steps are as follows. First, we use multilingual text embeddings as input to LSTM. It is employed to obtain a target and hidden states of words on the word level for targets and context respectively. Second, we calculate the average value of the targets' states and the contexts' *hidden states* to supervise the generation of attention vectors, with which the attention mechanism is adopted to capture the important information in the context by the target provided by human feedback. This type of architecture design [151] enables to capture the influence on the context from the identified target and the influence on the target from the context. This approach provides more clues to the modeling algorithm to pay attention to the contextually-relevant hate speech features and thus, allows to generate their effective data representations interactively. Finally, target and context representations are concatenated as a *final representation* for an input text that is fed to a softmax function for hate speech classification.

4.3.3 Transformer-Based Multilingual Embeddings

We use embeddings generated by two pre-trained transformer-based models for representing the input data for MLIAN: LASER [17] and Distilled version of multilingual BERT (DistilmBERT). The main difference between these two models is that LASER generates sentence-level embeddings while DistilmBERT generates word/token-level embeddings.

LASER. Results of previous large-scale analysis of multilingual hate speech detection in 9 languages from 16 different sources demonstrate that simple models such as LASER embeddings with machine learning algorithms perform with the best results [14]. LASER is based on an architecture to learn joint multilingual sentence representations for data in 93 languages. Given an input sentence, LASER provides sentence embeddings which are obtained by applying max-pooling operation over the

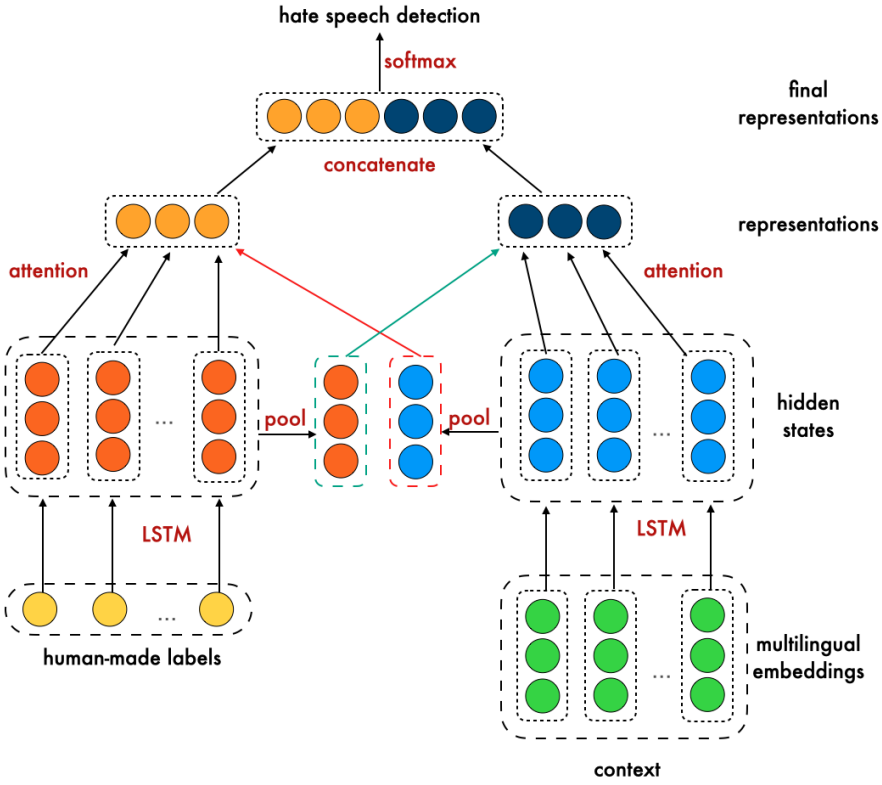


Figure 4.2: The overall architecture of MLIAN model.

output of a BiLSTM encoder. BiLSTM output is constructed by concatenating outputs of two individual LSTMs networks working in opposite directions (forward and backward). This way more contextual information is included in the output than a single LSTM reading text from left to right. The system uses a single BiLSTM encoder with a shared Byte-Pair Encoding (BPE) vocabulary for all languages, coupled with an auxiliary decoder, and trained on publicly available parallel corpora. In our experiments, all sentences are initialized by LASER in 1024-dimension fixed-size vector to represent the input textual post. The resulting embeddings

are computed using English annotated data only, and transferred to any of the 93 languages without any modification. Experiments in cross-lingual natural language inference (XNLI dataset), cross-lingual document classification (MLDoc dataset), and parallel corpus mining (BUCC dataset) have shown the effectiveness of the LASER approach [17].

DistilmBERT. Distilled version of multilingual BERT (DistilmBERT) model pre-trained by HuggingFace[1] is a distilled version of the multilingual BERT-base model [213]. The model is trained on the concatenation of Wikipedia in 104 different languages. The model has 6 layers, 768 dimension and 12 heads, totalizing 134M parameters (compared to 177M parameters for the multilingual BERT). On average DistilmBERT is 60% faster than multilingual BERT model. All DistilmBERT embeddings have 512-dimension fixed-size vector representation. BERT's key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modeling [59]. This is in contrast to previous efforts which looked at a text sequence either from left-to-right or combined left-to-right and right-to-left training. We use BERT because it shows high performance in many NLP-tasks including hate speech classification [167, 210, 64].

4.3.4 Human Feedback Guided by Frame Semantics Theory

While human could give a variety of feedback to the MLIAN model for hate speech detection, we propose to guide the nature of the feedback. One common approach of human-in-the-loop machine learning paradigm is to seek feedback from humans as the correct labels at the entire text level. Alternative to this approach can be the feedback at the level of word tokens in the text, but the question is how to create a principled approach to identify the special tokens for the feedback. We explore frame semantics theory from cognitive linguistics [74] that can help better explain and guide the types of special tokens to focus for the human feedback, in or-

[1]https://huggingface.co/models

der to rectify the model reasoning provided through attention weight map such as shown in Figure 4.1. According to the theory of frame semantics, word meanings are defined relative to frames in a given text. Given the varied ways in which hate speeches are expressed, human, rather than machine, could quickly identify the appropriate frame to interpret a given text. Thus, if the model can learn to give attention to the elements of the correct frame as per human interpretation, the model performance could improve.

A semantic frame is a set of statements that give "characteristic features, attributes, and functions of a denotatum (data object), and its characteristic interactions with things necessarily or typically associated with it." [13] Moreover, a semantic frame can be viewed as a coherent group of concepts such that complete knowledge of one of them requires knowledge of all of them [195]. Therefore, it provides a common representation to capture both knowledge and meaning of a given textual post. For example, a description of frame in FrameNet[2] (the popular knowledge base to understand human language) primarily contains following attributes: *Description* - a textual description of the frame including what it represents; *Frame Elements (FE)* - additional attributes for representing meaning of the frame in a sentence/context, such as the frame $Being\_born$ has FEs: Child, Time, Place, etc.; *Lexical Units (LU)*: the lemmatized form of words with their part-of-speech that invoke a frame; and lastly, *Example Sentences*.

Hate speech can be described by several frames [62], and thus, a common but essential pattern to guide the human feedback at the token-level could be the element of hate target group in a given text. To achieve this goal we propose a model described in the next subsection that can combine human feedback of the hate target interactively during the training process, within the specific language context of a textual post.

[2]https://framenet.icsi.berkeley.edu/fndrupal/

4.4 Experiment setup

Data: We evaluated our model on the dataset containing English and Spanish tweets provided by SemEval-2019 Task 5 — HatEval: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter [21]. This dataset has been labeled with two classes for determining whether a given tweet is hateful or not-hateful for a given target such as women or immigrants. Additionally, the data includes labels for identifying the harassed target as individual or generic (i.e. individual or group). Table 4.2 shows the quantity of training and test instances for each category. In the current work, we named the first type of labeling as hate speech labels, and the second type as target-type labels.

Table 4.2: Train and test data.

		Hate speech		Target type	
		Hate	Non-hateful	Individual	Group
Train	English	3783	5217	1341	2442
	Spanish	1857	2643	1129	728
Test	English	427	573	219	208
	Spanish	222	278	137	85

Schemes: To compare our proposed method, given there is not exactly comparable prior work for our multilingual problem setup tested on the same dataset, we construct multiple baseline schemes using classical machine learning models [178] that use LASER embeddings as input features. We did not use state-of-the-art models proposed during SemEval-2019, because participants had another task and tested their approaches on single-language data, while our setup includes only multilingual cases. Additionally, we also compare MLIAN with LSTM [154] based model, as used in the prior works (we use it with two hidden levels and trained for 20 epochs, similar to MLIAN). We evaluate our MLIAN model with both LASER and DistilmBERT embeddings. The full list of proposed modeling schemes for evaluation is the following (* denotes our proposed models and others are the baselines):

- **[*SVC+LASER*]**: This method uses pre-trained LASER embeddings, which are passed as input to a Linear Support Vector Classifier model.

- **[*RF+LASER*]**: This method uses pre-trained LASER embeddings, which are passed as input to a Random Forest model.

- **[*SGD+LASER*]**: This method uses pre-trained LASER embeddings, which are passed as input to a Stochastic Gradient Descent model.

- **[*MLP+LASER*]**: This method uses pre-trained LASER embeddings, which are passed as input to a Multi-Layer Perceptron (MLP) model.

- **[*LSTM+LASER*]**: This method uses pre-trained LASER embeddings, which are passed as input to a LSTM Network model.

- **[*LSTM+DistilmBERT*]**: This method uses pre-trained DistilmBERT embeddings, which are passed as input to a LSTM Network model.

- **[******MLIAN+LASER*]**: Multilingual Interactive Attention Network (MLIAN) method with LASER embeddings.

- **[******MLIAN+DistilmBERT*]**: Multilingual Interactive Attention Network (MLIAN) method with DistilmBERT embeddings.

To evaluate the performance of classification models, we adopt three metrics: Accuracy (ACC), Area Under the Receiver Operating Characteristic Curve (AUC), and weighted F-measure (F1), which is consistent with the prior works on hate speech detection.

Model Implementation: In MLIAN, we need to optimize all the parameters in LSTM networks: the attention layers, the softmax layer, and the text embeddings (LASER or DistilmBERT). Cross entropy with L2 regularization is used as the loss function. We use backpropagation to compute the gradients and update all the parameters of LSTM. The coefficient of L2 normalization in the objective function is set to 10^{-3}, the dropout rate is set to 0.2, and 20 epochs.

4.5 Result Analysis and Discussion

We first discuss the results of MLIAN model against the baseline schemes, followed by an in-depth analysis of the nature of human feedback, the amount of human feedback, and the analysis of cross-lingual and cross-target scenarios.

4.5.1 MLIAN Performance

Table 4.3 shows the performance comparison of *MLIAN*-based models with other baselines. We can observe that the deep learning models have higher performance in multilingual hate speech detection than classical machine learning based model schemes. Further, both the proposed models of *MLIAN+DistilmBERT* and *MLIAN+LASER* show higher performance than LSTM baselines in all metrics. *MLIAN+LASER* model scheme demonstrates better performance result in multilingual cases, which is perhaps contributed by the consideration of sentence-level context by the LASER embeddings and the human feedback in MLIAN.

Table 4.3: Comparison with baselines. Results of binary classification for the SemEval-2019 Task 5 (hate speech against immigrants and women). Best performances are in bold. Models were trained on multilingual data (10-fold CV). $*$ denotes the proposed models.

Model Scheme	ACC	AUC	F1
MLP+LASER	60.93±0.72	58.73±0.69	59.85±0.64
RF+LASER	70.20±0.61	67.63±0.53	68.85±0.62
SGD+LASER	69.38±0.43	70.07±0.99	70.07±0.52
SVC+LASER	71.87±0.71	71.27±0.45	71.85±0.63
LSTM+DistilmBERT	73.85±0.31	78.29±0.14	73.86±0.31
LSTM+LASER	71.59±0.43	79.38±0.11	71.58±0.43
**MLIAN+DistilmBERT*	81.24±0.59	79.84±0.61	81.00±0.55
**MLIAN+LASER*	**85.06±0.40**	**84.14±0.54**	**84.94±0.42**

Further, our *MLIAN* models demonstrate that emphasizing the im-

portance of human feedback through learning target and context representation interactively can be valuable for hate speech detection tasks. Compared with *LSTM* models, our architecture improves AUC performance by about 6% for LASER-based model implementation and 2% for DistilmBERT-based model implementation on multilingual data. The main reason for higher performance is that *MLIAN+DistilmBERT* and *MLIAN+LASER* use the additional feature contributed by the simulated human agent which can influence the learning of context in the attention network. Besides higher performance, in this design, we can learn the representations of targets and contexts whose collocation contributes to hate speech detection even in the posts where users resort to special language subtleties. This also inspires our future work to explore and research the semantics of a variety of target types.

4.5.2 Analysis of Frame Semantics Theory-based Human Feedback

In this analysis, we tested the theoretical justification of the impact of human feedback based on the connection between frame semantics theory described in Section 4.3.4 and attention weight maps resulting from the developed MLIAN model.

Specifically, to understand how MLIAN attention weights correlate with semantic frames, we examine the tweet text originally written in Spanish and its English translation. First, we can extract semantic frames for the text versions in both languages by employing a semantic parser. Semantic parsers are trained specifically to consider context when identifying frames in a text. In this exercise, we utilized SLING [195] to extract head frames from a textual post. Head frames are the frames directly evoked by a mention in text. Figure 4.3 illustrates the high-level frame extraction process with an example.

Second, we apply the trained MLIAN model with and without human feedback data on the two versions of the text to retrieve the attention weight maps. For this task, we used *MLIAN+DistilmBERT* model because it allows to extract word embeddings (while *MLIAN+LASER* ex-

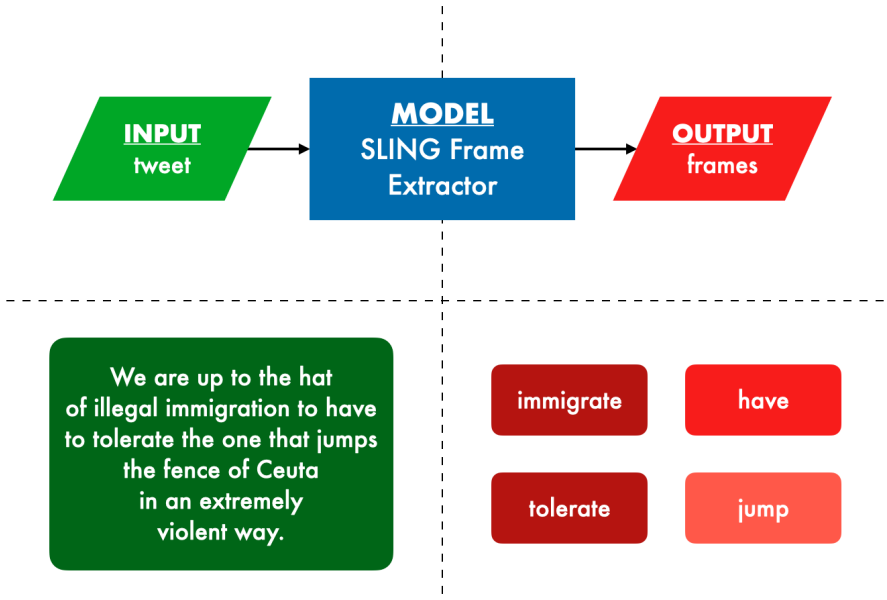


Figure 4.3: High Level Architecture of Frame Extraction Process.

tract only sentence embeddings that cannot be interpretable by human.) Figure 4.4 demonstrates that the attention weights for the MLIAN model with human feedback-based hate targets have more correlations with frames than the model without such principled feedback. Moreover, it is important to note that this correlation is observed for both languages, indicating the significance of relying on a principled approach of frame semantics theory to identify the type of targets to receive the human feedback.

4.5.3 Analysis of the Impact of Human Feedback

To measure the minimal required human feedback that could impact the MLIAN model performance, we start with a baseline model scheme without considering the target-types labels and then, design several model schemes that incrementally add a specific amount of the target-types labels as human feedbacks. Specifically, we analyze the baseline case against

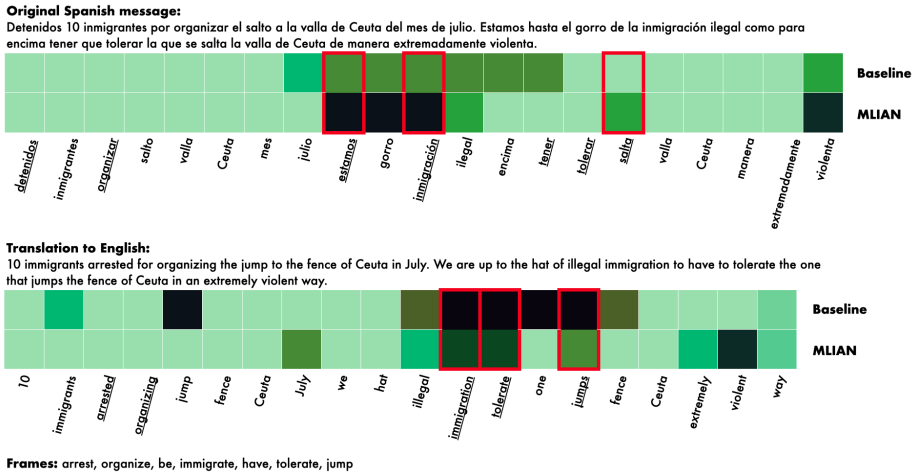


Figure 4.4: Attention weight maps of the original and translated texts in English and Spanish. The darker color of the word means the higher weight. Cells with red borders indicates words that match with semantic frames.

three schemes, with the gradual addition of randomly selected target-type labels – *100, 500, 1000*. For this task, we use the performance measures of ACC, F1, and AUC for assessing different model schemes.

The full set of results are presented in Table 4.4. Results show that the statistically significant improvement of accuracy and AUC was noticeable when increasing the number of target-type labels to just 500, which equals to approximately 4% of training data only. This analysis demonstrates how even the minimal use of human feedback could result into faster convergence of the model training, for better performance in the hate speech detection task.

Table 4.4: Minimal human impact in *MLIAN+LASER* model is studied for the model schemes with an increasing number of human feedback. The table shows p-value for each scheme's performance comparison with the baseline and the previous scheme. The best performing scheme's p-value is bold.

	baseline	100 feedbacks	500 feedbacks	1000 feedbacks
ACC	73.04	74.10	77.46	78.59
Compare:				
- baseline		0.297	0.001	0.000
- scheme		0.297	**0.028**	0.387
AUC	73.03	74.26	77.02	78.47
Compare:				
- baseline		0.206	0.002	0.000
- scheme		0.206	**0.060**	0.296
F1	73.15	74.16	77.44	78.62
Compare:				
- baseline		0.318	0.001	0.000
- scheme		0.318	**0.031**	0.368

4.5.4 Analysis of Cross-Lingual and Cross-Target Classification

Multilingual classification tasks also include cross-lingual classification settings — when languages in training and testing data are different.

For evaluation of cross-lingual capability of the proposed method, we train MLIAN model on one language and test on another, for English and Spanish. The full results of cross-lingual classification are presented in Table 4.5. The MLIAN models for both languages show better results comparing with LSTM baselines. The model achieves up to 68% AUC for training on English and testing on Spanish, while for the opposite scenario, it reaches up to 78% AUC. On the other hand, the LSTM-based baseline models show only 65% and 68% AUC for the two scenarios.

Table 4.5: Results of cross-lingual classification. Best performances are in bold (10 fold CV). $*$ denotes the proposed models.

Model Scheme	ACC	AUC	F1
EN $\rightarrow$ ES			
LSTM+LASER	61.90	65.67	61.74
LSTM+DistilmBERT	59.16	65.01	58.77
MLIAN+DistilmBERT*	71.33	**68.16**	69.78
MLIAN+LASER*	**71.46**	**68.16**	**69.91**
ES $\rightarrow$ EN			
LSTM+LASER	58.16	63.51	57.61
LSTM+DistilmBERT	63.45	68.56	63.13
MLIAN+DistilmBERT*	**81.76**	**78.95**	**81.01**
MLIAN+LASER*	81.28	78.57	80.57

Lastly, we evaluate DistilmBERT for cross-topic hate speech detection — when the type of targets in the training and testing data are different. For evaluation of cross-target capability of the model, we train the model on the dataset with hate speech targeted to one group of people (e.g. migrants) and tested on the dataset with another targeted group (e.g. women). Example of hate speech targeted at different groups are presented in Table 4.1. The full results of cross-topic classification is presented in Table 4.6. MLIAN models show the best results comparing with another baseline model schemes. MLIAN model reaches up to 80% AUC for training on migrants-as-target dataset and testing on women-as-target dataset, and in the opposite scenario, such a model reaches up to 82% AUC. In contrast, LSTM baselines were only able to achieve up to 74% AUC in both scenarios.

These results show that LSTM could reach good performance during both cross-lingual and cross-topic hate speech classification tasks and this analysis validates the benefits of deep learning model with human feedback for improving the task performance.

Table 4.6: Results of cross-target classification. Best performances are in bold (10 fold CV). $*$ denotes the proposed models.

Model Scheme	ACC	AUC	F1
migrants $\rightarrow$ women			
LSTM+LASER	68.16	74.35	68.16
LSTM+DistilmBERT	68.31	74.46	68.31
MLIAN+DistilmBERT*	71.41	69.37	71.05
MLIAN+LASER*	**81.89**	**80.31**	**81.51**
women $\rightarrow$ migrants			
LSTM+LASER	67.17	73.33	67.17
LSTM+DistilmBERT	67.23	74.12	67.23
MLIAN+DistilmBERT*	**84.06**	**82.70**	**83.80**
MLIAN+LASER*	70.40	77.25	69.39

4.6 Conclusion

In this work, we design a Multilingual Interactive Attention Network model for hate speech detection in social media posts, regardless of language. The core idea of MLIAN is to use two attention networks to model the context of content and the special tokens as targets interactively, where we employ the frame semantics theory to design a principled approach for appropriately guiding the human feedback to provide target labels. We use simulated human feedback by labeling posts that contain personal/group hate for identifying the special tokens as target labels. The model pays close attention to such important parts in the context and learns to give higher attention to the potential elements of the semantic frame characterizing the hate speech in the post. Experiments on SemEval-2019 Task 5 dataset demonstrate that MLIAN model performs better than several baselines and requires a minimal human feedback effort for improving the model performance. We present extensive analyses to show the value of modeling with human feedback, which can help adapt the model to different languages and tasks easily. The application of MLIAN model can inform future studies for multilingual hate

speech analytics.

4.6.1 Reproducibility

Datasets and code for the experiments described in this chapter are available for research purposes in a public repository.[3]

[3] https://github.com/vitiugin/mlian

Chapter 5

CROSS-LINGUAL INFORMATION EXTRACTION AND SUMMARIZATION

This chapter introduces a cross-lingual method for retrieving and summarizing crisis-relevant information from social media postings. The extraction of timely and relevant information from social media during crises poses significant challenges, especially when dealing with multiple languages. We present a flexible approach that utilizes multilingual transformer embeddings to enable the formulation of structured queries and the generation of comprehensive summaries. Our method allows experts to write queries in their preferred language, while the extracted sentences can be in any supported language. This approach facilitates the efficient processing of large volumes of information with human expertise in the decision-making process.

5.1 Introduction

Social media platforms such as Twitter are widely used to share information during disasters and mass convergence events [35]. During these situations, users, including eyewitnesses, media, governmental and non-profit organizations, post an enormous volume of diverse content, from personal opinions and commentary to reports and messages providing relevant information that could lead to better situational awareness. This chapter describes an approach to automatically summarize information posted in social media about an event, creating brief reports to help emergency response and recovery. These reports can help emergency managers better understand a developing situation and plan the following actions accordingly [52, 102].

The development of methods that automatically extract crisis-relevant information from social media has been an active line of work for many years [106]. Traditionally, crisis information extraction methods use linguistic and semantic resources mainly concentrated on one language [203]. However, there are many cases where a single crisis affects several countries or regions that speak different languages [236, 79, 147], or affects a region where the population speaks more than one language.

Previous work has shown that the information provided by social media postings is related to the language in which they are posted, and indeed messages in different languages about the same crisis often provide complementary information [238, 149]. Extracting and summarizing information from social media in only one language introduces the risk of missing valuable information. However, creating or adapting language-specific resources or methodologies for new languages is expensive and time-consuming. Therefore, current crisis informatics solutions need effective cross-lingual tools for extracting relevant information about appropriate categories of crisis-relevant information.

Our main contributions are:

- We describe a flexible, query-based, cross-lingual method for collecting from social media relevant postings in multiple languages about specific information categories. The method uses pre-trained

multilingual sentence embeddings (LASER [17]) to extract postings from a general collection of crisis-related messages.

- We describe an approach for crisis summarization that takes as input relevant postings about an information category and generates a summary using a transformer-based language model (Text to Text Transfer Transformer (T5) [192]). We use clustering and diversification operations to create less redundant, more information-rich summaries.

We perform empirical validation using both crowdsourcing annotators and emergency management experts and release a new annotated dataset to evaluate multilingual crisis informatics systems.

The remainder of this chapter includes a presentation of related work, followed by a description of the query-based method for crisis information extraction and cross-lingual classification and summarization. Next, we describe our experimental setupand the results of our analysis. Finally, we present our conclusions and envisioned future work.

5.2 Related Work

Mining the social web for crisis-relevant information has been an active and fruitful research topic for many years. Our coverage of it focuses on overviewing methods for mining,classification,and summarizationof crisis-relevant social media messages.

5.2.1 Mining Social Media for During Crises

Social media is a key communication channel during all kinds of crises, including natural and man-made disasters. Computational methods from many disciplines can contribute to creating mining and retrieval systems that can help emergency managers [35]. Crisis-related social spans many different categories of information, including timely messages about urgent needs from affected populations and damaged infrastructure such as

bridges or roads. Together, this information is relevant for emergency response, recovery management, and assessments of the costs of damages [117] Unfortunately, most methods for mining social media during disasters described in the extensive literature on the topic are monolingual, limiting the applicability in countries using languages other than English or even in English-speaking countries with increasing multilingual urban populations [149]. The response during the disasters could be significantly improved with the ability to employ social data mining methods on user-generated data across multiple languages [238]. Cross-lingual and multilingual classification and summarization methods provide an opportunity to gather complementary information across various languages spoken in affected areas.

5.2.2 Classification of Crisis-Related Messages

In the recent literature on this topic, "traditional" supervised learning methods such as Naive Bayes and SVM coexist with neural-network-based methods [214]. Indeed, SVM for the classification crisis-relevant social media has consistently shown to exhibit high performance, especially when combined with semantic features computed with the help of external knowledge bases [122].

Deep learning methods using various architectures have proven effective at detecting crisis-relevant messages; a popular architecture is CNN using word embeddings [172, 147]. The addition of information specific to an event type, such as hydrological information in the case of floods, has been shown to improve classification performance [53]. A particularly influential model has been BERT [191, 59], which is currently being used for various challenging NLP tasks, including classification. Recent papers on this topic describe end-to-end transformer-based models for crisis classification tasks, demonstrating promising results [145, 135].

LASER is an architecture to learn joint multilingual sentence representations for 93 languages. The system uses a single BiLSTM encoder with a shared byte-pair encoding vocabulary for all languages, coupled with an auxiliary decoder, and trained on publicly available parallel cor-

pora. The resulting embeddings are computed using English annotated data only and transferred to any of the 93 languages without any modification [17]. LASER embeddings have also been shown to be effective in multilingual classification tasks [183, 43].

5.2.3 Crisis-Related Information Summarization

Social media messages are usually short and thus tend to provide fragmented information hence consolidating and summarizing information is key [204, 205]. An informative summary can help stakeholders gain situational awareness and manage critical resources effectively [247].

The main approaches used for text summarization can be categorized as either extractive or abstractive [169]. *Extractive* approaches construct summaries by combining selected informative phrases or sometimes whole sentences from the source text [65]. *Abstractive* summarization, on the other hand, generates summaries from a representation of the semantics of a given text; an abstractive summary may contain words or sentences that do not appear in the source document(s). Abstractive summarization techniques usually employ a generative approach [142, 134].

Despite the benefits of abstractive approaches, extractive approaches are still considered state-of-the-art for summarization due to their simplicity and high performance [196, 113]. However, extractive approaches often fail to include key elements useful in a report, such as answers to "what," "who," "where," "when," and "how" questions. These are important elements in the domain of disaster and crisis management and need to be concisely incorporated into summaries [130]. Query-based approaches have been described as a helpful manner of incorporating this information to improve the quality of reports [197]. In general, abstractive methods may facilitate the generation of more informative summaries, not restricted to sentences that directly take sequences of words from the source text [171]. State-of-the-art abstractive summarization methods tend to adopt transformers and pre-trained models that have demonstrated great performance in other NLP tasks: BART [40], T5 [94], Pre-training with Extracted Gap-sentences for Abstractive Summarization (PEGASUS) [207].

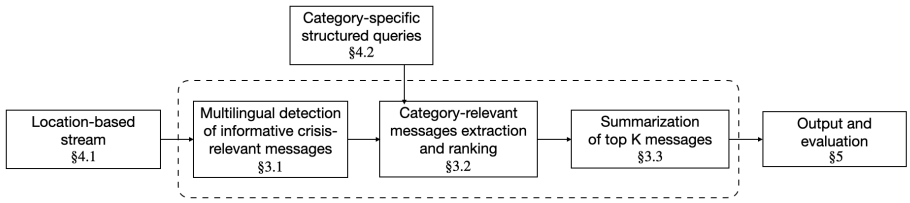


Figure 5.1: Overview of our cross-lingual information summarization framework.

Our research builds upon previous work and contains two key innovations. To the best of our knowledge, (1) we are the first to describe a summarization method that retrieves crisis-relevant information using a query-based approach, and (2) we are the first to propose a transformer-based summarization model for crisis-related messages.

5.3 Method overview

In this section, we provide an overview of the proposed method, named Cross-LIngual Query-based Summarization of Crisis Messages (CLiQS-CM). An overview of the method is shown in Figure 5.1. First, an automatic classification model is used for detecting crisis-relevant, informative messages. Second, cross-lingual ranking is performed on these messages. Third, the top k ranked messages are given as input to a summarization model.

5.3.1 Classification Model

A large fraction of messages posted in social media in response to a crisis event doesn't include any informative claims beyond merely announcing that a crisis situation is developing. Hence, a key step is detecting crisis-relevant *informative* messages. We model this as a binary classification task and create an automated classification model that we name Cross-

LIngual Query-based Classification of Crisis Messages (CLiQC-CM).

Pre-processing. Messages are preprocessed by replacing URLs and account mentions ("@user") by specific tokens and turning hashtags into words. We preserve punctuation and stopwords, as in the next steps, we use sentence embeddings and dependency parsing.

Feature extraction. We include morphological and syntactical features. Using the Stanza Part Of Speech (POS) tagger [189], we count the number of numerals/numbers, nouns, verbs, adverbs, and adjectives in the messages. Using the Stanza dependency parser, we extract and count syntactic features indicating the presence/absence of claim-containing sentences, such as subjective nouns, compounds, roots, and modality. For both tasks, we use pre-trained models in each of the languages we work with. We normalize the count of occurrences of each type of element in a message, such as "contains N numerals/numbers," "contains N subjective nouns," and so on, using min-max scaling to be in $[0, 1]$. We also consider Named Entity Recognition (NER) features extracted using the SpaCy library [99]. These are binary features indicating whether a message contains persons' names, the name of a place, organization, or a date. We use SpaCy's pre-trained models for each of the languages we work with. Off-the-shelf, SpaCy supports 15 languages, including all the ones we work with, except Croatian and Tagalog. For Croatian, we use a contributed model for the Stanza package; for Tagalog, we use an open-source pre-trained model.[1] Finally, we include *message-specific features* indicating the (min-max scaled) number of URLs and user mentions in messages.

Embeddings. For representing the input data, we used sentence embeddings generated by the pre-trained transformer-based model, LASER [17]. LASER sentence-level embeddings are obtained by applying max-pooling over the output of BiLSTM sentence encoder. The BiLSTM output is constructed by concatenating the outputs of two LSTMs working in opposite directions (forward and backward). The bidirectional encoder captures more contextual information than a single-direction LSTM en-

[1] https://github.com/matthewgo/FilipinoStanfordPOSTagger

Table 5.1: Values of hyperparameters.

Hyperparameter	Text features	LASER embeddings	Similarity features
LSTM layers	-	1	-
MLP layers	2	3	2
MLP neurons	128;24	1024;256;128	128;24
Dropout	-	0.5	-
Activation	relu	sigmoid	relu

coder (e.g., a left-to-right one). In our experiments, we used LASER to embed all tweet sentences into fixed-size vectors of length 1,024.

Architecture. In our classification architecture, the embeddings are passed to an LSTM-layer and then combined with additional features. This architecture is inspired by one proposed for the detection of fake news articles [25], which has also been used for emotion detection [237].

The proposed scheme, depicted in Figure 5.2 (minus the query-based features, which are only used by the ranking step), computes the feature vectors separately and then combines them with the help of MLP layer. We use binary cross-entropy as the loss function to optimize and include a soft-max layer to classify social media text into one of two classes ("crisis relevant" or "not crisis relevant"). The hyperparameter settings of the feature extractor portions are shown in Table 5.1. The feature combination layer uses the softmax activation function with Adam optimizer, the learning rate of 0.001, batch size of 100, and binary cross-entropy loss.

5.3.2 Cross-lingual Ranking Model

The next step in our method is to retrieve, from the informative crisis-related messages, a series of messages that are relevant for various informational categories. Specifically, we retrieve and rank the top-k most relevant messages from each category to pass them to the summarization model.

To make information extraction more adaptable to different needs of emergency managers, our method is based on structured queries. Each

query is related to a specific information need and contains keywords, templates, and prototypes; a sample query is found in Table 5.4. *Keywords* are words used frequently in messages of a category; *templates* are fragments containing key crisis-relevant facts; and *examples* are entire sentences or even entire messages corresponding to each category.

Each query is written in one language (English in our case) and used to extract information across all languages. Queries are, for the most part, agnostic of the type of event, but in some cases, they may include elements that are specific to a type of disaster, for instance, in the case of earthquakes, their magnitude, or in the case of storms, rainfall. We remark that context-based semantics allow our system to work even if these elements provided by the user are not 100% complete (e.g., we can find messages containing related keywords or messages with similar semantics to the examples provided but using different wording).

To calculate query similarity features, we measure average and maximum cosine similarity between the query's keywords, templates, prototypes, and each message. As a result, we have six similarity features. For ranking messages, we use basically the same architecture as in the previous step (S5.3.1), with the addition of query similarity features. This is depicted in Figure 5.2. After removing duplicates, we pass the top 100 candidates to the summarization step. We tested with the top 20, 50, and 100 candidates, and observed that the top 100 provided the highest recall.

5.3.3 Summarization Model

The final component of our method creates a category-specific summary from the retrieved messages for each category. These summaries are created by T5, a pre-trained[2] transformer model widely used for summarization tasks [192]. In our preliminary experiments, this model performed better than a similarly pre-trained BART-based model.

We tested two different configurations for the summarization model: a regular condition and a diversified condition. In the *regular* condition, we gave T5 as input the texts from the top-100 most relevant candidates

[2]HuggingFace - https://huggingface.co

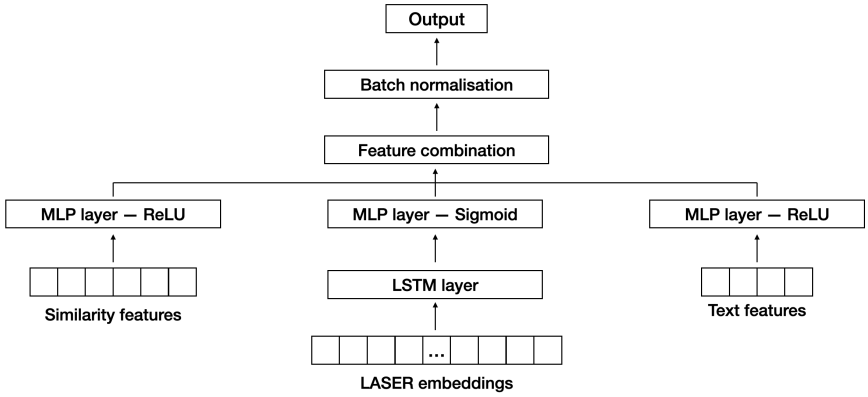


Figure 5.2: Combining the transformer embeddings with morphological, syntactic, message-specific features and query similarity features using deep MLP

and limited the length of the output text. In the *diversified* condition, we clustered the top-100 most relevant candidates and gave to T5 as input the texts of all the messages on each cluster, one cluster at a time; the resulting summary is the concatenation of the per-cluster summaries. For the diversification step, we first automatically find an appropriate number of clusters using the Silhouette Score. Next, we cluster texts by the K-Means method. Because we want to keep all summaries comparable in length for our experiments, we set a maximum number of clusters to four. Also, during experiments, we found that the heuristic of summarizing clusters in decreasing order by size (i.e., starting with the largest cluster) helps to generate more relevant summaries. This is akin to following the "inverted pyramid" style typically used in journalism.

5.4 Experimental Setup and Evaluation

In this section, we describe our data collection (§5.4.1) and the queries used to retrieve relevant messages in various information categories (§5.4.2). Then, we describe the baselines for classification (§5.4.3) and summariza-

Table 5.2: Number of annotated messages for each event, including total number of messages, and number of messages labeled as informative by a human annotator. Local languages appear in bold.

		lang. 1 (en)		lang. 2 (es)		lang. 3 (fr)		language 4		language 5	
	dates	total	info	total	info	total	info	total	info	total	info
Australia bushfires	06-31.01.2020	**2000**	**233**	2000	435	2000	460	2000	285 (ja)	2000	167 (id)
Fukushima earthquake	13.02.2021	2000	266	2000	529	2000	227	**3000**	**101** (ja)	3000	153 (id)
Gloria storm	17-25.01.2020	703	393	**571**	**210**	**517**	**168**	**542**	**233** (ca)	-	-
Taal eruption	12-17.01.2020	551	123	691	202	610	114	**1500**	**258** (tl)	458	151 (pt)
Zagreb earthquake	22-24.03.2020	537	162	509	243	520	187	**1500**	**282** (hr)	542	163 (de)

tion (§5.4.4), and the evaluation metrics used to compare the proposed method against the baselines (§5.4.5).

5.4.1 Multilingual Data Collection

Our data collection followed standard practices to collect crisis-related social media messages from Twitter. We collected public tweets using Twitter's public API, filtering by location-related keywords and date, without using any additional filtering (e.g., we did not restrict the query to specific languages). We considered five disaster events between January 2020 and February 2021 that received substantial news coverage internationally:

- *Australian bushfires* (2019-2020): period of bushfires in many parts of Australia, which, due to its unusual intensity, size, duration, and uncontrollable dimension, was considered a "megafire";[3]
- *Fukushima earthquake* (February 2021): a 7.1 M_w or 7.3 M_{JMA} earthquake that struck offshore east of Tōhoku, Japan and caused significant structural damage across the Tōhoku and Kanto regions;[4]
- *Gloria storm* (January 2020): a Mediterranean storm that affected eastern Spain and southern France with high winds and heavy rainfall;[5]

[3]https://en.wikipedia.org/wiki/2019–20_Australian_bushfire_season
[4]https://en.wikipedia.org/wiki/2021_Fukushima_earthquake
[5]https://en.wikipedia.org/wiki/Storm_Gloria

- *Taal volcano eruption* (January 2020): a phreatomagmatic eruption from its main crater that spewed ashes across Calabarzon, Metro Manila, and some parts of Central Luzon and the Ilocos Region in the Philippines, resulting in temporary closures of schools and workplaces, and disruptions of flights in the area;[6]

- *Zagreb earthquake* (March 2020): an earthquake of magnitude 5.3 M_w, 5.5 M_L, which hit the capital of Croatia, causing severe damage to hundreds of buildings in its historical center.[7]

All messages include a "language" field computed by Twitter using a language detection model developed specifically for tweets. We counted the number of messages per language in each event. Three of the top languages were common to all of the studied events: English (ISO 639-1 code: en), Spanish (es), and French (fr). Additionally, we found several hundred messages for each event in other languages, including Catalan (ca), Tagalog (tl), Croatian (hr), German (de), Japanese (ja), Indonesian (id), and Portuguese (pt). After collecting the data, we labelled tweets or their translation to English that contained potentially informative factual information. We name this group of tweets "informative messages." One of the authors created the ground truth by reviewing each event and hand labeling these tweets. Another author reviewed a portion of the classified tweets, and adjustments to the classification task were agreed upon when needed. Additionally, "informative messages" were reviewed by crowdworkers during the categorization task and excluded if they did not contain information related to any category. The number of annotated messages is shown in Table 5.2.

Next, we used crowdsourcing to further categorize the messages into various informational categories. Specifically, we employed crowdworkers through a crowdsourcing platform,[8] paying the standard rate recommended by the platform. We asked three different workers to label each of the approximately 5,700 informative messages across languages. The

[6]https://en.wikipedia.org/wiki/2020_Taal_Volcano_eruption

[7]https://en.wikipedia.org/wiki/2020_Zagreb_earthquake

[8]SurgeHQ - https://www.surgehq.ai

Table 5.3: Categories for multilingual information extraction, based on the ontology from TREC-IS 2018 [159]. Example messages have been paraphrased for anonymity.

Category	Description	Example message
Casualties	Affected or injured people	*Around 150 injured people*
Damage	Built or natural environment damage	*Destroying orange trees and rice paddies*
Danger	Messages of caution or alerts	*RED WARNING Barcelona - Danger to life*
Government	Official report by public agencies	*Local authorities continuing the search for ...*
Sensor	Seismic activity	*Zargeb hit by 5.3 magnitude earthquake*
Service	Providing a service or help	*Local org. provides shelter for more than 1,000 people*
Water	Water-related messages	*floods in Catalonia*
Weather	Weather updates	*heavy rainfall and flooding across region*

target categories were based on an ontology from TREC-IS 2018 [159], where we grouped some low-level ontology categories into higher-level ones. In total, we defined nine high-level classes of information, shown in Table 5.3.

5.4.2 Queries

We use a set of queries covering the nine information categories listed in Table 5.3. As described in in previous subsection, a query for an information category includes keywords, templates, and prototypes. Creating a query requires some degree of familiarity with social media messages posted during emergency situations.

Keywords are nouns and verbs usually present in messages containing a specific category of information. Practitioners could complete this task with scripts or programs to find frequent words or phrases present in previous collections of messages from past events. *Templates* are small frag-

ments of text describing crisis-relevant facts. The kind of information that we seek is in situation reports or in Wikipedia disaster-related *infoboxes*, which are templates that Wikipedia editors use to summarize crisis information. Users can provide such templates by copy-pasting passages from these sources, replacing the numbers or locations found there with the tokens `NUMBER` or `LOCATION`. Finally, users can provide *prototypes* – example messages or central passages typical in category-related texts, which can be obtained by sampling diverse, informative messages from past events. We envision a specialized user interface may assist users in formulating such queries, and we plan to explore that in future work. The scope of this work is to demonstrate the approach and provide an initial set of easily extended and refined queries. One such query is shown in Table 5.4.

Table 5.4: Example query. Each query includes keywords, templates, and prototypes.

Query for category: Weather	
keywords:	*snow, weather, rain, wind, coast, mph, kmh, forecast*
templates:	*batter parts of LOC, damages from winds, pummelling the region, NUMBER km/h winds, weather forecast, bad weather, heavy snow, strong wind, storm is hitting, wind gust*
prototypes:	*Wind, rain and snow batter parts of country; Storm brought around NUMBER m of snow and affect rivers; Heavy rainfall, strong wind and more than NUMBER of snow across LOC; Storm is hitting eastern LOC, with high winds and heavy rain; Storm has battered parts of LOC and reportedly brought worth of rain; Maximum gusts of wind in LOC NUMBER km / h; Tonight, terrible rains in LOC; Organisation has so far done NUM health care due to strong winds; Gusts of wind left fallen trees*

5.4.3 Message Classification Schemes

To compare our proposed method for informative messages detection, we construct baseline models using one classical machine learning scheme

(SVM) and one deep learning scheme (LSTM) that uses LASER embeddings as input features. Also, we compare our model with a cross-lingual LinearSVC-based model that uses semantic features extracted with the BabelNet knowledge base.[9] The complete list of proposed modeling schemes for evaluation is the following:

- **LASER+SVM**: this method uses pre-trained LASER embeddings; the embeddings are then classified by a Linear SVM model;
- **LASER+LSTM**: this method uses pre-trained LASER embeddings; the embeddings are then classified by a LSTM model;
- **Khare** [122]: this is a cross-lingual classification approach that uses additional semantic features extracted from external knowledge bases;
- **CrisisBERT** [145]: this is an end-to-end transformer-based model for crisis classification tasks (our implementation uses the DistilBERT [213] architecture);
- (ours): this is our method for classification, using a combination of LASER embeddings and tweet-related features. classified by a LASER model.

5.4.4 Summarization Methods

We compare our *CLiQS-CM* model and its diversified variation *Diversified Cross-LIngual Query-based Summarization of Crisis Messages (CLiQS-D-CM)* against several state-of-the-art summarization models. With the exception of the *LASER+LSTM+T5* method, all of the baselines use only category-related tweets as input, i.e., we simulate the best scenario in which the input is received from a perfect classifier. In our proposed models, we use the query-based model we described. The complete list of baselines for summarization that we used is the following:

[9] https://babelnet.org

- **LASER+LSTM+T5**: this method uses pre-trained LSTM embeddings, which are passed as input to a LSTM model for category classification and then to a T5 model for summarization;
- **C-SKIP** [196]: this is a centroid-based method using a FastText skipgram model trained on the CrisisLexT26 dataset [177], improved by the use of T5 pre-trained model (originally, the method used a corpus extracted from Google News);
- **CX_DB8** [197]: this is a queryable word-level unsupervised extractive summarizer, which is based on the text embedding framework Flair [10]. We tested this with different pre-trained embeddings, including transformer-based such as BERT and XLNet; for this task and datasets, the best results were obtained with Global Vectors for Word Representation (GLOVE) embeddings;
- **NAFI** [171]: this is an abstractive text summarization method developed specifically for crisis events;
- **CLiQS-CM** (ours): we use a combination of LASER embeddings with tweet-related features and query similarities features that are passed to a LSTM model for the ranking step and then uses a T5 model for the summarization step;
- **CLiQS-D-CM** (ours): this is the same as CLiQS-CM but retrieves diversified top-k candidates in the ranking step.

5.4.5 Evaluation Metrics

To evaluate the performance of the *classification* models, we use three standard metrics: ACC, AUC, and weighted F-measure (F1). These metrics are typically used in research on social media for emergency management (e.g., [122, 147, 238]).

To evaluate the *summarization* models, we considered four methods. First, we annotated all summaries for factual claims and then computed, for each summary, the fraction of factual claims it contained out of the

total factual claims mentioned across all summaries. Second, we computed the BERTScore [251] of each summary, which is a metric for evaluation of a text that compares them against a ground truth; in our case, an official report about the event. Third, we performed a crowdsourced evaluation of the readability of each summary across five dimensions: grammaticality, non-redundancy, referential clarity, focus, and structure and coherence [110]; five crowdsourcing workers were asked to compare summaries across each dimension. Fourth, we asked three experts in emergency management to perform a side-by-side comparison of the summaries and computed the number of times each summary was preferred.

5.5 Results

In this section, we present the results of our evaluation and comparison with state-of-the-art methods. First, we present an evaluation of our classification method. Next, we consider the extent to which summaries are comprehensive in terms of factual claims. Then, we ask crowdsourcing workers to evaluate the readability of summaries. Finally, we ask experts to perform a side-by-side comparison of the summaries.

5.5.1 Cross-lingual Classification

The first experiment is a "leave-one-language-out" evaluation: for each event, the classifier is trained on data from 3 or 4 languages and tested on the last language. What we simulate here is a scenario in which we have labeled data in several languages and extract information in a new language. Table 5.5 shows the performance comparison of our method *CLiQC-CM* with other baselines and state-of-the-art. We also perform a "leave-one-event-out" evaluation, in which we train on multilingual data for all events except one and test on the event that was left out. Results are shown on Table 5.6.

We can observe that in general methods based on multilingual trans-

formers perform better than the semantic-based model by *Khare*. However, there are a few differences between schemes based on LASER embeddings; SVM and LSTM achieve in general the best performance, with some variations across datasets. The performance of *CrisisBERT* is comparable to that of the proposed method *CLiQC-CM* in some cases and across some metrics, but the average performance of *CLiQC-CM* is better. We also observe that across all methods, "leave-one-event-out" seems to pose a more difficult problem than "leave-one-language-out." This suggests that most of the multilingual methods we tested capture fairly well event-specific concepts (such as specific places, impacts, or needs) of each crisis but do not generalize so well across events.

Table 5.5: Results of cross-validation evaluation of message classification across languages ("leave-one-language-out"), with 4-5 languages per event: the test data contains all of the messages for an event in one language, while the training data contains messages in other language for the same event.

	Australia bushfires			Fukushima earthquake			Gloria storm			Taal eruption			Zagreb earthquake			Average		
Schemes	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
LASER+SVM	92.3	91.1	70.7	97.4	**97.3**	85.4	83.6	81.6	66.6	95.4	**94.4**	77.9	92.9	**92.9**	90.6	92.3	91.5	78.2
LASER+LSTM	91.9	74.6	92.7	97.5	92.7	99.5	83.2	66.9	86.2	94.5	84.3	96.7	93.7	83.0	97.2	92.2	80.3	94.4
Khare	88.8	85.8	64.1	88.1	85.3	65.9	56.3	50.1	70.5	90.6	86.1	50.0	41.6	30.2	33.3	73.1	67.5	58.0
CrisisBERT	91.4	87.1	96.2	**97.6**	96.6	**99.4**	87.7	83.4	94.2	95.3	92.6	98.0	**94.1**	90.6	98.3	93.2	90.1	97.2
***CLiQC-CM**	**95.9**	**93.6**	**99.2**	**97.6**	96.3	**99.4**	**93.0**	**90.9**	**97.8**	**95.6**	93.3	**98.1**	93.2	88.5	**98.6**	**95.1**	**92.5**	**98.6**

Table 5.6: Results of cross-validation evaluation of message classification across events ("leave-one-event-out"): the test data contains all of the messages for one event, while the training data contains messages from all of the other events.

	Australia bushfires			Fukushima earthquake			Gloria storm			Taal eruption			Zagreb earthquake			Average		
Schemes	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
+SVM	**90.1**	87.8	63.0	96.4	**96.2**	84.8	**87.3**	**85.8**	71.4	93.1	91.6	65.6	85.7	84.0	74.0	90.5	89.1	71.8
LASER+LSTM	87.5	66.4	87.5	96.6	83.4	99.0	80.3	63.5	80.3	93.6	81.4	96.9	84.2	75.4	92.5	88.4	74.0	91.2
Khare	86.2	83.3	54.8	93.4	93.4	81.9	80.8	77.1	57.9	90.2	88.4	58.8	80.6	77.4	65.9	86.2	83.9	63.9
CrisisBERT	88.0	88.0	93.3	**96.7**	95.9	**99.3**	83.3	82.2	91.7	**94.6**	93.4	**98.0**	87.0	82.1	94.4	89.9	88.3	95.3
***CLiQC-CM**	89.1	**88.2**	**94.0**	96.1	95.9	98.4	86.4	84.8	**93.3**	93.7	**93.5**	96.3	**88.4**	**84.4**	**95.0**	**90.7**	**89.4**	**95.4**

5.5.2 Recall of Factual Claims

One way of measuring how informative are different summaries are, is to consider the extent to which they contain factual claims related to an information category for an event. To perform this evaluation, we manually coded each category-related factual claim in each of the generated summaries across all methods and then counted the number of claims in every summary compared to the overall claims.[10] The fraction of claims contained in summary is divided by the total number of claims across all summaries if what we call the *recall of factual claims.* Table 5.7 shows the performance comparison of *CLiQS-CM* with other baselines and state-of-the-art. *CLiQS-CM* and *C-SKIP* outperform the other methods in both the cross-lingual and English-only evaluation, with a small advantage for *CLiQS-CM*.

The diversified method *CLiQS-D-CM* produces summaries with less factual claims than *CLiQS-CM*; in our observations, this is partially explained by diversification leading to more low-ranking claims to be included, and those claims are more likely to be incorrectly associated to the category under analysis. In other words, the lower we go on the list of retrieved messages for a category, the more likely we are to find messages that actually belong to other categories. As we explain in subsection 5.3, we create clusters from top-100 candidates, and often the number of category-related candidates is much less than 100; in this case, clusters other than the first one are likely to be non-related to a category.

5.5.3 BERTScore: Similarity with Official Reports

To perform this evaluation, we retrieve summaries for each event prepared by the Emergency Response Coordination Centre (ERCC).[11] These summaries are created at the level of entire events and not divided by category. Hence, we create an event-level summary using each method by combining the summaries from all categories. Each event-level summary

[10] These annotated summaries are part of our data release.

[11] ERCC Portal - https://erccportal.jrc.ec.europa.eu/

Table 5.7: Factual claims present in each summary, on average, as a percentage of the total number of factual claims across all summaries for a category and event. We consider cross-lingual and English-only evaluations. The top two methods are extractive, while the remaining four are abstractive; our methods are marked with an asterisk.

Scheme	Cross-lingual	English
C-SKIP	28.8%	29.0%
CD_DB8	24.1%	17.8%
LASER+LSTM+T5	10.0%	16.0%
Nafi	25.1%	28.8%
***CLiQS-CM**	**29.4%**	**31.2%**
***CLiQS-D-CM**	23.8%	23.5%

is compared against the ERCC one using BERTScore.

Table 5.8 shows the results of the performance comparison of *CLiQS-CM* against other models. Both proposed methods *CLiQS-CM* and *CLiQS-D-CM* show better performance than the baselines in all datasets except one. The exception is *C-SKIP*, a centroid-based extractive method, which demonstrates higher performance for one of the analyzed events (Zagreb earthquake). The BERTScore evaluation also helps us interpret the results regarding the recall of factual claims, as often the precision of *CLiQS-CM* is higher than the one of *CLiQS-D-CM*.

Table 5.8: Comparison of cross-lingual summaries against reports by using BERTScore: precision (P), Recall (R), and F1 measure (F1). The top two methods are extractive, while the remaining four are abstractive; ours are marked with an asterisk.

	Australia bushfires			Fukushima earthquake			Gloria storm			Taal eruption			Zagreb earthquake			Average		
Schemes	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
C-SKIP	78.9	79.5	79.2	79.9	84.0	81.9	79.5	81.5	80.5	80.6	80.7	80.7	**82.5**	**82.8**	**82.6**	80.3	81.7	81.0
CX_DB8	75.2	77.4	76.3	78.5	80.8	79.6	76.3	78.1	77.2	77.3	78.8	78.0	77.6	78.9	78.2	77.0	78.8	77.9
LASER+LSTM+T5	79.5	78.6	79.0	81.8	83.8	82.8	79.9	81.5	80.7	75.5	78.3	76.9	82.3	81.9	82.1	79.8	80.8	80.3
Nafi	76.0	78.6	77.3	78.9	82.1	80.5	77.6	79.8	78.7	77.0	79.8	78.4	80.3	82.4	81.3	78.0	80.5	79.2
***CLiQS-CM**	**81.0**	**80.6**	**80.8**	80.6	**85.4**	82.9	**81.2**	82.0	**81.6**	**82.7**	**81.2**	**81.9**	81.8	82.2	82.0	**81.5**	**82.3**	**81.9**
***CLiQS-D-CM**	80.5	80.1	80.3	**83.1**	84.8	**83.9**	80.9	**82.2**	81.5	82.4	81.1	81.7	80.6	82.6	81.6	**81.5**	82.2	81.8

Additionally, the comparison of *CLiQS-CM* against other models, con-

sidering only English messages as input for all models, is presented in Table 5.9. In this monolingual evaluation, methods are closer to each other in terms of BERTscore similarity with the reference. *CLiQS-D-CM* models show slightly better average performance. In the Australia bushfires dataset, *C-SKIP* performs better, and in the Taal volcano eruption dataset *CLiQS-CM* performs slightly better.

Table 5.9: Comparison of English-only summaries against reports by using BERTScore: precision (P), Recall (R), and F1 measure (F1). The top two methods are extractive, while the remaining four are abstractive; ours are marked with an asterisk.

	Australia bushfires			Fukushima earthquake			Gloria storm			Taal eruption			Zagreb earthquake			Average		
Schemes	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
C-SKIP	**78.2**	**79.2**	**78.7**	79.8	80.7	80.3	77.4	78.9	78.2	77.9	80.1	79.0	80.2	81.0	80.6	78.7	80.0	79.3
CX_DB8	75.8	77.2	76.5	79.9	79.2	79.5	76.1	77.1	76.6	75.7	77.6	76.7	79.2	78.9	79.0	77.3	78.0	77.7
LASER+LSTM+T5	75.9	78.1	77.0	**81.9**	80.8	81.3	78.0	79.1	78.5	**78.0**	79.7	78.9	81.4	81.3	81.3	79.0	79.8	79.4
Nafi	75.0	78.2	76.6	77.6	80.9	79.2	75.3	77.8	76.6	72.9	78.0	75.4	77.5	79.7	78.6	75.7	78.9	77.3
***CLiQS-CM**	77.9	78.8	78.3	80.4	80.7	80.5	77.6	79.4	78.5	77.9	**80.2**	**79.0**	80.8	82.2	81.5	78.9	80.3	79.6
***CLiQS-D-CM**	78.2	78.8	78.5	81.7	**81.4**	**81.6**	**79.0**	**79.8**	**79.4**	77.7	80.0	78.8	**81.6**	**82.5**	**82.1**	**79.6**	**80.5**	**80.1**

5.5.4 Readability evaluation

The *readability* of crisis reports is crucial to provide information to practitioners in an understandable way [230]. We considered five dimensions of readability: grammaticality, non-redundancy, referential clarity, focus, structure, and coherence [110]. We performed this evaluation through a crowdsourcing platform,[12] and computed our results by aggregating the assessments of five different annotators. Annotators were shown an explanation of each annotation dimension before starting the evaluation.

A total of 43 evaluation rounds were performed, and in each round, the five annotators were shown independently the six summaries in random ordering. They were asked to pick one of them as the best in terms of each evaluation dimension. Then, we computed the best method for each round by majority voting among the five annotators. The results

[12] SurgeHQ - https://www.surgehq.ai/

of the evaluation, shown in Table 5.10, indicate that annotators considered summaries generated by *CLiQS-D-CM* as less redundant, more referentially clear, more focused, and more structured and coherent than the summaries generated by other methods. On the other hand, *C-SKIP* summaries, which are extractive summaries generated by a centroid-based method, were considered as more grammatically correct.

Table 5.10: Readability evaluation of summaries, expressed as the percentage of times a method was chosen as the best for a given dimension (column). The top two methods are extractive, the remaining four are abstractive; ours are marked with an asterisk.

Scheme	Grammaticality	Non-Redundancy	Referential Clarity	Focus	Structure and Coherence
C-SKIP	**33.5%**	14.9%	14.9%	10.2%	13.0%
CD_DB8	0.5%	4.2%	2.8%	2.8%	1.4%
LASER+LSTM+T5	19.5%	25.6%	20.0%	23.7%	25.6%
Nafi	1.4%	1.4%	5.6%	4.2%	1.4%
***CLiQS-CM**	20.0%	20.9%	27.0%	24.7%	24.2%
***CLiQS-D-CM**	25.1%	**33.0%**	**29.8%**	**34.4%**	**34.4%**

5.5.5 Expert Evaluation

The last evaluation involved three experts in emergency management, none of them a co-author of this work, working in three different EU countries: (1) an operations coordinator with a Virtual Operations Support Teams (VOST) organization, (2) a program manager at a Civil Protection Department, and (3) a project manager at an Emergency Management System. Experts were shown 43 pairs of summaries randomly selected from the five events; in each pair, which one of the summaries was generated by our method and the other by one of the baselines. Each summary was accompanied by references (links) to source tweets related to each sentence or passage in the summary, as in the following example, in which underlined letters represent links:

at least 20,000 people have taken refuge in evacuation centers. evacuees need masks. there are also many evacuees in need. evacuees need food, water, shelter and medical help.[a] more than 30 thousand people evacuated due to the eruption of the Taal volcano in the Philippines. Taal volcano eruption threatens the lives of more than 900,000 inhabitants. eruption of the Taal volcano in the Philippines has caused more than 24,000 people to be evacuated. [b,c,d]

The evaluation was performed through an online form and was "blind" in the sense that the experts did not know, in each pair, which summary was generated by which method; the ordering of each pair was randomly chosen. We performed two evaluation rounds, the first one comparing *CLiQS-D-CM* against *Nafi*, and the second one comparing *CLiQS-D-CM* against *LASER+LSTM+T5*. Experts were asked to chose in a 5-points scale whether they (1) preferred summary 1, (2) had a slight preference for summary 1, (3) considered both summaries equally preferable, (4) had a slight preference for summary 2, or (5) preferred summary 2. We additionally asked respondents to optionally comment on the quality of both summaries, if considered appropriate.

Table 5.11: Expert evaluation results: percentage of answers received, aggregated across three experts.

	← Prefers	Prefers slightly	Both equal	Prefers slightly	Prefers →	
***CLiQS-D-CM**	63.6%	14.7%	5.4%	8.5%	7.8%	**Nafi**
***CLiQS-D-CM**	55.8%	14.0%	12.4%	9.3%	8.5%	**LASER+LSTM+T5**

The results of the evaluation, shown in Table 5.11, indicate that the consulted practitioners clearly preferred summaries generated by the proposed method in 64% of the cases when compared with *Nafi*, and in 56% of the cases when compared with *LASER+LSTM+T5*. Practitioners' opinions about summaries mentioned that *CLiQS-D-CM* summaries were "more accurate and with fewer repetitions," "more information and better organized," "better explained,""more understandable," and they contain "fewer mistakes and more data." In comparison, according to their

comments, they tended to reject summaries that contain "contradictory information," "a lot of repetitions," "more mistakes and subjects mixed." Inspecting the evaluation, we noticed that in general *CLiQS-D-CM's* had lower performance in summaries related to the *Zagreb earthquake* event. In contrast, in other events, our method was often preferred.

5.6 Conclusions and Limitations

We have described a method for generating informative reports about crises from multilingual social media. This method is based on structured queries, which are matched against messages that potentially contain the information we are interested on. Queries are straightforward to construct, which means this method can be extended to a large variety of information needs. Experiments with five different disaster events indicate that we can generate high-quality, readable reports from the messages and that practitioners might prefer the summaries generated by *CLiQS-D-CM* to those generated using state-of-the-art methods.

In the work, we generated only English summaries which were useful for practitioners' evaluation. The generation of summaries in other languages could show different results. The proposed approach is flexible and allows including additional categories of information with help of queries, but we have not tested that at this point. Finally, the use of sentence embeddings (LASER) allows using the same approach for other social media (Facebook, Reddit, etc.) but this would require additional experiments for performance evaluation.

5.6.1 Reproducibility.

All of the data and code used for the experiments presented in this study is freely available in a public repository.[13]

[13]https://github.com/vitiugin/CLiQS-CM

Chapter 6

MULTILINGUAL SERVICEABILITY MODEL FOR DETECTING AND RANKING HELP REQUESTS ON SOCIAL MEDIA DURING DISASTERS

During emergencies, social media users turn to these platforms to seek quick and high-quality assistance from emergency services. However, the overwhelming influx of information on social media and the limited resources of these organizations pose challenges in effectively identifying and prioritizing critical requests. This problem becomes even more complex when users communicate in different languages, which is often the case during disasters. The delay in detecting and addressing urgent help requests can greatly impact the overall effectiveness of disaster response efforts.

This chapter presents a knowledge distillation framework, which leverages the strengths of task-related and behavior-guided models. By com-

bining these models, we train a model capable of efficiently detecting serviceable request posts across various languages on social media during natural disasters. The implementation of the framework has the potential to alleviate the cognitive load on emergency service personnel during disaster events while also being adaptable to different languages and regions worldwide.

6.1 Introduction

Social media is instrumental in connecting the public with various organizations, such as governments, non-profits, and for-profit companies [11]. In the case of for-profit companies, there has been a growing recognition of the value of providing customer service through social media. These companies often respond promptly to social media inquiries from both current and potential customers. Likewise, recent research indicates that the public expects timely responses to their social media queries directed at governments and non-profit organizations [60, 49, 126].

Table 6.1: Examples of multilingual messages with varied serviceability characteristics that were directed at emergency services' accounts on a social media platform.

	Event	Serviceability	Message
M1	Turkey-Syria Earthquake 2022	serviceable (help request)	@SERVICE Hayrullah mahallesi 16. sokak'taki Ferhat apartmanında acil yardıma ihtiyaç var! **[EN]***@SERVICE Urgent help needed at Ferhat apartment in 16th street, Hayrullah neighborhood!*
M2	Hurricane Sandy 2012	serviceable (information request)	@SERVICE how I can volunteer to help clean up after the hurricane?
M3	Catalonia Fires 2019	non-serviceable (gratitude and complaints)	@SERVICE Realizáis un gran trabajo y no os pagan lo suficiente por ello, de verdad muchas gracias **[EN]***@SERVICE You guys do a great job, and you don't get paid enough for it, really thank you so much*

To meet these expectations poses significant challenges for emergency services and non-profit organizations. During emergencies, the public

posts an enormous number of messages on social media at a high velocity, leading to information overload for emergency services that have limited human resources [118]. Further, the value of these messages for operational response varies greatly, ranging from specific requests for information or resources and concrete offers of help to unsubstantiated rumors, concerns, and prayers that may not be serviceable requests [186]. Consequently, there is an urgent need for communication departments in emergency services to quickly prioritize messages that require a timely response and have a help-seeking intent [220]. Further, there is a limited research on helping emergency services in regions with low-resource languages, or multilingual, non-English speaking populations on social media during disasters.

Table 6.1 demonstrates examples of various messages addressed to emergency services in different regions, cultures, and languages during disaster events. M1 is a prototypical serviceable message containing a concrete help request (informing the address where people need rescue). M2 is also serviceable that has a request for relevant information (asking how a user can become a volunteer). Finally, M3 is not a serviceable request for help from the perspective of operational response, but a message expressing gratitude and complaints. Capturing these various nuances of human behavior, along with understanding multilingual content, makes the task of automatically detecting a serviceable help request on social media challenging.

Our study investigates the following research questions:

- RQ 4.1. How can we teach a classification and ranking model of multilingual serviceable requests to learn different types of human behaviors when seeking help on social media during disasters?

- RQ 4.2. To what extent does the performance improvement of the proposed framework depend on the type of behavior-guided models used?

- RQ 4.3. Are there any differences in attention on various parts of a request content resulting from behavior fine-tuning, to analyze the

model's understanding of relevant human behaviors in multilingual requests?

To address these questions, our framework relies on the popular knowledge distillation process [98] for designing a computational framework called *Multiple Teachers Model for Ranking (MulTMR)*, which can detect and prioritize multilingual serviceable help requests on social media during disasters. This process aims to transfer knowledge from one or more complex models (like a *teacher*) to a simpler model (like a *student*) for a task, in order to train it to mimic the teacher models. It creatively leverages behavior-guided teacher models in the knowledge distillation process for achieving higher performance on a task. We utilize pre-trained language models that have been fine-tuned to identify sarcasm behavior and questioning behavior, which allows for more understanding of diverse user behaviors in help-seeking messages. The automated decision-making of the MulTMR is analyzed by comparing the distribution of attention weight maps within the textual posts. This novel framework enables the creation of an efficient classification and ranking system of multilingual serviceable help requests that utilizes multiple teachers, demonstrating a high level of performance to capture different human behaviors in help seeking as well as being applicable across languages and regions.

6.2 Related Work

In this section, we discuss studies that have been conducted on filtering and ranking serviceable help requests. We will also provide an overview of related literature on multilingual text classification methods for disaster-related social media posts and the teacher-student approach.

6.2.1 Social Media Requests

The literature offers insights into modeling requesting behavior or information-seeking intent across various domains, such as forums, email communi-

cation, and social media platforms. Researchers have identified request behavior in online forums across diverse contexts, such as urgency, informational intent, and social support. Furthermore, social media has emerged as a widely used channel for seeking help when individuals face challenging situations, such as health problems [95, 121], mental disorders [185], and public health emergencies [150, 136].

During natural disasters, social media has become a popular platform for users to seek help from emergency services [174, 247, 41, 60]. Whether it is for rescue, supplies, or critical information, social media is often the first point of contact for those in need. Unlike other online scenarios, posts during disasters require immediate attention and need to be directed to the intended target, such as rescue teams, for timely offline responses. As a result, special strategies have been developed to ensure that serviceable posts requesting help receive the necessary attention [221, 187, 108].

Researchers have studied the factors that influence the spread and response of requests on social media. They have focused on two categories of features: content characteristics and creator characteristics. Relevant features, such as content type, emotional tone, proximity, depth of self-disclosure, and social capital of help seekers, have been explored to determine how they affect the popularity and effectiveness of posts that contain requests [137, 139, 97]. Studies on characterizing various types of user behavior when posting messages to seek help has also been conducted through theory-driven approaches. For example, researchers have explored how theories such as the negativity bias theory [198] can be applied to help-seeking scenarios [141].

6.2.2 Knowledge Distillation and Teacher-Student Model

The Teacher-Student model is a knowledge distillation approach [98] that aims to transfer knowledge from a complex model ($Teacher$) to a simpler model ($Student$) and has been utilized for various tasks such as reducing the dimension of word embeddings [216], self-knowledge distillation [100], or contrastive learning [39]. However, the knowledge learned

from a single teacher may be limited and biased, which can result in a low-quality student model. To address this, a multi-teacher knowledge distillation framework has been proposed for pre-trained language model compression, enabling the training of high-quality student models from multiple teachers LLMs [243]. Recently, a multilingual knowledge distillation approach has been proposed that transfers knowledge from high-performance monolingual models to a multilingual model using a Teacher-Student approach, which enables the model to learn from multiple monolingual models simultaneously, resulting in improved performance [246]. Furthermore, the teacher models need not be limited to LLMs, as task-specific models can also be used to transfer specific behavioral knowledge to the student model [125]. Inspired from the last two approaches, we propose our multiple task-related teachers model for ranking serviceable requests for help.

6.3 Method

This section introduces the framework of the *MulTMR* and describe how it can be used for detection and prioritization of multilingual serviceable help requests during disasters. First, we present MulTMR framework for collaborative teaching of the student model. Second, we describe the method for behavioral fine-tuning of pre-trained multilingual LLMs using question type and sarcasm classification tasks to learn relevant user behaviors for detecting serviceable help requests.

6.3.1 *MulTMR*: Multiple Teachers Model for Ranking

Our framework is inspired from the task-related language model distillation process [125] using a diverse set of multiple teachers [243]. Its architecture presented in Figure 6.1 has two loss functions for knowledge distillation: multiple teacher hidden loss and multiple teacher distillation loss.

The multi-teacher hidden loss transfers knowledge between hidden

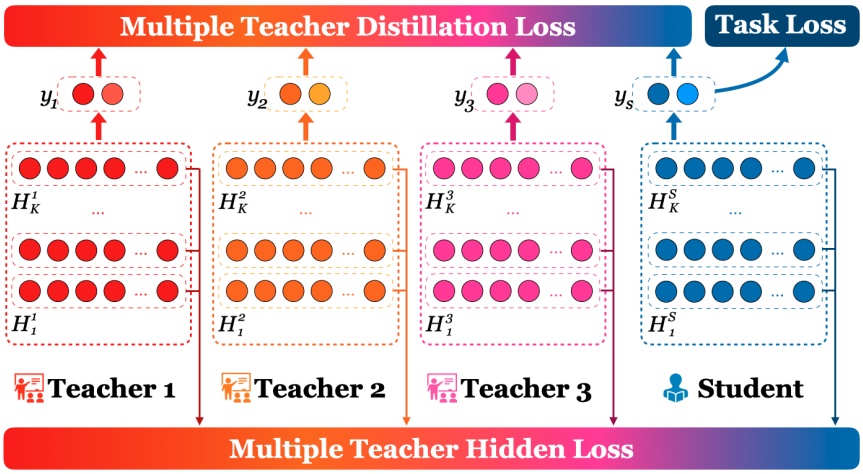


Figure 6.1: The overall architecture of MulTMR.

states of multiple teachers. Suppose there are N teacher models, and each of them has T Transformer layers. They collaboratively teach a student model with T layers, and each j-th layer in the student model corresponds to j-th layer in a teacher model.[1] Denote the hidden states output by the j-th layer of the student model as H_j^s, and the corresponding hidden states output by the j-th layer of the i-th teacher model as H_j^i. We apply the mean squared error (MSE) to the hidden states of corresponding layers in the student and teacher models to encourage the student model to have similar functions with teacher models [224]. The multi-teacher hidden loss L_{hid} is formulated as:

$$L_{hid} = \sum_{i=1}^{N} \sum_{j=1}^{T} MSE(H_j^S H_j^i W) \tag{6.1}$$

where W is a set of hyperparameters.

The multi-teacher distillation loss aims to transfer the knowledge in the soft labels output by multiple teachers to student. The predictions

[1] We assume that all teacher and student models have the same number of layers with the same size.

of different teachers on the same sample may have different correctness and confidence. Since in task-related knowledge distillation the labels of training samples are available, we used a distillation loss weighting method to assign different weights to different teachers by grid search. The multi-teacher distillation loss L_{dis} is formulated as follows:

$$L_{dis} = \sum_{i=1}^{N} CE(y_s/t, y_i/t) \tag{6.2}$$

where $CE(\cdot, \cdot)$ stands for the cross-entropy loss, y_s and $y_i$are predictions by student and teachers models respectively, t is the temperature coefficient.

Next, we incorporate gold labels y to compute the task-related loss on the predictions of the student model: $L_{task} = CE(y, y_s)$. The final loss function L for learning the student model is a summation of the multi-teacher hidden loss, multi-teacher distillation loss and the task-related loss, which is formulated as follows:

$$L = \alpha L_{task} + (1 - \alpha) L_{dis} + \beta L_{task} + (1 - \beta) L_{hid} \tag{6.3}$$

where α and β are hyperparameters.

6.3.2 Behavioral Fine-Tuning of Pre-Trained Models

Behavioral fine-tuning [200] refers to the process of teaching a model relevant capability that are useful for performing well on a target task. This is accomplished by fine-tuning the model on tasks that are related to the target task. It is called "behavioral" fine-tuning because it emphasizes the acquisition of practical behaviors, as opposed to adaptive fine-tuning. Particularly, behavioral fine-tuning using labeled data has proven effective in teaching models about various linguistic features such as named entities [29], paraphrasing [15], syntax [89], answer sentence selection [82], and question answering [123]. A recent study on fine-tuning a model on nearly 50 labeled datasets in a massively multitask environment yielded

the observation that a comprehensive and varied selection of tasks is crucial for achieving optimal transfer performance [6].

Classification of serviceable requests posted in social media during disasters is a challenging task because of extreme variety of content presented in text data for expressing diverse user behaviors. Further, ranking of multilingual serviceable help requests is a more challenging task because of increasing syntactic and semantic redundancies, i.e., multilingual model should be more context-sensitive and consider differences presented in user-generated content on different languages [238]. Our approach is based on an intuition of detecting and ranking serviceable requests for help using a behavioral fine-tuning approach, i.e., use of models for detecting specific behavior of users in a disaster relevant for region or culture, or type of disaster.

At first, we fine-tuned multilingual transformer-based model (Multilingual BERT [59] and XML-RoBERTa [46]) for detecting serviceable requests (task-related model). Next, we conducted an error analysis of common mistakes made by this task-related model. Table 6.2 demonstrates examples of detected errors. Hence, we decided to address these mistakes by use of additional teacher models as behavior-guided models during the distillation step: question type classification for the 1st, 2nd, and 3rd mistake types, sarcasm classification for the 4th type. For the 5th type of mistake, we used named entity recognition model during pre-processing step and change all location names by $LOCATION$ tag. Based on our findings, we fine-tuned the same pre-trained language model (Multilingual BERT and XML-RoBERTa) for the two behavioral tasks. Finally, we had 3 fine-tuned models with the same architecture, tokenizers, and number of output classes. We describe the implementation details in the next section.

6.4 Experiment Setup

In the section, we first describe the datasets used for the experiment and behavioral fine-tuning, the baselines and model variations, and finale im-

Table 6.2: Error examples in serviceable messages detection.

Behavior type	**Example**	**Type**
Imperative mood requests	*Prohibit rockets, firecrackers, and dangerous activities with...*	False Negative
Imperative mood requests	*Many of us have no choice... I have to drive 160km to go to work*	False Positive
Sarcastic questions	*Now if you want the Spanish army to come in and get your chestnuts out of the fire, right?*	False Positive
Sarcastic questions	*What does it say? Sorry but I don't speak Catalan and I want to find out*	False Negative
Short question	*Is there no more fire?*	False Negative
Information requests in a form close to complain	*You say we are strong together, you prevent aid. You cannot rule alone. People die while waiting for instructions. There are voices coming from under the buildings but you are passing by.. Is this unity????*	False Negative
Contextual requests	*Gazi Mustafa Kemal street No:50/A Opposite Güneşli mosque Elbistan, Kahramanmaraş*	False Negative

plementation details.

6.4.1 Data

The data for Twitter platform for serviceable requests across multiple disasters in English were presented in a recent study [186], while messages posted during Chile earthquake 2014 in Spanish were presented by CrisisNLP [107].

We also collected additional data in Spanish and Turkish via Twitter API. All collected tweets were labeled by one human assessor with

Table 6.3: Summary of datasets.

Event (start-end month/day)	Serviceable	Non-Serviceable
English		
Hurricane Sandy 2012 (10/27-11/07)	30	30
Oklahoma Tornado 2013 (05/20-06/10)	28	24
Louisiana Floods 2016 (08/14-09/29)	19	37
Alberta Floods 2013 (06/21-07/05)	190	624
Nepal Earthquake 2015 (04/15-05/15)	40	198
Hurricane Harvey 2017 (08/29-09/15)	209	1323
Spanish		
Catalonia Fires 2019 (06/04-06/30)	28	163
Chile Earthquake 2014 (04/02-04/07)	358	1197
Gloria Storm 2020 (01/26-01/28)	32	44
Turkish		
Turkey-Syria Earthquake 2023 (02/05-02/07)	980	701

language proficiency in the target language. The labeling task was to assign one of the two classes for determining whether a given tweet is serviceable or non-serviceable for a target (such as emergency services like emergenciescat, AFADBaskanlik, houstonpolice, etc.), using the similar setup as provided in prior studies [186]. Before labelling datasets, we conducted a simple preprocessing step (replaced mentions and URLs by corresponding special tokens), to filter out all uninformative tweets (with length ≤ 4 words after removing special tokens). For uncertain labelled texts, authors consulted with an emergency service practitioner. Table 6.3 presents the quantity of train and test instances for each category.

To fine-tune pre-trained LLMs for knowledge distillation using behavior-guided models, we used existing public datasets:

- Sarcasm and irony detection dataset – contains 99000 English Tweets, 33000 of which contain the hashtag #irony or #ironic and 33000 contain #sarcasm or #sarcastic [143]. We modified the dataset for fine-tuning the pre-trained model into binary. All posts from classes "sarcasm", "irony" and "figurative" were labelled as $sarcasm$, while

the last class $regular$ stayed unmodified.

- Question type classification dataset – contains 5500 questions in 6 coarse classes ("abbreviation", "entity", "description", "human", "location" and "numeric value") [140]. Based on question class description, we labelled "description" and "location" as $serviceable$, while other classes were labelled as $non - serviceable$.

During fine-tuning, we used the same parameters and number of frozen layers as during task-related fine-tuning (BERT and Robustly Optimized BERT approach (RoBERTa) models).

6.4.2 Schemes

In order to evaluate our proposed method, we compared it to commonly-used[2] pre-trained multilingual LLMs, and built a neural baseline model that utilized LSTM with DistilmBERT embeddings as input features. Both Multilingual BERT and XLM-RoBERTa models were used to evaluate the performance of our MulTMR framework. The full list of proposed modeling schemes for evaluation is the following (* denotes our proposed models and others are the baselines):

- ***[LSTM + DistilmBERT]*** – method uses pre-trained DistilmBERT embeddings,[3] which are passed as input to a LSTM Network model;
- ***[BERT]*** – multilingual BERT base model (cased);[4] were fine-tuned on the dataset with 5 frozen layers
- ***[XLM-RoBERTa]*** – XLM-RoBERTa (large-sized model;[5] were fine-tuned on the dataset with 20 frozen layers
- ***[* MulTMR-BERT]*** – MulTMR based on fine-tuned multilingual BERT model;

[2]Based on HuggingFace.com downloads statistics.
[3]https://www.sbert.net/docs/pretrained$_m odels.html$
[4]https://huggingface.co/bert-base-multilingual-cased
[5]https://huggingface.co/xlm-roberta-large

- ***[* MulTMR-RoBERTa]*** – MulTMR based on fine-tuned multilingual RoBERTa model.

We utilize three metrics to assess the effectiveness of classification models for serviceable requests detection, which are ACC, AUC, and weighted F-measure (F1), in alignment with previous studies.

In order to compare the various schemes in learning to rank task, we utilized the *normalized Discounted Cumulative Gain (nDCG)* metric, which provides a more significant weight to the discrepancies in the top positions compared to those occurring farther down the ranking outputs. Specifically, for each event/query:

$$nDCG(k) = G^{-1}_{msx,i}(k) \sum_{j:\pi_i(j)\leq k} \frac{2^{y_{i,j}} - 1}{log_2(1 + \pi_i(j))} \tag{6.4}$$

where

- $\pi_i(j)$ – position of the document d^i_j in ranking list π_i;
- $G^{-1}_{msx,i}(k)$ – normalizing factor at position k;
- $y_{i,j}$ – label of the document d^i_j in ranking list π_i.

We evaluated $nDCG$ for the top-5, top-10, and top-20 ranked message posts.

6.4.3 Model Implementation

For LLMs' fine-tuning, we used $0.5 \cdot 10^{-5}$ learning rate, 10 epochs. The number of frozen layers for each model were found by grid search. For knowledge distillation, we applied $0.6 \cdot 10^{-5}$ learning rate, 10 epochs. Based on results of grid search, we used the next hyperparameter values: $\alpha = 0.6$, $\beta = 0.5$, $t = 2$. The models were trained on NVIDIA A100-SXM4 with 40Gb GPU RAM via Google Colab.[6]

[6]https://colab.research.google.com/

Table 6.4: Comparison with baselines. Results of binary classification. Best performances are in bold. Models were trained on multilingual data: train (67%) – validation (13%) – test (20%). 5-fold CV. $*$ denotes the proposed models.

Model Scheme	ACC	F1	AUC
LSTM+DistilmBERT	80.92±0.65	62.21±5.74	85.74±3.40
BERT	80.85±2.38	81.19±2.09	80.04±0.99
XLM-RoBERTa	82.69±1.35	83.04±1.09	82.69±1.39
* *MulTMR-BERT*	88.59±1.87	88.76±1.76	88.04±1.50
* *MulTMR-RoBERTa*	**88.97±1.23**	**89.07±1.19**	**88.23±1.16**

6.5 Result Analysis and Discussion

We first discuss the results of the proposed MulTMR schemes against the baseline schemes for research question RQ 4.1, followed by an in-depth analysis of behavior-guided teacher models for RQ 4.2, and the analysis of model interpretability for RQ 4.3. Finally, we describe the analysis of cross-lingual classification scenarios and present the results for the learning to rank task as well.

6.5.1 MulTMR Performance

Table 6.4 displays the performance evaluation of MulTMR-based models against other schemes. It is evident that both proposed models, MulTMR-BERT and MulTMR-RoBERTa, exhibit superior performance compared to the baselines across all metrics. Notably, the MulTMR-RoBERTa model performs better in multilingual settings, which could be attributed to the larger size of the underlying RoBERTa model in terms of the number of layers and parameters, as compared to BERT.

Furthermore, MulTMR emphasizes the importance of behavior-guided models by incorporating additional features for serviceability help requests detection tasks. In comparison with the LSTM+DistilmBERT model, our architecture demonstrates superior performance, with an improve-

ment of approximately 3% in AUC on multilingual data. Additionally, there are significant enhancements in ACC and F1, with improvements of 8% and 27%, respectively. The primary reason for the higher performance is attributed to the knowledge distillation method used in MulTMR, which allows for the combination of hidden states and logits from multiple teachers, resulting in more accurate attention weights to achieve effective behavioral fine-tuning.

After analyzing the errors made by MulTMR, we discovered that the posts most commonly classified as false negatives were those that:

- were related to volunteering, including both offers and requests;
- were long messages that contained multiple thoughts, such as greetings and requests for information simultaneously;
- were tweets related to donations.

Similarly, the types of posts most commonly classified as false positive were those that:

- were long messages that contained multiple thoughts, such as reports and unclear help requests simultaneously;
- contained complaining about the work of emergency services.

Apart from the higher performance, our framework design enables the highlighting of behavior-related tokens, which enhances the model's sustainability against various informative types, as described in the previous section. We believe that using other behavior-guided models could improve the performance by avoiding the errors.

6.5.2 Impact of Behavior-Guided Models

In order to evaluate the impact of different teachers on the performance of the MulTMR model, we began with a baseline model that did not include any behavior-guided teacher models. We then designed two models by incorporating one behavior-guided teacher each for knowledge distillation.

Table 6.5: Behavior-specific models impact in *MulTMR-BERT* model is studied for the model schemes with the different set of teachers. The table shows p-value for each scheme's performance comparison with the baseline (*BERT*) and all-teachers model (*MulTMR-BERT*).

	baseline	+ questions	+ sarcasm	all-teachers
ACC	80.85	87.82	87.94	88.59
Compare:				
- baseline		0.00021	0.00017	0.00014
- all-teachers		0.4244	0.5225	–
F1	81.19	87.83	88.03	88.76
Compare:				
- baseline		0.00013	0.00010	0.00008
- all-teachers		0.3191	0.4099	–
AUC	80.04	85.91	86.90	88.04
Compare:				
- baseline		0.00001	0.00000	0.00000
- all-teachers		0.02878	0.4268	–

We used ACC, F1, and AUC as performance metrics to compare different model schemes, and the complete results can be found in the Table 6.5.

The results indicated that adding a second teacher to the knowledge distillation pipeline led to a statistically significant improvement in all measures. Adding just one behavior-guided teacher enhanced the model performance by 6-7%. We also compared the significance of the third teacher relative to the second. Our findings suggest that incorporating the sarcasm-detection teacher model is more significant in the AUC measure, implying that the MulTMR gained more knowledge about sarcasm than question types. This might indicate the model's effective understanding of user behavior for what to filter out in detecting serviceable help requests. Despite this, the use of the third teacher still led to an improvement of around 1% in the final model performance.

Based on our analysis, we conclude that the use of different behavior-guided teacher models results in faster convergence of model training and improved performance in the serviceable help requests classification task.

6.5.3 Analysis of Behavior-Guided Modeling

In this part of our study, we examined the relationship between behavior-guided modeling and attention weight maps generated by the MulTMR. Our objective was to investigate how MulTMR attention weights were related to behavior-guided modeling by analyzing the tweet text written in each language from the dataset.

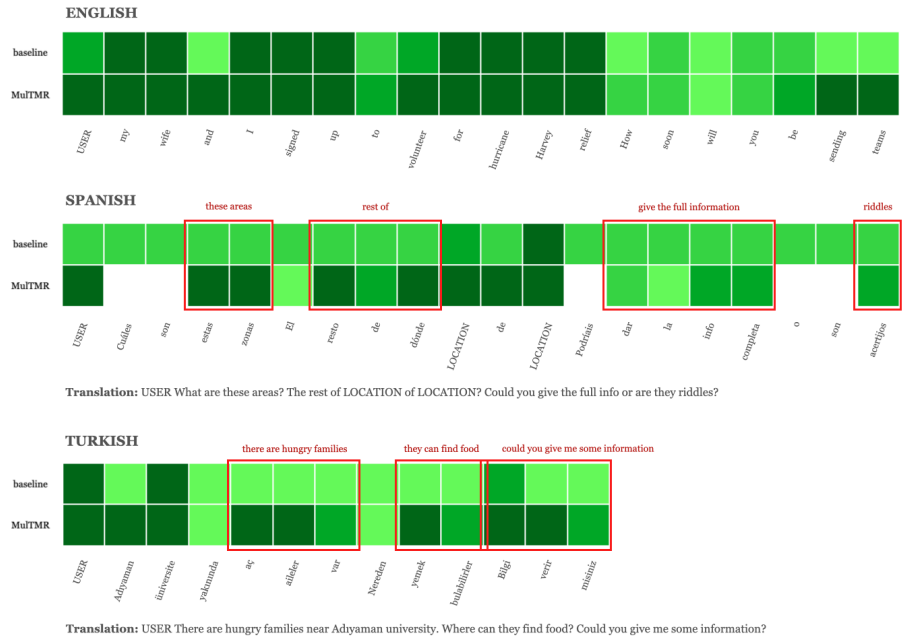


Figure 6.2: Attention weight maps of the texts in English, Spanish, and Turkish. The darker color of the word means the higher weight. Translations of target phrases are identified with red rectangles.

We utilized the MulTMR+BERT model to retrieve the attention weight maps by applying the MulTMR (with behavior-guided teachers) and without (baseline) on the texts. MulTMR+BERT model was chosen because it is faster, uses less memory, and has comparable performance to MulTMR+RoBERTa. Our findings, as depicted in the Figure 6.2, show that the attention weights

Table 6.6: Results of evaluation of post classification across 3 languages ("leave-one-language-out"): the test data contains all the posts in one language, while the training data contains messages in other languages. 5-fold CV. $*$ denotes the proposed models.

	EN & ES → TR			EN & TR → ES			ES & TR → EN			Average		
Schemes	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
LSTM+DistilmBERT	60.07	54.61	63.34	77.14	65.38	66.65	77.12	58.89	65.21	71.44	59.63	65.07
BERT	62.06	61.25	62.89	76.40	75.16	63.92	75.31	74.52	63.76	71.26	70.31	63.52
XML-RoBERTa	70.25	70.30	70.29	79.60	77.58	65.53	79.50	77.11	64.23	76.45	75.00	66.68
* *MulTMR-BERT*	67.44	67.05	66.53	77.88	76.66	67.26	77.68	76.27	67.27	74.33	73.32	67.02
* *MulTMR-RoBERTa*	**73.93**	**73.82**	**73.58**	**81.71**	**80.03**	**69.86**	**81.14**	**79.63**	**67.10**	**78.92**	**77.83**	**70.18**

of the MulTMR model focus more on valuable details. Additionally, this result was consistent for all three languages, highlighting the importance of knowledge obtained from behavior-guided teachers.

6.5.4 Cross-Lingual Performance

In addition to multilingual classification tasks, there are also cross-lingual classification settings where the languages in the training and testing data are different. To assess the proposed framework's cross-lingual capability, we utilize a "leave-one-language-out" setting, where we train and validate the MulTMR on a pair of languages and test it on the third language. To ensure unbiased results, we shuffled the training and validation data instances. The complete findings of the cross-lingual classification are outlined in Table 6.6.

In comparison to the baselines, the MulTMR exhibit improved performance for both languages. The Turkish testing yields an AUC of up to 73%, while Spanish yields almost 70% AUC, and English yields 67% AUC. The MulTMR result in a 3.5% AUC improvement compared to task-related models, and the MulTMR-RoBERTa model shows increasing in 5% compared to LSTM+DistilmBERT.

6.5.5 Learning to Rank Serviceable Help Requests

Finally, we present a supervised learning approach for automatically ranking serviceable requests using MulTMR framework. The objective of automatic ranking is to prioritize a list of posts based on their serviceability characteristics. We used the learning-to-rank methodology to achieve this goal. The learning-to-rank method aims to learn a ranking model that can associate each query with a permutation of documents that matches the training labels for relevance as closely as possible. The documents that are deemed more serviceable receive higher graded labels and are associated with higher positions in the ranking. While the proposed method can accommodate any relevance levels of serviceability, but given the labeled data, we used binary levels in this experiment. It should be noted that our approach is applicable to any relevance levels. To accomplish this, we have employed the LambdaMART algorithm [244, 31, 190], which relies on Gradient-Boosted Decision Trees (GBDT).

To train the ranking models, we utilized the data from all events except one, which was reserved for testing the model using the "leave-one-event-out" approach. We utilized sentence embeddings, which were the sum of word embeddings, using classification baselines and models as features. We then obtained rankings for each event through the 5-fold cross-validation setting.

Table 6.7 compares the performance of different schemes in terms of $nDCG$ of the first 5 positions ($nDCG@5$), 10 positions ($nDCG@10$), and 20 positions ($nDCG@20$). The results prove that MulTMR performs better than baseline models. MulTMR-RoBERTa outperforms in $nDCG@5$ and $nDCG@10$, while MulTMR-BERT shows the best results in $nDCG@20$. As expected, the MulTMR-based ranking models exhibit superior performance compared to the pre-trained DistilmBERT-based ranking model.

Table 6.7: Comparison of the average $nDCG@5$, $nDCG@10$, and $nDCG@20$. 5-fold CV. $*$ denotes the proposed models.

Scheme	$nDCG@5$	$nDCG@10$	$nDCG@20$
DistilmBERT	57.32	50.51	44.73
BERT	92.94	95.94	93.34
XML-RoBERTa	89.71	90.80	86.97
* *MulTMR-BERT*	96.14	96.34	**94.12**
* *MulTMR-RoBERTa*	**99.48**	**97.70**	93.58

6.6 Conclusions

This chapter introduced a design of the knowledge distillation framework, Multiple Teachers Model for Ranking (MulTMR), to detect and rank multilingual serviceable requests for help on social media during disasters. The core idea of MulTMR is to combine task-related and behavior-guided fine-tuned LLMs as teacher models for distilling knowledge to train a student model by optimizing hidden and distillation losses. The utilization of behavior-guided models helps to drop uncertainty of results produced by a task-related teacher model alone. MulTMR pays close attention to important parts in the context and learns to give higher attention to the potential elements of behavioral characteristics in serviceable requests classification and ranking. Experiments on the dataset of 10 events in 3 languages show that the proposed model outperforms several baselines in classification and ranking tasks. We presented extensive analyses to show the value of modeling with multiple teachers, which can help adapt the model to different languages, event types, and tasks easily. The application of the MulTMR can inform future studies on the classification and ranking of multilingual serviceable requests for help during disasters.

6.6.1 Reproducibility

Datasets and code implementation for the experiments described in this work are available for research purposes at the public repository.[7]

[7] https://github.com/vitiugin/multismc

Chapter 7

CONCLUSIONS AND FUTURE WORK

In this manuscript, we explore the potential role that multilingual social media can play in crisis management. Our work introduces novel solutions based on transfer learning, paving the way for future advancements in this field. In the following pages, we restate and summarize answers to key research questions from Chapter 1.

RQ1. What are the main actors and their respective roles in social media during crisis across various languages?

In Chapter 3 we can recognize the difference in primary actors across various language. During three war-related events, only one actor appears in all analyzed events and languages but in different roles. Other, less popular and notable actors, appear in the same roles or have similar narratives. Moreover, we examined how tones could differ, especially in such contrastive events, like war. Within a stance and a language, we found a variety of narrative constructions, encompassing encouragement of friends and criticism of foes, propagation of content from other sources, and building a larger context.

RQ2. What potential enhancements and improvements could bring

a combination of context features and a human-in-the-loop approach in the classification of multilingual social media?

Chapter 4 demonstrates how combining deep learning techniques with indirect human guidance not only enhances the overall model performance, but also increases its explainability. Even minimal human effort proves beneficial in tracing informal out-of-vocabulary words and phrases specific to country or even region. Additionally, we provide evidence that incorporating human feedback enables the establishment of connections between the attention weights assigned by the deep learning model and semantic frames.

RQ3. How does the use of a transfer learning approach contribute to enhancing the performance of detecting, classifying, and summarizing multilingual information from social media?

The flexible approach presented in Chapter 5 enables the extraction of information from crisis-related social media posts in multiple languages and generation of summaries that might be preferred by disaster experts and practitioners. State-of-the-art LLMs and transfer learning allows successful processing of messages in numerous natural languages, which used by our classification and summarization models to outperform various baseline methods. This approach is a promising way to effectively handle the vast volume of diverse content shared during disasters and mass convergence events.

RQ4. Could implementation of behavior-guided models improve detecting and ranking of multilingual help-seeking requests on social media during disasters?

In Chapter 6, we present a model based on Teacher-Student knowledge distillation technique to detect and rank multilingual serviceable messages. Through our research, we discovered that combining a specified classification model with behavior-guided models improves overall performance in zero-shot learning conditions. Furthermore, this approach directs the model's attention to specific tokens in the relevant requests and enhances explainability of decisions for practitioners.

7.1 Limitations

There are certain limitations to the research that future work could address.

First, we tested our approaches and models based on several multilingual datasets, while they are comprised of different languages and event types could affect the results of presented experiments. On the other hand, new datasets could make future models more robust and universal.

Second, human-in-the-loop methods proposed in the manuscript were tested using a simulated environment of human feedback and queries as we planned to conduct several analyses presented in the work. In future work, our approaches could be tested easily with real human agents.

Third, the applicability of our methods depends on some multilingual resources, but that do not cover all existing languages. We use Stanza for POS tagging, and dependency parsing [189] (50 languages available through community contributions), LASER embeddings to represent sentences (93 languages available), multilingual BERT (104 languages available) and XLM-RoBERTa (supports 94 languages) as LLMs for fine-tuning. Using the proposed method for unsupported languages may require training new models for the additional languages.

Last, our proposed models were tested using data collected over eleven years. During these years, the platforms and usage of social media have changed in ways that are relevant from an NLP perspective, e.g., the length of tweets was increased from 140 to 280 symbols (and during the last changes even bigger). Future research could be accomplished with more recent or streaming data to test the approaches in close-to-real conditions.

7.2 Future directions

7.2.1 Multilingual Large Language Models

Chapter 5 and Chapter 6 provide a great example of how LLMs integrate into information extraction pipelines. These LLMs serve as robust

tools capable of processing and generating texts across various languages. However, despite the significant advancements made, there are remaining challenges and promising opportunities within the field of multilingualism that warrant further exploration.

Datasets. One of the most significant challenges in multilingual research revolves around the scarcity of available data for the world's languages. Astonishingly, 88% of the world's languages lack access to text data, while for 5% of languages, the available data is extremely limited. To effectively leverage the advancements in language technology, it becomes important to collect data that is relevant for real-world applications and has the potential to positively impact speakers of underrepresented languages. Such data collections could be applicable for the development of assistive language technology for various domains such as humanitarian crises, healthcare, education, legal services, and finance. Standardized languages, as well as contact languages like creoles and regional language varieties, stand to benefit from these technologies [26]. By creating real-world datasets, researchers can establish a stronger foundation for their studies, resulting in a more substantial impact. Additionally, this approach helps bridge the gap between research and practical scenarios, increasing the likelihood that models trained on academic datasets will prove useful in real-world production surroundings. However, it is crucial for researchers and dataset creators to navigate the challenges associated with responsible Artificial Intelligence (AI) when collecting data and developing technology for underrepresented languages. Considerations such as data governance, safety, privacy, and ensuring meaningful participation of communities need to be at the forefront of these attempts [3, 27]. By addressing these challenges, researchers can contribute to the development of ethical and inclusive language technologies that empower and benefit underrepresented language communities.

Efficiency. The challenges faced by applications targeting underrepresented languages extend beyond data scarcity. Limited access to mobile data, computational resources, and expensive infrastructure poses

additional constraints [9]. To optimize the utilization of limited compute, it becomes crucial to develop more efficient methods. An exploration of efficient Transformer architectures and general efficient NLP techniques provides a comprehensive overview of the field [233]. One of the promising directions is the adaptation of these models through parameter-efficient methods. These approaches have demonstrated greater effectiveness compared to in-context learning methods [144].

Specialization. The world's languages exhibit a rich typological diversity, yet languages within one region typically exhibit shared linguistic features. For example, African languages predominantly belong to a few major language families. However, computational resources and data for most under-represented languages are limited. In light of this, it becomes crucial to integrate knowledge into language models to enhance their usefulness for these languages. Furthermore, we can leverage the fact that many under-represented languages belong to clusters of closely related languages. LLMs that focus on such language groups can effectively exchange information across languages, facilitating knowledge sharing. While recent models have primarily focused on related languages [116], future models have the potential to encompass not only related languages but also language variants and dialects. Such inclusion can foster positive transfer of knowledge from related languages, benefiting these variants and dialects. Moreover, model architectures can be adapted to incorporate information about language morphology, a crucial aspect of linguistic structure [175]. By incorporating morphological information, models can capture the intricate patterns and features unique to each language, enhancing performance and accuracy.

Controllable Text Generation. One crucial and ongoing challenge in Natural Language Generation (NLG) is achieving Controllable Text Generation (CTG). It is important for NLG systems to generate texts that adhere to specific controllable constraints as desired by humans [249]. These constraints are often task-specific, especially in the context of crisis reports, where the generated texts need to meet the requirements of

crisis practitioners [149]. CTG for the summarization of crisis social media presents an opportunity for improvement. One approach to achieving control is through fine-tuning of LLMs, which can be tailored to specific constraints and objectives. However, the current limitation lies in the availability of correspondent datasets that accurately capture the desired controllable aspects of text generation.

7.2.2 Crisis Information Extraction

Geolocation. The issue of information extraction quality has been extensively addressed in previous research [149, 158, 138], and it continues to pose a challenge today. Although deep learning algorithms have significantly improved the accuracy and efficiency of textual information extraction, determining whether a specific post corresponds to the observed time and location remains a difficult task. While knowing the exact location of a place is one of the most important aspects for emergency responders, precise GPS coordinates in social media posts are scarce, as users often provide abstract location names rather than their actual coordinates. One area of study focuses on utilizing NER techniques to identify mentioned locations in social media texts [226]. Typically, users describe locations at the administrative unit level in their posts. During emergencies, they tend to provide highly detailed location descriptions, including specific house numbers and road intersections [101]. However, accurately identifying and locating place names, especially in different languages, still poses a significant challenge for existing geoparsing services. In addition, efforts have been made to leverage visual information for inferring geographic coordinates [168]. However, the current accuracy of geolocation estimation approaches is still insufficient for robust analysis and mapping of natural disasters.

Misinformation and disinformation. The rapid spread of fake news and content on social media poses significant challenges, as they are often retweeted by numerous users, leading to the potential misallocation of resources and even endangering lives [206, 231]. Existing research has

primarily focused on detecting fake news by interpreting content [34]. However, effective detection requires not only an understanding of the current social media content provided by users but also a comparison and verification of information from various sources and users. One area of concern is the advanced capabilities of current LLMs, which can generate highly sophisticated texts that closely resemble human-authored content. This raises concerns about the potential misuse of such synthetic texts, including the dissemination of misinformation. Addressing this concern necessitates the development of robust algorithms that can effectively distinguish between genuine and fake content. These algorithms should be tailored to the unique characteristics and context of disaster-related information.

Multimodal data. Integrating the analysis of crisis-related images and videos, alongside text analysis, can greatly enhance decision-making and task prioritization for humanitarian response organizations. By leveraging annotated imagery and extracting unified semantic and visual features from social media posts, new avenues for crisis detection and damage assessment can be explored [108]. Visual media from diverse sources like Instagram, YouTube, and TikTok can be harnessed to detect and assess the severity and impact of natural disasters such as floods, fires, landslides, earthquakes, as well as man-made crises like industrial accidents or conflict fallout. This approach can also aid in evaluating infrastructure damage and its consequences on the affected population. Furthermore, the integration of information from multiple sources can provide a comprehensive and detailed overview to support effective disaster responses. Previous studies have proposed solutions that combine social media with other information sources [148, 73]. However, determining the appropriate integration approaches poses a challenge due to its subjective nature. Therefore, the development of guidelines is necessary to select integration methods based on the characteristics of different information sources. It is also important to consider the varying degrees of confidence and time-location quality associated with individual messages, which should be considered in future research.

Ultimately, to effectively address the challenge of multilingualism in information retrieval from crisis social media, it is key to incorporate cutting-edge solutions from all the mentioned domains. This necessitates a collaborative effort between researchers and practitioners to ensure the integration of state-of-the-art techniques. By employing a diverse set of tools, we can enhance the effectiveness of information retrieval and better navigate the complexities posed by disaster-related social media.

Bibliography

[1] Digital 2020: Global overview report. https://datareportal.com/reports/digital-2020-global-overview-report. [Accessed 2023-06-29].

[2] Digital 2023: Global overview report. https://datareportal.com/reports/digital-2023-global-overview-report. [Accessed 2023-06-29].

[3] R. Abebe, K. Aruleba, A. Birhane, S. Kingsley, G. Obaido, S. L. Remy, and S. Sadagopan. Narratives and counternarratives on data sharing in africa. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 329–341, 2021.

[4] F. Abrahams. When War Crimes Evidence Disappears: Social Media Companies Can Preserve Proof of Abuses. ReliefWeb. https://reliefweb.int/report/ukraine/when-war-crimes-evidence-disappears-social-media-companies-can-preserve-proof-abuses, 2022. [Accessed 15-11-2022].

[5] P. Adams. How Ukraine is winning the social media war. BBC. https://www.bbc.com/news/world-europe-63272202, 2022. [Accessed 15-11-2022].

[6] A. Aghajanyan, A. Gupta, A. Shrivastava, X. Chen, L. Zettlemoyer, and S. Gupta. Muppet: Massive multi-task representations with

pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, 2021.

[7] K. Ahmad, K. Pogorelov, M. Riegler, N. Conci, and P. Halvorsen. Social media and satellites. *Multimedia Tools and Applications*, 78(3):2837–2875, 2019.

[8] D. Ainsworth. Funding tracker: Who's sending aid to Ukraine? https://www.devex.com/news/funding-tracker-who-s-sending-aid-to-ukraine-102887, 2022. [Accessed 15-11-2022].

[9] A. F. Aji, G. I. Winata, F. Koto, S. Cahyawijaya, A. Romadhony, R. Mahendra, K. Kurniawan, D. Moeljadi, R. E. Prasojo, T. Baldwin, et al. One country, 700+ languages: Nlp challenges for underrepresented languages and dialects in indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, 2022.

[10] A. Akbik, D. Blythe, and R. Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1638–1649, 2018.

[11] H. Albanna, A. A. Alalwan, and M. Al-Emran. An integrated model for using social media applications in non-profit organizations. *International Journal of Information Management*, 63:102452, 2022.

[12] I. Alimova and V. Solovyev. Interactive attention network for adverse drug reaction classification. In *Conference on Artificial Intelligence and Natural Language*, pages 185–196. Springer, 2018.

[13] K. Allan. Natural language semantics. 2001.

[14] S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*, 2020.

[15] Y. Arase and J. Tsujii. Transfer fine-tuning: A BERT case study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5393–5404, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[16] P. Arora. *The next billion users: Digital life beyond the West*. Harvard University Press, 2019.

[17] M. Artetxe and H. Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019.

[18] P. Azunre. *Transfer learning for natural language processing*. Simon and Schuster, 2021.

[19] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760, 2017.

[20] A. Baruah, K. Das, F. Barbhuiya, and K. Dey. Aggression identification in English, Hindi and bangla text using BERT, RoBERTa and SVM. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 76–82, Marseille, France, May 2020. European Language Resources Association (ELRA).

[21] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics, 2019.

[22] BBC. Ukraine war: Kremenchuk shopping centre attack claims fact-checked. https://www.bbc.com/news/61967480, 2022. [Accessed 15-11-2022].

[23] I. Bermudez, D. Cleven, R. Gera, E. T. Kiser, T. Newlin, and A. Saxena. Twitter response to munich july 2016 attack: Network analysis of influence. *Frontiers in big Data*, 2:17, 2019.

[24] A. Bhatia. Election Disinformation in Different Languages is a Big Problem in the U.S. Center for Democracy & Technology. https://cdt.org/insights/election-disinformation-in-different-languages-is-a-big-problem-in-the-u-s/, 2022. [Accessed 15-02-2023].

[25] G. Bhatt, A. Sharma, S. Sharma, A. Nagpal, B. Raman, and A. Mittal. On the benefit of combining neural, statistical and external features for fake news identification. *arXiv preprint arXiv:1712.03935*, 2017.

[26] S. Bird. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, 2022.

[27] A. Birhane, W. Isaac, V. Prabhakaran, M. Diaz, M. C. Elish, I. Gabriel, and S. Mohamed. Power to the people? opportunities and challenges for participatory ai. *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8, 2022.

[28] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[29] S. Broscheit. Investigating entity knowledge in bert with simple neural end-to-end entity linking. *arXiv preprint arXiv:2003.05473*, 2020.

[30] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. 2020.

[31] C. J. Burges. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11(23-581):81, 2010.

[32] P. Burnap and M. L. Williams. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5(1):11, 2016.

[33] A. Capozzi, G. D. Francisci Morales, Y. Mejova, C. Monti, A. Panisson, and D. Paolotti. Facebook ads: Politics of migration in italy. In *International Conference on Social Informatics*, pages 43–57. Springer, 2020.

[34] N. Capuano, G. Fenza, V. Loia, and F. D. Nota. Content based fake news detection with machine and deep learning: a systematic review. *Neurocomputing*, 2023.

[35] C. Castillo. *Big crisis data: social media in disasters and time-critical situations*. Cambridge University Press, 2016.

[36] A. D. R. Center. Sendai framework for disaster risk reduction 2015–2030. *United Nations Office for Disaster Risk Reduction: Geneva, Switzerland*, 2015.

[37] J. M. Chamberlain, F. Spezzano, J. J. Kettler, and B. Dit. A network analysis of twitter interactions by members of the us congress. *ACM Transactions on Social Computing*, 4(1):1–22, 2021.

[38] M. Chance, N. Hodge, T. Lister, L. Smith-Spark, and I. Kottasová. Peace in europe 'shattered' as russia invades ukraine. *CNN*, 2022.

[39] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[40] Y. Chen and Q. Song. News text summarization method based on bart-textrank model. In *Proceedings of the IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 5, pages 2005–2010. IEEE, 2021.

[41] S. Cheng, L. Liu, and K. Li. Explaining the factors influencing the individuals' continuance intention to seek information on weibo during rainstorm disasters. *International Journal of Environmental Research and Public Health*, 17(17):E6072–E6072, 2020.

[42] D. Cherepnalkoski and I. Mozetič. Retweet networks of the european parliament: Evaluation of the community structure. *Applied network science*, 1(1):1–20, 2016.

[43] Z. Chi, L. Dong, S. Ma, S. Huang, S. Singhal, X.-L. Mao, H.-Y. Huang, X. Song, and F. Wei. mt6: Multilingual pretrained text-to-text transformer with translation pairs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1671–1683, 2021.

[44] D. Ciuriak. The role of social media in russia's war on ukraine. *Available at SSRN*, 2022.

[45] D. Ciuriak. Social Media Warfare Is Being Invented in Ukraine. Centre for International Governance Innovation. https://www.cigionline.org/articles/social-media-warfare-is-being-invented-in-ukraine/, 2022. [Accessed 15-11-2022].

[46] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[47] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22, 2020.

[48] G. Crupi, Y. Mejova, M. Tizzani, D. Paolotti, and A. Panisson. Echoes through time: Evolution of the italian covid-19 vaccination debate. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 102–113, 2022.

[49] L. Dahal, M. S. Idris, and V. Bravo. "it helped us, and it hurt us" the role of social media in shaping agency and action among youth in post-disaster nepal. *Journal of Contingencies and Crisis Management*, 29(2):217–225, 2021.

[50] A. Dahou, A. Mabrouk, A. A. Ewees, M. A. Gaheen, and M. Abd Elaziz. A social media event detection framework based on transformers and swarm optimization for public notification of crises and emergency management. *Technological Forecasting and Social Change*, 192:122546, 2023.

[51] T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017.

[52] J. P. De Albuquerque, B. Herfort, A. Brenning, and A. Zipf. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, 29(4):667–689, 2015.

[53] J. A. de Bruijn, H. de Moel, A. H. Weerts, M. C. de Ruiter, E. Basar, D. Eilander, and J. C. Aerts. Improving the classification of flood tweets with contextual hydrological information in a multimodal neural network. *Computers & Geosciences*, 140:104485, 2020.

[54] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, 2018.

[55] C. de Saint Laurent, V. P. Glăveanu, and I. Literat. Internet memes as partial stories: Identifying political narratives in coronavirus memes. *Social Media+ Society*, 7(1):2056305121988932, 2021.

[56] T. De Smedt. Profanity & offensive words (pow). 2020.

[57] A. Dejaifve and A. Bamas. Fresh round of fake videos claim the Bucha massacre was staged. France 24. https://observers.france24.com/en/europe/20220408-fresh-round-of-fake-videos-claim-the-bucha-massacre-was-staged, 2022. [Accessed 15-02-2023].

[58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[59] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[60] M. F. DiCarlo and E. Z. Berglund. Use of social media to seek and provide help in hurricanes florence and michael. *Smart Cities*, 3(4):1187–1218, 2020.

[61] P. Dickinson. Russia in retreat: Putin appears to admit defeat in the Battle for Kyiv. Atlantic Council.

https://www.atlanticcouncil.org/blogs/ukrainealert/russia-in-retreat-putin-appears-to-admit-defeat-in-the-battle-for-kyiv/, 2022. [Accessed 15-11-2022].

[62] M. R. Dixon, S. Dymond, R. A. Rehfeldt, B. Roche, and K. R. Zlomke. Terrorism and relational frame theory. *Behavior and Social Issues*, 12(2):129–147, 2003.

[63] L. Doroshenko and J. Lukito. Trollfare: Russia's disinformation campaign during military conflict in ukraine. *International Journal of Communication*, 15:28, 2021.

[64] S. Dowlagar and R. Mamidi. Hasocone@ fire-hasoc2020: Using bert and multilingual bert models for hate speech detection. *arXiv preprint arXiv:2101.09007*, 2021.

[65] S. Dutta, V. Chandra, K. Mehra, S. Ghatak, A. K. Das, and S. Ghosh. Summarizing microblogs during emergency events: A comparison of extractive summarization algorithms. In *Emerging Technologies in Data Mining and Information Security*, pages 859–872. Springer, 2019.

[66] Education First. EF English proficiency Index for Ukraine. https://www.ef.com/wwen/epi/regions/europe/ukraine/, 2022. [Accessed 15-11-2022].

[67] A. Erlich and C. Garner. Is pro-kremlin disinformation effective? evidence from ukraine. *The International Journal of Press/Politics*, page 19401612211045221, 2021.

[68] B. Evkoski, I. Mozetič, N. Ljubešić, and P. Kralj Novak. Community evolution in retweet networks. *Plos one*, 16(9):e0256175, 2021.

[69] B. Evkoski, I. Mozetič, P. K. Novak, and N. Ljubešić. The russian invasion of ukraine through the lens of ex-yugoslavian twitter. In *Information Society*, 2022.

[70] Facebook. Community standards. objectionable content.

[71] F. M. Federici. Translating hazards: multilingual concerns in risk and emergency communication, 2022.

[72] S. Fekih, N. Tamagnone, B. Minixhofer, R. Shrestha, X. Contla, E. Oglethorpe, and N. Rekabsaz. Humset: Dataset of multilingual information extraction and classification for humanitarian crisis response. *arXiv preprint arXiv:2210.04573*, 2022.

[73] Y. Feng, X. Huang, and M. Sester. Extraction and analysis of natural disaster-related vgi from social media: review, opportunities and challenges. *International Journal of Geographical Information Science*, 36(7):1275–1316, 2022.

[74] C. J. Fillmore et al. Frame semantics. *Cognitive linguistics: Basic readings*, 34:373–400, 2006.

[75] C. for Research on the Epidemiology of Disasters. 2022 disasters in numbers. https://www.un-spider.org/news-and-events/news/cred-publication-2022-disasters-numbers, 2023.

[76] P. Fortuna and S. Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

[77] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science*, pages 105–114, 2019.

[78] L. Franceschi-Bicchierai. Ukraine Accuses Russia of Using WhatsApp Bot Farm to Ask Military to Surrender. VICE. https://www.vice.com/en/article/5dgemn/ukraine-accuses-russia-of-using-whatsapp-bot-farm-to-ask-military-to-surrender, 2022. [Accessed 15-11-2022].

[79] M. J. Fuadvy and R. Ibrahim. Multilingual sentiment analysis on social media disaster data. In *2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, volume 6, pages 269–272, 2019.

[80] A. Galassi, M. Lippi, and P. Torroni. Attention in natural language processing.

[81] L. Gao and R. Huang. Detecting online hate speech using context aware models, 2017.

[82] S. Garg, T. Vu, and A. Moschitti. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7780–7788, 2020.

[83] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27, 2018.

[84] N. Gaumont, M. Panahi, and D. Chavalarias. Reconstruction of the socio-semantic dynamics of political activist twitter networks—method and application to the 2017 french presidential election. *PloS one*, 13(9), 2018.

[85] A. Ghermandi, J. Langemeyer, D. Van Berkel, F. Calcagni, Y. Depietri, L. E. Vigl, N. Fox, I. Havinga, H. Jäger, N. Kaiser, et al. Social media data for environmental sustainability: A critical review of opportunities, threats, and ethical use. *One Earth*, 6(3):236–250, 2023.

[86] S. Ghosh, S. Maji, and M. S. Desarkar. Gnom: Graph neural network enhanced language models for disaster related multilingual text classification. In *14th ACM Web Science Conference 2022*, pages 55–65, 2022.

[87] K. Giaxoglou. # jesuischarlie? hashtags as narrative resources in contexts of ecstatic sharing. *Discourse, context & media*, 22:13–20, 2018.

[88] E. Giuliano. The social bases of support for self-determination in east ukraine. *Ethnopolitics*, 14(5):513–522, 2015.

[89] G. Glavaš and I. Vulić. Is supervised syntactic parsing beneficial for language understanding? an empirical investigation, 2021.

[90] M. Grčar, D. Cherepnalkoski, I. Mozetič, and P. Kralj Novak. Stance and influence of twitter users regarding the brexit referendum. *Computational social networks*, 4(1):1–25, 2017.

[91] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, and N. Asokan. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, pages 2–12, 2018.

[92] Guardian. Russia protests: more than 1,300 arrested at anti-war demonstrations. https://www.theguardian.com/world/2022/sep/22/russia-protests-more-than-1300-arrested-at-anti-war-demonstrations-ukraine, 2022. [Accessed 15-11-2022].

[93] R. Guermazi, M. Hammami, and A. B. Hamadou. Using a semi-automatic keyword dictionary for improving violent web site filtering. In *2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System*, pages 337–344. IEEE, 2007.

[94] A. Gupta, D. Chugh, R. Katarya, et al. Automated news summarization using transformers. *arXiv preprint arXiv:2108.01064*, 2021.

[95] P. Gupta, A. Khan, and A. Kumar. Social media use by patients in health care: a scoping review. *International Journal of Healthcare Management*, 15(2):121–131, 2022.

[96] H. W. Hanley, D. Kumar, and Z. Durumeric. " a special operation": A quantitative approach to dissecting and comparing different media ecosystems' coverage of the russo-ukrainian war. *arXiv preprint arXiv:2210.03016*, 2022.

[97] C. He, Y. Deng, W. Yang, and B. Li. " help! can you hear me?": Understanding how help-seeking posts are overwhelmed on social media during a natural disaster. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–25, 2022.

[98] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[99] M. Honnibal and I. Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1), 2017.

[100] C.-Y. Hsieh, C.-L. Li, C.-K. Yeh, H. Nakhost, Y. Fujii, A. Ratner, R. Krishna, C.-Y. Lee, and T. Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes, 2023.

[101] Y. Hu and J. Wang. How do people describe locations during a natural disaster: An analysis of tweets from hurricane harvey. In *11th International Conference on Geographic Information Science (GIScience 2021)-Part I*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

[102] Q. Huang and Y. Xiao. Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery. *ISPRS International Journal of Geo-Information*, 4(3):1549–1568, 2015.

[103] X. Huang, L. Xing, F. Dernoncourt, and M. Paul. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1440–1448, 2020.

[104] Human Rights Watch. Russia: Growing Internet Isolation, Control, Censorship. https://www.hrw.org/news/2020/06/18/russia-growing-internet-isolation-control-censorship, 2022. [Accessed 15-11-2022].

[105] S. Ilić, E. Marrese-Taylor, J. A. Balazs, and Y. Matsuo. Deep contextualized word representations for detecting sarcasm and irony. *arXiv preprint arXiv:1809.09795*, 2018.

[106] M. Imran, C. Castillo, F. Diaz, and S. Vieweg. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):1–38, 2015.

[107] M. Imran, P. Mitra, and C. Castillo. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA).

[108] M. Imran, F. Ofli, D. Caragea, and A. Torralba. Using ai and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions, 2020.

[109] Interfax. Russia restricting Proton VPN, similar services - Roskomnadzor. https://interfax.com/newsroom/top-stories/79803/, 2022. [Accessed 15-11-2022].

[110] N. Iskender, A. Gabryszak, T. Polzehl, L. Hennig, and S. Möller. A crowdsourcing approach to evaluate the quality of query-based extractive text summaries. In *Proceedings of the 11th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019.

[111] P. Ivanova. Russians struggle to make sense of Ukraine war after Kherson retreat. Financial Times. https://www.ft.com/content/2aa769e3-c0f1-430b-bb26-c6cf10e4b529, 2022. [Accessed 15-11-2022].

[112] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one*, 9(6):e98679, 2014.

[113] A. Jadhav and V. Rajan. Extractive summarization with swap-net: Sentences and words from alternating pointer networks. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics (ACL)*, pages 142–151, 2018.

[114] M. Jamali, A. Nejat, S. Ghosh, F. Jin, and G. Cao. Social media data and post-disaster recovery. *International Journal of Information Management*, 44:25–37, 2019.

[115] S. Jhaver, I. Birman, E. Gilbert, and A. Bruckman. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):1–35, 2019.

[116] D. Kakwani, A. Kunchukuttan, S. Golla, N. Gokul, A. Bhattacharyya, M. M. Khapra, and P. Kumar. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, 2020.

[117] A. Karami, V. Shah, R. Vaezi, and A. Bansal. Twitter speaks: A case of national disaster situational awareness. *Journal of Information Science*, 46(3):313–324, 2020.

[118] M.-A. Kaufhold, N. Rupp, C. Reuter, and M. Habdank. Mitigating information overload in social media during conflicts and crises: design and evaluation of a cross-platform alerting system. *Behaviour & Information Technology*, 39(3):319–342, 2020.

[119] E. S. Kayi, L. Nan, B. Qu, M. Diab, and K. McKeown. Detecting urgency status of crisis tweets: A transfer learning approach

for low resource languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4693–4703, 2020.

[120] J. Keating. Did smartphones get dozens of Russian soldiers killed? Armies around the world are struggling to keep troops off their phones. https://www.grid.news/story/global/2023/01/04/did-smartphones-get-dozens-of-russian-soldiers-killed-armies-around-the-world-are-struggling-to-keep-troops-off-their-phones/, 2023. [Accessed 15-02-2023].

[121] M. I. Khan and J. Loh. Benefits, challenges, and social impact of health care providers' adoption of social media. *Social Science Computer Review*, 40(6):1631–1647, 2022.

[122] P. Khare, G. Burel, D. Maynard, and H. Alani. Cross-lingual classification of crisis data. In *Proceedings of the International Semantic Web Conference*, pages 617–633. Springer, 2018.

[123] D. Khashabi, S. Min, T. Khot, A. Sabharwal, O. Tafjord, P. Clark, and H. Hajishirzi. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online, Nov. 2020. Association for Computational Linguistics.

[124] S. Khatoon, A. Asif, M. M. Hasan, and M. Alshamari. Social media-based intelligence for disaster response and management in smart cities. In *Artificial Intelligence, Machine Learning, and Optimization Tools for Smart Cities: Designing for Sustainability*, pages 211–235. Springer, 2022.

[125] Y. J. Kim and H. Hassan. Fastformers: Highly efficient transformer models for natural language understanding. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 149–158, 2020.

[126] C. C. Knox. Local emergency management's use of social media during disasters: a case study of hurricane irma. *Disasters*, 47(2):247–266, 2023.

[127] R. Koshy and S. Elango. Multimodal tweet classification in disaster response systems using transformer-based bidirectional attention model. *Neural Computing and Applications*, 35(2):1607–1627, 2023.

[128] I. Krastev and M. Leonard. Peace versus Justice: The coming European split over the war in Ukraine. European Council on Foreign Relations https://ecfr.eu/publication/peace-versus-justice-the-coming-european-split-over-the-war-in-ukraine/, 2022. [Accessed 15-11-2022].

[129] J. Krishnan, A. Anastasopoulos, H. Purohit, and H. Rangwala. Cross-lingual text classification of transliterated hindi and malayalam. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1850–1857. IEEE, 2022.

[130] J. Kropczynski, R. Grace, J. Coche, S. Halse, E. Obeysekare, A. Montarnal, F. Benaben, and A. Tapia. Identifying actionable information on social media for emergency dispatch. *Proceedings of ISCRAM Asia Pacific*, page 11, 2018.

[131] R. Kumar, G. Bhanodai, R. Pamula, and M. R. Chennuru. Trac-1 shared task on aggression identification: Iit (ism)@ coling'18. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 58–65, 2018.

[132] Kyiv Post. Ukraine's success in war depends on Western support – President. https://www.kyivpost.com/ukraine-politics/ukraines-success-in-war-depends-on-western-support-president.html, 2022. [Accessed 15-11-2022].

[133] L. W. Levy, L. W. Levy, K. L. Karst, and A. Winkler. Encyclopedia of the american constitution. 2000.

[134] C. Li, W. Xu, S. Li, and S. Gao. Guiding generation for abstractive text summarization based on key information guide network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 55–60, 2018.

[135] H. Li, D. Caragea, and C. Caragea. Combining self-training with deep learning for disaster tweet classification. In *Proceedings of the 18th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, 2021.

[136] L. Li, A. Aldosery, F. Vitiugin, N. Nathan, D. Novillo-Ortiz, C. Castillo, and P. Kostkova. The response of governments and public health agencies to covid-19 pandemics on social media: a multi-country analysis of twitter discourse. *Frontiers in Public Health*, page 1410, 2021.

[137] L. Li, J. Tian, Q. Zhang, and J. Zhou. Influence of content and creator characteristics on sharing disaster-related information on social media. *Information & Management*, 58(5):103489, 2021.

[138] L. Li, Y. Wang, and Q. Cui. Exploring the potential of social media data to support the investigation of a man-made disaster: What caused the notre dame fire. *Journal of Management in Engineering*, 39(5):04023028, 2023.

[139] X. Li, A. Bahursettiwar, and M. Kogan. Hello? is there anybody in there? analysis of factors promoting response from authoritative sources in crisis. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[140] X. Li and D. Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[141] L. Lifang, W. Zhiqiang, Q. Zhang, and W. Hong. Effect of anger, anxiety, and sadness on the propagation scale of social media posts

after natural disasters. *Information Processing & Management*, 57(6):102313, 2020.

[142] H. Lin and V. Ng. Abstractive summarization: A survey of the state of the art. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9815–9822, 2019.

[143] J. Ling and R. Klinger. An empirical, quantitative analysis of the differences between sarcasm and irony. In *The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29–June 2, 2016, Revised Selected Papers 13*, pages 203–216. Springer, 2016.

[144] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. A. Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

[145] J. Liu, T. Singhal, L. T. Blessing, K. L. Wood, and K. H. Lim. Crisisbert: a robust transformer for crisis classification and contextual crisis embedding. In *Proceedings of the 32nd ACM conference on hypertext and social media*, pages 133–141, 2021.

[146] R. Liu, Y. Shi, C. Ji, and M. Jia. A survey of sentiment analysis based on transfer learning. *IEEE access*, 7:85401–85412, 2019.

[147] V. Lorini, C. Castillo, F. Dottori, M. Kalas, D. Nappo, and P. Salomon. Integrating social media into a pan-european flood awareness system: A multilingual approach. In *Proceedings of ISCRAM*. 2019.

[148] V. Lorini, C. Castillo, D. Nappo, F. Dottori, and P. Salamon. Social media alerts can improve, but not replace hydrological models for forecasting floods. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 351–356. IEEE, 2020.

[149] V. Lorini, C. Castillo, S. Peterson, P. Rufolo, H. Purohit, D. Pajarito, J. P. de Albuquerque, and C. Buntain. Social media for emergency management: Opportunities and challenges at the intersection of research and practice. In *Proceedings of the 18th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, page 772–777, 2021.

[150] C. Luo, Y. Li, A. Chen, and Y. Tang. What triggers online help-seeking retransmission during the covid-19 period? empirical evidence from chinese social media. *Plos one*, 15(11):e0241465, 2020.

[151] D. Ma, S. Li, X. Zhang, and H. Wang. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4068–4074, 2017.

[152] N. Macfarquhar. Putin's name absent as Russians discuss retreat in Ukraine. Indian Express. https://indianexpress.com/article/world/putin-absent-as-russians-discuss-ukraine-retreat-8262804/, 2022. [Accessed 15-11-2022].

[153] A. Mahajan, D. Shah, and G. Jafar. Explainable ai approach towards toxic comment classification. Technical report, EasyChair, 2020.

[154] S. Malmasi and M. Zampieri. Detecting hate speech in social media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 467–472, 2017.

[155] S. Malmasi and M. Zampieri. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202, 2018.

[156] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, and A. Patel. Overview of the hasoc track at fire 2019: Hate speech

and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, pages 14–17, 2019.

[157] R. Martins, M. Gomes, J. J. Almeida, P. Novais, and P. Henriques. Hate speech classification in social media using emotional analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 61–66. IEEE, 2018.

[158] R. McCreadie and C. Buntain. Crisisfacts: Buidling and evaluating crisis timelines. 2023.

[159] R. McCreadie, C. Buntain, and I. Soboroff. Trec incident streams: Finding actionable information on social media. In *Proceedings of ISCRAM*. 2019.

[160] McSweeney, Sinéad. Our ongoing approach to the war in Ukraine. Twitter. https://blog.twitter.com/en$_us/topics/company/2022/our-ongoing-approach-to-the-war-in-ukraine, 2022. [Accessed 15-11-2022].$

[161] U. A. Mejias and N. E. Vokuev. Disinformation and the media: the case of russia and ukraine. *Media, culture & society*, 39(7):1027–1042, 2017.

[162] Y. Mejova and N. Kourtellis. Youtubing at home: Media sharing behavior change as proxy for mobility around covid-19 lockdowns. In *13th ACM Web Science Conference 2021*, pages 272–281, 2021.

[163] D. Milmo. Russia blocks access to Facebook and Twitter. Guardian. https://www.theguardian.com/world/2022/mar/04/russia-completely-blocks-access-to-facebook-and-twitter, 2022. [Accessed 15-11-2022].

[164] S. Min, M. Seo, and H. Hajishirzi. Question answering through transfer learning from large fine-grained supervision data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 510–517, 2017.

[165] H. Mouzannar, Y. Rizk, and M. Awad. Damage identification in social media posts using multimodal deep learning.

[166] M. Mozafari, R. Farahbakhsh, and N. Crespi. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pages 928–940. Springer, 2020.

[167] M. Mozafari, R. Farahbakhsh, and N. Crespi. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861, 2020.

[168] E. Müller-Budack, K. Pustu-Iren, and R. Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In *European Conference on Computer Vision*, pages 575–592. Springer, 2018.

[169] N. Munot and S. S. Govilkar. Comparative study of text summarization methods. *International Journal of Computer Applications*, 102(12), 2014.

[170] A. Nacher. # blackprotest from the web to the streets and back: Feminist digital activism in poland and narrative potential of the hashtag. *European Journal of Women's Studies*, 28(2):260–273, 2021.

[171] N. M. Nafi, A. Bose, S. Khanal, D. Caragea, and W. H. Hsu. Abstractive text summarization of disaster-related document. In *Pro-*

ceedings of the 17th International Conference on Information Systems for Crisis Response and Management (ISCRAM), 2020.

[172] D. T. Nguyen, K. A. Al Mannai, S. Joty, H. Sajjad, M. Imran, and P. Mitra. Robust classification of crisis-related data on social networks using convolutional neural networks. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, 2017.

[173] T. H. Nguyen and K. Rudra. Learning faithful attention for interpretable classification of crisis-related microblogs under constrained human budget. In *Proceedings of the ACM Web Conference 2023*, pages 3959–3967, 2023.

[174] S. Nishikawa, N. Tanaka, K. Utsu, and O. Uchida. Time trend analysis of "# rescue" tweets during and after the 2017 northern kyushu heavy rain disaster. In *2018 5th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, pages 1–4. IEEE, 2018.

[175] A. Nzeyimana and A. N. Rubungo. Kinyabert: a morphology-aware kinyarwanda language model. *arXiv preprint arXiv:2203.08459*, 2022.

[176] R. Ogie, S. James, A. Moore, T. Dilworth, M. Amirghasemi, and J. Whittaker. Social media use in disaster recovery: A systematic literature review. *International Journal of Disaster Risk Reduction*, 70:102783, 2022.

[177] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2014.

[178] O. Oriola and E. Kotzé. Evaluating machine learning techniques for detecting offensive and hate speech in south african tweets. *IEEE Access*, 8:21496–21509, 2020.

[179] B. Ortutay. Twitter unveils version of site that can bypass Russia block. https://apnews.com/article/russia-ukraine-technology-business-europe-media-f1da10285a1631542b332597c5d35c29, 2022. [Accessed 15-11-2022].

[180] C. Y. Park, J. Mendelsohn, A. Field, and Y. Tsvetkov. Voynaslov: A data set of russian social media activity during the 2022 ukraine-russia war. *arXiv preprint arXiv:2205.12382*, 2022.

[181] J. Phengsuwan, T. Shah, N. B. Thekkummal, Z. Wen, R. Sun, D. Pullarkatt, H. Thirugnanam, M. V. Ramesh, G. Morgan, P. James, et al. Use of social media data in disaster management: a survey. *Future Internet*, 13(2):46, 2021.

[182] F. Pierri, A. Artoni, and S. Ceri. Investigating italian disinformation spreading on twitter in the context of 2019 european elections. *PloS one*, 15(1):e0227821, 2020.

[183] T. Pires, E. Schlinger, and D. Garrette. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4996–5001, 2019.

[184] A. Polese. Language and identity in post-1991 ukraine: was it really nation-building? *Studies of Transition States and Societies*, 3(3):36–50, 2011.

[185] C. Pretorius, D. McCashin, N. Kavanagh, and D. Coyle. Searching for mental health: a mixed-methods study of young people's online help-seeking. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[186] H. Purohit, C. Castillo, M. Imran, and R. Pandey. Social-eoc: Serviceability model to rank social media requests for emergency operation centers. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, pages 119–126. IEEE, 2018.

[187] H. Purohit, C. Castillo, and R. Pandey. Ranking and grouping social media requests for emergency services using serviceability model. *Social Network Analysis and Mining*, 10:1–17, 2020.

[188] H. Purohit and S. Peterson. Social media mining for disaster management and community resilience. *Big data in emergency management: Exploitation techniques for social and mobile data*, pages 93–107, 2020.

[189] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL)*, 2020.

[190] Z. Qin, L. Yan, H. Zhuang, Y. Tay, R. K. Pasumarthi, X. Wang, M. Bendersky, and M. Najork. Are neural rankers still outperformed by gradient boosted decision trees? 2021.

[191] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training, 2018.

[192] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

[193] Rating Group. The Sixth National Poll: The Language Issue in Ukraine (March 19th, 2022). https://ratinggroup.ua/en/research/ukraine/language_issue_in_ukraine_march_19th_2022.html, 2022. [Accessed 30-11-2022].

[194] C. Restrepo-Estrada, S. C. de Andrade, N. Abe, M. C. Fava, E. M. Mendiondo, and J. P. de Albuquerque. Geo-social media as a proxy for hydrometeorological data for streamflow estimation and to improve flood monitoring. *Computers & Geosciences*, 111:148–158, 2018.

[195] M. Ringgaard, R. Gupta, and F. C. Pereira. Sling: A framework for frame semantic parsing. *arXiv preprint arXiv:1710.07032*, 2017.

[196] G. Rossiello, P. Basile, and G. Semeraro. Centroid-based text summarization through compositionality of word embeddings. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pages 12–21, 2017.

[197] A. Roush. Cx db8: A queryable extractive summarizer and semantic search engine. *arXiv preprint arXiv:2012.03942*, 2020.

[198] P. Rozin and E. B. Royzman. Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, 5(4):296–320, 2001.

[199] F. S. Rubin, A. C. Alvim, R. P. dos Santos, and C. E. R. de Mello. Detecting influential communities in twitter during brazil oil field auction in 2019. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 897–904. IEEE, 2020.

[200] S. Ruder. Recent Advances in Language Model Fine-tuning. http://ruder.io/recent-advances-lm-fine-tuning, 2021.

[201] S. Ruder. The State of Multilingual AI. http://ruder.io/state-of-multilingual-ai/, 2022.

[202] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, pages 15–18, 2019.

[203] K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, and S. Ghosh. Extracting situational information from microblogs during disaster events: a classification-summarization approach. In *Proceedings*

of the 24th ACM International on Conference on Information and Knowledge Management (CIKM), pages 583–592, 2015.

[204] K. Rudra, P. Goyal, N. Ganguly, P. Mitra, and M. Imran. Identifying sub-events and summarizing disaster-related information from microblogs. In *Proceedings of the 41st International Conference on Research anb Development in Information Retrieval (SIGIR)*, pages 265–274, 2018.

[205] K. Rudra, A. Sharma, N. Ganguly, and M. Imran. Classifying and summarizing information from microblogs during epidemics. *Information Systems Frontiers*, 20(5):933–948, 2018.

[206] G. Ruffo, A. Semeraro, A. Giachanou, and P. Rosso. Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language. *Computer science review*, 47:100531, 2023.

[207] A. Sahu and S. G. Sanjeevi. Better fine-tuning with extracted important sentences for abstractive summarization. In *Proceedings of the International Conference on Communication, Control and Information Sciences (ICCISc)*, volume 1, pages 1–6. IEEE, 2021.

[208] V. Salama. Republican Opposition to Helping Ukraine Grows, WSJ Poll Finds. The Wall Street Journal. https://www.wsj.com/articles/republican-opposition-to-helping-ukraine-grows-wsj-poll-finds-11667467802, 2022. [Accessed 15-11-2022].

[209] D. Salza, E. Arnaudo, G. Blanco, and C. Rossi. A'glocal'approach for real-time emergency event detection in twitter. In *ISCRAM 2022 Conference Proceedings-19th International Conference on Information Systems for Crisis Response and Management*, 2022.

[210] N. S. Samghabadi, P. Patwa, P. Srinivas, P. Mukherjee, A. Das, and T. Solorio. Aggression and misogyny detection using bert: A

multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, 2020.

[211] C. Sánchez, H. Sarmiento, J. Pérez, A. Abeliuk, and B. Poblete. Cross-lingual and cross-domain crisis classification for low-resource scenarios. *arXiv preprint arXiv:2209.02139*, 2022.

[212] D. L. Sánchez, J. R. Herrero, F. de la Prieta, and C. Dang. Analysis and visualization of social user communities. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, 4(3):11–18, 2015.

[213] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[214] A. Saroj and S. Pal. Use of social media in crisis management: A survey. *International Journal of Disaster Risk Reduction*, 48:101584, 2020.

[215] Y. Senarath and H. Purohit. Evaluating semantic feature representations to efficiently detect hate intent on social media. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 199–202. IEEE, 2020.

[216] B. Shin, H. Yang, and J. D. Choi. The pupil has become the master: teacher-student model-based word embedding distillation with ensemble learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3439–3445, 2019.

[217] I. Singh, K. Bontcheva, X. Song, and C. Scarton. Comparative analysis of engagement, themes, and causality of ukraine-related debunks and disinformation. In *International Conference on Social Informatics*, pages 128–143. Springer, 2022.

[218] B. Smart, J. Watt, S. Benedetti, L. Mitchell, and M. Roughan. # istandwithputin versus# istandwithukraine: The interaction of bots

and humans in discussion of the russia/ukraine war. In *International Conference on Social Informatics*, pages 34–53. Springer, 2022.

[219] D. Smeltz, E. Sullivan, L. Wojtowicz, D. Vokov, and S. Goncharov. Russian Public Accepts Putin's Spin on Ukraine Conflict. The Chicago Council on Global Affairs. https://globalaffairs.org/sites/default/files/2022-04/Final%20Russia%20Brief%20V3.pdf, 2022. [Accessed 15-11-2022].

[220] L. S. Snyder, M. Karimzadeh, C. Stober, and D. S. Ebert. Situational awareness enhanced through social media analytics: A survey of first responders. In *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*, pages 1–8. IEEE, 2019.

[221] C. Song and H. Fujishiro. Toward the automatic detection of rescue-request tweets: Analyzing the features of data verified by the press. In *2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, pages 1–4. IEEE, 2019.

[222] I. Stanley-Becker and D. Rosenzweig-Ziff. A TikTok video, an apology and allegations of disinformation: How Russia's assault on Ukraine is inflaming German fears of hybrid warfare. The Washington Post. https://www.washingtonpost.com/world/2022/03/22/russian-video-fake-disinformation-germany/, 2022. [Accessed 30-11-2022].

[223] P. Suciu. Is Russia's Invasion Of Ukraine The First Social Media War? https://www.forbes.com/sites/petersuciu/2022/03/01/is-russias-invasion-of-ukraine-the-first-social-media-war/?sh=5dd83f2c1c5c, 2022. [Accessed 15-11-2022].

[224] S. Sun, Y. Cheng, Z. Gan, and J. Liu. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*, 2019.

[225] W. Sun, P. Bocchini, and B. D. Davison. Applications of artificial intelligence for disaster management. *Natural Hazards*, 103(3):2631–2689, 2020.

[226] R. Suwaileh, T. Elsayed, M. Imran, and H. Sajjad. When a disaster happens, we are ready: Location mention recognition from crisis tweets. *International Journal of Disaster Risk Reduction*, 78:103107, 2022.

[227] TASS. Azov battalion militants blow up Mariupol theater building — Defense Ministry. https://tass.com/world/1423275, 2022. [Accessed 15-11-2022].

[228] TASS. Lavrov slams situation in Bucha as fake attack staged by West and Ukraine. https://tass.com/world/1432013, 2022. [Accessed 15-11-2022].

[229] A. Tegos, A. Ziogas, V. Bellos, and A. Tzimas. Forensic hydrology: A complete reconstruction of an extreme flood event in data-scarce area. *Hydrology*, 9(5):93, 2022.

[230] I. Temnikova, S. Vieweg, and C. Castillo. The case for readability of crisis communications in social media. In *Proceedings of the 24th International Conference on World Wide Web (WWW)*, pages 1245–1250, 2015.

[231] S. Torpan, S. Hansson, M. Rhinard, A. Kazemekaityte, P. Jukarainen, S. F. Meyer, A. Schieffelers, G. Lovasz, and K. Orru. Handling false information in emergency management: A cross-national comparative study of european practices. *International Journal of Disaster Risk Reduction*, 57:102151, 2021.

[232] Translators without boarders. Language data for Ukraine. https://translatorswithoutborders.org/language-data-for-ukraine, 2022. [Accessed 15-11-2022].

[233] M. Treviso, T. Ji, J.-U. Lee, B. van Aken, Q. Cao, M. R. Ciosici, M. Hassid, K. Heafield, S. Hooker, P. H. Martins, et al. Efficient methods for natural language processing: A survey. *arXiv preprint arXiv:2209.00099*, 2022.

[234] A. C. Valdez and M. Ziefle. Human factors in the age of algorithms. understanding the human-in-the-loop using agent-based modeling. In *International Conference on Social Computing and Social Media*, pages 357–371. Springer, 2018.

[235] M. Vengattil and E. Culliford. Facebook allows war posts urging violence against Russian invaders. Reuters. https://www.reuters.com/world/europe/exclusive-facebook-instagram-temporarily-allow-calls-violence-against-russians-2022-03-10/, 2022. [Accessed 15-11-2022].

[236] R. J. M. Ventayen. Multilingual detection and mapping of emergency and disaster-related tweets. *MATTER: International Journal of Science and Technology*, 3(2), 2017.

[237] F. Vitiugin and G. Barnabo. Emotion detection for spanish by combining laser embeddings, topic information, and offense features. In *Proc. of EmoEvalEs at IberLEF*, 2021.

[238] F. Vitiugin and C. Castillo. Comparison of social media in english and russian during emergencies and mass convergence events. In *ISCRAM*, 2019.

[239] T. I. Walker, S. Putin confirms crimea annexation as ukraine soldier becomes first casualty. Guardian. https://www.theguardian.com/world/2014/mar/18/putin-confirms-annexation-crimea-ukrainian-soldier-casualty, 2014. [Accessed 6-07-2023].

[240] C. Wang. Interpreting neural network hate speech classifiers. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 86–92, 2018.

[241] E. Woo. Twitter will stiffen moderation policies in response to the war in Ukraine. https://www.nytimes.com/2022/04/05/business/twitter-policy-ukraine.html, 2022. [Accessed 15-11-2022].

[242] I. D. Wood, J. Glover, and P. Buitelaar. Community topic usage in online social media. *ACM Transactions on Social Computing*, 3(3):1–21, 2020.

[243] C. Wu, F. Wu, and Y. Huang. One teacher is enough? pre-trained language model distillation from multiple teachers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4408–4413, 2021.

[244] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 13:254–270, 2010.

[245] K.-C. Yang, O. Varol, P.-M. Hui, and F. Menczer. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1096–1103, 2020.

[246] Z. Yang, Y. Cui, Z. Chen, and S. Wang. Cross-lingual text classification with multilingual distillation and zero-shot-aware training. *arXiv preprint arXiv:2202.13654*, 2022.

[247] H. Zade, K. Shah, V. Rangarajan, P. Kshirsagar, M. Imran, and K. Starbird. From situational awareness to actionability: Towards improving the utility of social media data for crisis response. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–18, 2018.

[248] C. Zhang, C. Fan, W. Yao, X. Hu, and A. Mostafavi. Social media for intelligent public information and warning in disasters: An interdisciplinary review. *International Journal of Information Management*, 49:190–207, 2019.

[249] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song. A survey of controllable text generation using transformer-based pre-trained language models. *arXiv preprint arXiv:2201.05337*, 2022.

[250] L. Zhang, Z. Liu, S. Zhang, X. Yang, H. Qiao, K. Huang, and A. Hussain. Cross-modality interactive attention network for multispectral pedestrian detection. *Information Fusion*, 50:20–29, 2019.

[251] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. In *Proceedings of the International Conference on Learning Representations*, 2019.