An NLP Case Study on Predicting the Before and After of the Ukraine–Russia and Hamas–Israel Conflicts

Jordan Miner
Hofstra University
Hempstead, New York
jminer4@pride.hofstra.edu

John E. Ortega
Hofstra University
Northeastern University
john@naturallang.com

Abstract

We propose a method to predict toxicity and other textual attributes through the use of natural language processing (NLP) techniques for two recent events: the Ukraine–Russia and Hamas–Israel conflicts. This article provides a basis for exploration in future conflicts with hopes to mitigate risk through the analysis of social media before and after a conflict begins. Our work compiles several datasets from Twitter and Reddit for both conflicts in a *before* and *after* separation with an aim of predicting a future state of social media for avoidance. More specifically, we show that: (1) there is a noticeable difference in social media discussion leading up to and following a conflict and (2) social media discourse on platforms like Twitter and Reddit is useful in identifying future conflicts before they arise. Our results show that through the use of advanced NLP techniques (both supervised and unsupervised) toxicity and other attributes about language before and after a conflict is predictable with a low error of nearly 1.2 percent for both conflicts.

1 Introduction

In the past decade, social media has had a massive impact on how we communicate as a society in its ability to sway public opinion and shape political landscapes (Dylko et al., 2018). In particular, the nature of the algorithms used in social networking platforms will oftentimes amplify extremist perspectives and provide users who hold these views a platform in which they can connect and share ideas (Church et al., 2022). It is our hypothesis that through the use of natural language processing (NLP) we could potentially help avoid social media becoming a catalyst for conflict as it has in the past.

In this study, we use NLP to examine interactions from social media on two well-known, recent conflicts: Ukraine–Russia and Hamas–Israel. We examine the role of social media in the emergence of both conflicts by gathering data from Reddit[1] and Twitter[2] and then segmenting the data into four main datasets based on date posted: (1) *before* Ukraine–Russia (2) *after* Ukraine–Russia (3) *before* Hamas–Israel and (4) *after* Hamas–Israel.

We first reveal important insights on the segmented datasets using unsupervised techniques during development that lead to further exploration of predictive capabilities. For prediction, we use toxicity scores as a method of determining the type of language that leads up to and is used after a conflict begins based on the unsupervised results. By recognizing toxic language patterns leading up to a conflict, we can use these toxicity scores as a tool for avoidance—defined as a mechanism to prevent the escalation of a conflict by addressing or mitigating factors before they trigger or exacerbate a conflict.

Our findings show that avoidance through the use of state-of-the-art NLP techniques can be achieved on the two conflicts studied. To better illustrate our work we show that other work has not studied the more recent conflicts or used toxicity for prediction in Section 2. We then illustrate the details of our dataset segmentation and methods in Section 3. Next, in Section 4 and Section 5 we provide results and discussion from our experimentation. Finally, in Section 6 we conclude with comments about achievements and next steps.

2 Related Work

When used as a source of information, social media platforms' user-driven model has been known to lead to self-reinforcing polarization, a method to shape specific narratives, and act as echo chambers containing negative rhetoric to describe political or social events (Dylko et al., 2018; Natali Helberger and D'Acunto, 2018; Church et al., 2022;

[1] https://www.reddit.com

[2] https://www.twitter.com

Proceedings of the Third Workshop on NLP for Positive Impact, pages 140–151
November 15, 2024 ©2024 Association for Computational Linguistics

Kaiser and Rauchfleisch, 2020) As of this paper, research in the context of both the ongoing Ukraine–Russia and Hamas–Israel have not been compared. In the past, there has been investigations about the intricacies of toxic language on social media with the Detoxify model (Sheth et al., 2022; Taleb et al., 2022; Nagavi and S., 2021; He et al., 2024), but many of this research identified toxic content that spanned a variety of categories, rather than focusing on discussions surrounding a potential or ongoing conflict. While previous literature observes public discourse of the Ukraine–Russia conflict through the use of Latent Dirichlet Allocation (LDA) for topic modeling (Aslan, 2023; Sazzed, 2022; Chang et al., 2023; Maathuis and Kerkhof, 2023), many of these are used in combination with sentiment analysis only to gain an understanding of the opinions of perception of users on social media platforms like Twitter (now known as X). Additionally, LDA has also been used to observe Russian state-sponsored accounts on Twitter and their influence 2016 United States Elections (Zannettou et al., 2019), and has been compared with alternative methods to estimate latent topics (Golino et al., 2021).

Other investigations of public sentiment surrounding the Hamas–Israel conflict have taken place using sentiment analysis prior to its beginning (Nurlela et al., 2023; Gangwar and Mehta, 2023). Likewise, Chen et al. (2024) utilizes an innovative keyword extraction framework on Reddit posts created before and after the Hamas–Israel conflict, and the sentiment for a given comment was assessed using emotions like fear or sadness. Our works compares two major conflicts on a *before* and *after* data segmentation. Previous research has also been carried out by Celiku and Kraay (2017) focusing on conflict prediction, but, to our knowledge, other work has not compiled the same corpora into four segmented datasets. Additionally, we provide two major aspects of prediction: topic discovery and conflict prediction for avoidance as described in Section 3.

3 Methodology

In this section we focus on the data collection and preparation necessary to repeat our experiments along with the model preparation for both *unsupervised* discovery and *supervised* prediction for avoidance. The work is made publicly available[3] for others to consume with the aim of somehow “sounding the alarm” for future conflicts through social media.

3.1 Data Collection and Processing

A total of four dataset were obtained to examine the role social media has in avoiding future conflicts. We again denote the datasets as the following, this time adding additional acronyms for reference purposes: (1) *before* Ukraine–Russia (**URB**) (2) *after* Ukraine–Russia (**URA**) (3) *before* Hamas–Israel (**HIB**) and (4) *after* Hamas–Israel (**HIA**).

It is noteworthy to take into account that we only processed posts in English and we feel that additional bias may have been introduced by doing so, as both conflicts took place between populations whose primary language is not English. Nonetheless, we would not want to get *lost in translation* due to language differences as shown in the past (Van Nes et al., 2010). Furthermore, the work obtained from this investigation is still helpful as it provides insight the perspectives of the international audience. In the 2014 Gaza War, social media allowed "Israel and Hamas to tailor their message to international supporters, and monitor their feedback extremely quickly" (Zeitzoff, 2018). In doing so, these international supporters can then pressure their governments to choose a side in a dispute and even change the dynamics and scope. Therefore, while international audiences might not be the directly involved, their opinions can garner political or social support in ongoing disputes that can escalate tensions into a conflict.

URB and URA are described in the following. The first Ukraine–Russia dataset (URB) consisted of tweets posted before the conflict began with dates ranging from 31 December 2021 to 23 February 2022 (Purtova, 2022) that contained 835,142 documents gathered from searches including "ukraine war", "ukraine NATO", "StandWithUkraine", and "russian border ukraine" to name a few. The second Ukraine–Russia dataset (URA) was composed of tweets posted after the conflict began ranging from 24 February 2022 to 25 March 2022 (BwandoWando, 2024), and contained 8,268,526 documents gathered using hashtags such as "ukraineunderattack", "RussianConflict", "StopPutinNow" and "UkraineConflict" among others.

[3] https://naturallang.com/conflict/conflict.html

Table 1: Top 5 N-grams for Each Topic by Dataset.
https://naturallang.com/conflict/conflict.html

Dataset	Topic	Top 5 Bigrams/Trigrams
HIB	Topic 1	"fifa worldcup", "palestine flag", "good morning", "support palestine"
	Topic 2	"human right", "world cup", "palestine action", "palestinian flag"
	Topic 3	"free palestine", "palestine free", "israel palestine, "israeli apartheid"
	Topic 4	"gaza strip", "palestinian people", "solidarity palestine", "day solidarity"
HIA	Topic 1	"sub reddit", "action performed", "bot action", "action performed automatically"
	Topic 2	"word news", "gaza strip", "hamas terrorist", "sub reddit"
	Topic 3	"west bank","middle east", "support hamas", "israeli government"
	Topic 4	"state solution", "make sense", "human shield", "sound like"
URB	Topic 1	"near ukraine border", "ukraine case", "troop surrounding", "nato troop"
	Topic 2	"russian star", "ukraine case", "twitter come time", "twitter com time status"
	Topic 3	"ukraine believe", "war prevent", "news euro", "twitter com time"
	Topic 4	"ukraine case", "twitter com time", "twitter com time status", "russia threat invade"
URA	Topic 1	"russia ukraine", "ukraine war", "ukraine russian", "ukraine ukraine"
	Topic 2	"urkaine russia", "russia war", "ukraine russia war", "war ukraine"
	Topic 3	"ukraine need", "airlift ukraine", "safe airlift", "safe airlift ukraine"
	Topic 4	"stand ukraine", "slava rain", "people ukraine", "president lensky"

The remaining datasets (HIB and HIA) contained posts from Twitter and Reddit discussing the Hamas–Israel conflict. HIB was composed of tweets posted on Twitter before the war began with dates ranging from 1 September 2022 to 30 December 2022 (Erroukrma, 2023), with a total of 24,251 documents generated from keywords mentioning "Palestine" or "Gaza." The HIA dataset consisted of posts made on Reddit from 7 October 2023 to 29 October 2023 (Asaniczka, 2024) and contained 436,725 documents gathered from subreddits like /WorldNews and /IsraelPalestine.

All four of the datasets were first tokenized using the natural language toolkit[4] (NLTK). We removed URLs, non-alphabetical characters, accents, and English stopwords. Additionaly, we tokenized the text and lemmatized using NLTK's WordNetLemmatizer[5]. Likewise, since Twitter is known for using hashtags, any hashtags were deconstructed into separate words using WordNinja[6].

Since the datasets for the Ukraine–Russia conflicts were quite large and we were limited to one GPU Tesla A100 machine with 20GB of ram, we decided to use a smaller dataset which consisted of the 174,292 URB and 1,240,279 URA documents. Size reduction was done using random sampling and stratification. Contrastingly, the HIB and HIA datasets were smaller with 20–400k documents.

Feature vocabularies for the four datasets were first vectorized using a count vectorizer. For URB and URA, we had to limit the vocabulary to a minimum of document frequency of 5 and a maximum of 85 percent. On the other hand, the HIB and HIA datasets were set to a minimum document frequency of 5 percent and a maximum document frequency of 90. These settings resulting in a vocabulary of 5,000 terms for each corpus based on n-grams ranging from size 2 to 4. Terms for each document were combined in a term-document matrix and used in for experimentation in the *unsupervised* setting that follows.

3.2 LDA Topic Modeling

The unsupervised topic modeling based on LDA was used to determine whether certain documents could be grouped together based on their textual data. The optimal number of topics were obtained through experimentation to find which parameters yielded the most distinct topics and minimize any overlapping as much as possible. This yielded a total of 9 topics for the Ukraine–Russia conflict, and 7 topics for the Hamas–Israel conflict.

3.3 Toxicity Prediction

In order to better understand how the term "avoidance" is deemed in this article, we present the idea of *toxicity* as a prediction task. In the context of this investigation and its relevance to conflict, we define toxicity as content that fosters polarization between opposing sides, spreads distrust, and re-

[4] https://www.nltk.org/

[5] https://www.nltk.org/_modules/nltk/stem/wordnet.html

[6] https://github.com/keredson/wordninja

inforces an 'us vs. them' narrative, which further encourage division and hostility. Toxic content of this nature is oftentimes used to promote the radicalization of individuals online, shape narratives about one's own group, and mobilize supporters to act (Zeitzoff, 2017).

It is our belief that the datasets in the two conflicts studied seem to become more toxic after a conflict had begun. This makes our task somewhat distinct from a sentiment task by digging deeper into the language, like hate speech and more, that seem to provoke and sway sentiment.

In our experiments, we used numeric toxicity values to provide an approximation between zero and one where a 0.00 toxicity score signifies not toxic at all and a score of 1.00 means extremely toxic. We use the toxicity value because it provides a fundamental assessment of whether the text content was negative or harmful in nature so that we could examine its relevance in conflict causation. To do this, we assigned each bag-of-word feature a toxicity score using the Detoxify[7] (Hanu and Unitary team, 2020) library that identifies toxic content as "obscenity, threats, and identity-based hate. The toxicity scores were calculated in batches of 100, and then stored in a dictionary where each n-gram was given corresponding toxicity scores between 0.00 and 1.00.

3.4 Linear Regression

We chose a *supervised* linear regressor (LR) to establish a baseline toxicity prediction where URB and HIB were used to predict the toxicity scores of URA and HIA, respectively. Section 4 provide more insight into the original LDA results that helped show the before/after toxicity analytics. For instance, if the model predicts higher toxicity scores for social media posts after a conflict starts, toxicity and even later sentiment can be used as a mechanism of avoidance **before** a conflict hits a highly toxic point. For that reason, we attempt to predict URA and HIA toxicity with the aim of accurately predicting a future toxicity.

Independent variables for the LR model were created using document matrices similar to the *unsupervised* LDA experiment. A document's toxicity score was calculated by collecting the toxicity scores of terms present in a given document, with each term associated with a calculated toxicity score described in 3.3, and then calculating the average of these scores. In doing so, the LR models then used the average document scores from URB and HIB and the term-frequency matrices to predict the average toxicity scores for each document in the URA and HIA. For the entire set URB, URA, HIB, HIA prediction was done for individual conflicts such that URB–>URA and HIB–>HIA; we left mixing of the conflicts for future work.

To evaluate the performance of the models, the predicted toxicity scores and actual toxicity scores were compared using the mean squared error (MSE) and mean absolute error (MAE) The mean square error is the squared difference between the actual values and the predicted values, and the mean absolute examines the absolute difference; both being used to indicate how close the line of best fit is to the set of points (Tatachar, 2021). We also employed RobustScaler to scale URB and URA due to the presence of many outliers in those datasets, and MaxAbScaler for HIB and HIA as those did not contain many outliers. Both the RobustScaler and the MaxABScaler were from SciKit Learn's latest stable release, version 1.4.2.The default LR is used, which has the *fit intercept* value to true. According to SciKit Learn, the default LR is: “just plain Ordinary Least Squares (scipy.linalg.lstsq) or Non Negative Least Squares (scipy.optimize.nnls) wrapped as a predictor object”.

3.5 BERT

For comparison purposes, we compared the LR to a tranformer-basesd (Vaswani et al., 2017) model. The transformer model is a state-of-the-art model based on the BERT (Devlin et al., 2019) architecture. This allowed us to use a pre-trained language model with the aim of transfer learning to include data from external sources along with fine-tuning on our data.

We selected the BERT model created by Mishra et al. (2020a) that had been trained on posts taken from Twitter and Youtube with the purpose of distinguishing instances of Trolling, Agression and Cyberbullying (Mishra et al., 2020b). The hyperparameters used for fine-tunig/training are listed in Table 2.

We illustrate the two machine learning tasks for conflict avoidance based first on a *unsupervised* technique for hypothesis approbation and then secondly with two *supervised* regressors to better understand how valid our conflict avoidance hypothesis works.

[7] https://github.com/unitaryai/detoxify

Hyperparameter	Value
fp16	True
epochs	3
per device train batch size	16
per device eval batch size	16
weight decay	0.01
learning rate	.00002
save total limit	10
evaluation strategy	epoch

Table 2: Hyperparameters for training the Bert-based model using a before–>after conflict method.

4 Results

4.1 LDA Topic Modeling

After calculating the toxicity scores of the n-grams, we wanted to inspect how the toxicity scores varied from one cluster to another. To do this, we utilized a topic-document matrix that classified documents based on their predominant topics, and a document could only be assigned to a topic so long as its highest association score was at least 80 percent. The results from this were then stored in a dictionary where each topic index was associated with a list of strongly linked documents.

Subsequently, by mapping the documents to the topics, and the n-grams to documents, we were then able to create a dictionary mapping topics to the n-grams, or terms, they encompassed. The resulting clusters can be visualized online by clicking here. Ultimately, in using this method, we obtained the toxicity scores of each topic by extracting each term in the topic-term dictionary and matching it to the toxicity scores in the term-toxicity dictionary. The collected toxicity scores were then aggregated to compute the average, total, maximum and minimum toxicity scores for each topic as illustrated in Figures 1 and 2.

For URB and URA, it appears that the minimum toxicity scores were mostly consistent across topics, and the minimum toxicity scores for before the conflict were slightly higher but still very close to 0.[8] The average and total toxicity scores experienced a significant increase once the conflict began, as indicated by the higher scores for URA.The difference in toxicity were the most dramatic for Topic 6 in URB and Topic 6 in URA. Interestingly, it appeared that the URB dataset seemed to contain many extreme values since most of their maximum and minimum toxicity were higher in the URB dataset even though, on average, the URB toxicity scores were lower. Overall, it appears that, on average, most of the toxicity scores seemed to have increased upon the emergence of the war.

Toxicity levels between HIB and HIA also saw an increase once the conflict began. The minimum toxicity scores all appear to be the same across all topics, with the HIB toxicity scores being much lower than the HIA scores. That being said, with an exception for Topic 6 (https://naturallang.com/conflict/conflict.html), all of the maximum toxicity scores increased after the initial start of the conflict. The average and total toxicity scores, for the most part, were also much higher after the start of the conflict.

4.2 Linear Regression and BERT

In this section we compare the result of the two supervised models for accuracy according to the regression task as a manner of avoiding future conflict. We demonstrate accuracy differences for both regressors at different threshold along with the initial error in Table 3.

Despite the differences in the size and content of the datasets, both models (LR and BERT) exhibit similar behaviors based on the results of the evaluation metrics. The MSE was quite small in both cases, but the lower MSE values in the Hamas–Israel conflict suggests that the model was able to achieve a better fit to the data as it had less errors. Similarly, for MAE, the lower the value indicates that the model also performed well with less errors, and the Hamas–Israel sets again performed better than on the Ukraine–Russia data.

In both scatter plots from Figures 3 (LR) and 4 (BERT), the majority of the data points cluster near the bottom-left, suggesting that the majority of the actual and predicted toxicity scores were low and closer to 0.2. For the LR model, as the actual toxicity scores increased, the Ukraine–Russia prediction scores was less likely to identify the increasing toxicity levels. This can be seen by the frequency of points that fell below the toxicity diagonal line when the actual toxicity scores were above 0.4. Thus, it can be understood that the LR model has a tendency to underestimate the magnitude of the toxicity scores, resulting in the prediction scores to be slightly lower than the actual toxicity scores.

[8] https://naturallang.com/conflict/conflict.html

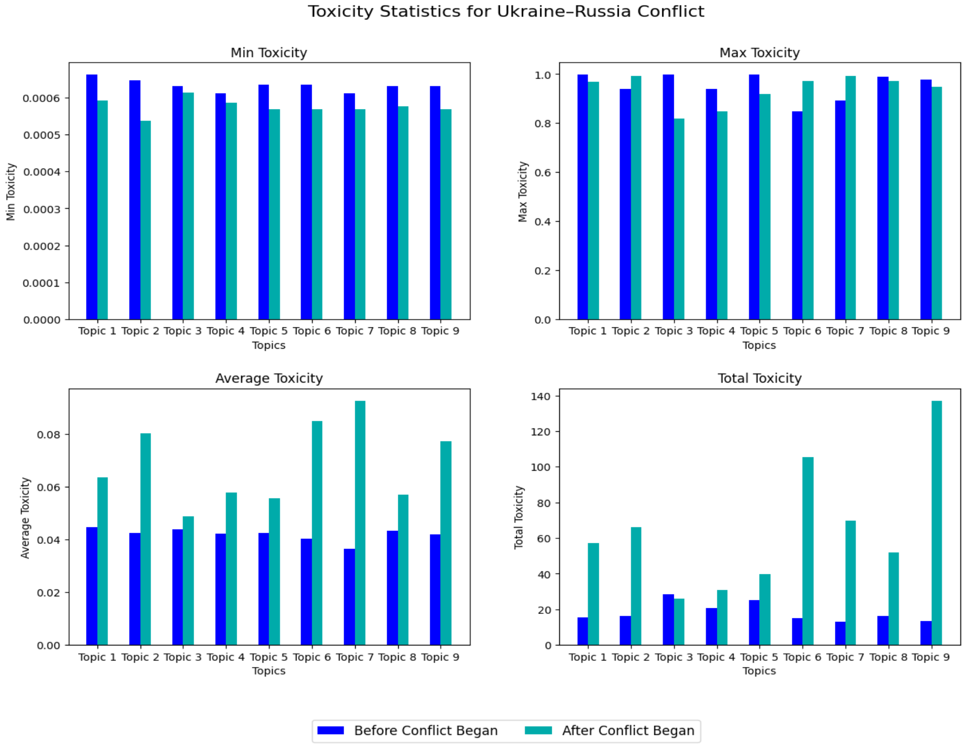


Figure 1: Ukraine–Russia minimum, maximum, average and total toxicity of topics created with Latent Dirichlet Allocation

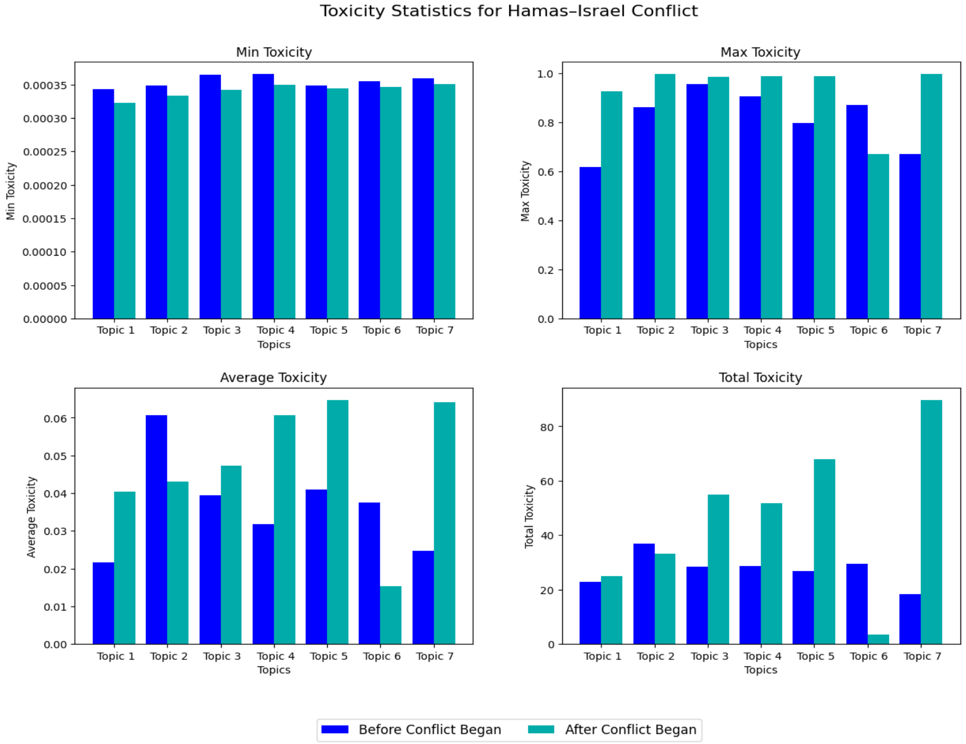
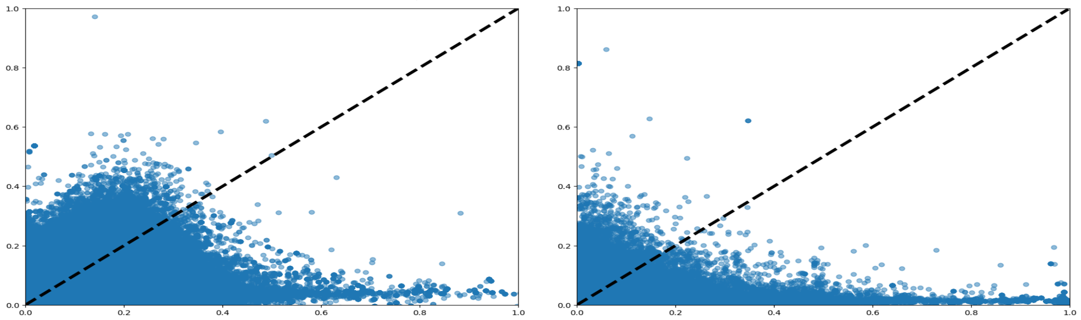


Figure 2: Hamas–Israel minimum, maximum, average and total toxicity of topics created with Latent Dirichlet Allocation

In comparison, the LR model based on the Hamas–Israel data performed better with fewer deviations than the Ukraine-Russia model, but still struggles in identifying the highest toxicity scores. In either case, this underestimation underlines the increase in toxicity that occurred after the conflicts.



(a) Ukraine–Russia

(b) Hamas–Israel

Figure 3: Prediction capability with Linear Regression on both conflicts using actual (x-axis) vs. predicted (y-axis) toxicity.

	Ukr–Rus	Ham–Isr
LR MSE	0.0124	0.0120
LR MAE	0.0753	0.0461
BERT MSE	0.0172	0.0144
BERT MAE	0.0805	0.0494

Table 3: Error (as low as 1.2%) for both the Linear Regressor and BERT models on predicting after-conflict toxicity.

Performance of the BERT model on the two conflicts is depicted in Figure 4. For the Ukraine-Russia content, while the BERT model performs worse when measured by error alone (see Table 3), its performance prediction based on the scatter plot exhibits a stronger central clustering tendency where predictions did not vary even as the actual scores changed. The better performance of our model on both datasets may speak to the relevant and informative topics extracted during the unsupervised learning portion of the investigation, as they would contribute to a better understanding of the text data. However, both models possessed a shared tendency to underestimate higher toxicity scores, as indicated by the fact that a majority of the points fell below the diagonal line. Both models would benefit from additional fine-tuning and other methods to improve feature representation. In particular, our model may benefit from further refinement during the unsupervised portion by altering the alpha and beta parameters, or using other forms of topic modeling to improve feature quality. We save those tasks for future work.

The LR models outperform BERT in our experiments. We believe that this can be attributed to the power of small models and their objective function that has to search a smaller, more distilled space. We chose the closest pre-trained language model to our data but it could be the case that other models BERT-based or hybrid models could outperform the LR.

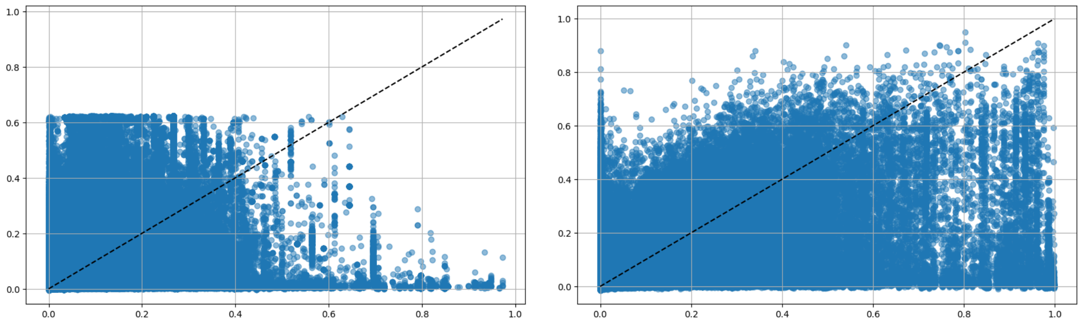
4.3 Accuracy Comparison and Thresholds

Various thresholds were evaluated to determine the accuracy of the model. We determined this to be the best form to measure accuracy on the level of classification alone. We believe that this would be beneficial for future use, and using one threshold over another can help balance the trade-offs between false positives and false negatives, depending on the objective of future tasks.

Based on the results in Figures 5 and 6, both models (LR and BERT) performed better as the threshold increased, allowing for more flexibility when it comes to determining what is considered a toxic post. For the Ukraine–Russia model, it appeared that the most optimal threshold value was the sum of the standard deviation and mean, or 0.157, and the optimal value for the Hamas–Israel model was the standard deviation of around 0.099. Hence, the optimal thresholds allow for a balance between identifying toxic posts without flagging non-toxic posts toxic or vice versa. These thresholds can serve as the foundation for further studies using more complex techniques to improve model reliability and accuracy. Integration of semantic analysis would also be beneficial to refine predictions that are over or under-looked using neural networks or other methods that are sensitive to complex patterns of language use.

5 Discussion

By incorporating LDA topic modeling, the model should have ability to detect how users' language



(a) Ukraine–Russia

(b) Hamas–Israel

Figure 4: Prediction capability with BERT on both conflicts using actual (x-axis) vs. predicted (y-axis) toxicity.

changes during times of crisis. We believe that the increase in total and average toxicity scores during the *unsupervised* method is reflective of the overall emotion and thoughts of social media users after a conflict has begun. For instance, in the Ukraine–Russia data, the top salient terms discussed Russian troops being stationed near the eastern border and NATO's involvement to curtail war, while the post-conflict discussions focused on detailed events from the conflict and user's reactions to those events. Moreover, toxicity of certain topics experienced a noteworthy growth in comparison to others; thus indicating that certain topics were more divisive and probably elicited a stronger emotional response from user. This was seen in the case of Topic 6 (https://naturallang.com/conflict/conflict.html) in the Hamas–Israel data which contained n-grams such as "war crime" before the conflict, but was more heavily discussed after the conflict began.

Furthermore, in the time leading up to the conflicts, we observed clear patterns that highlighted social media's role as an amplifier for pre-existing grievances and polarization. For the Hamas–Israel conflict, the discourse showed an increase in inflammatory content from both sides with terms like "islamic jihad" and "anti semite" to describe both sides. These terms and similar content displayed the growing distrust amid both parties, which work to feed narratives and feed existing tensions using phrases like "ethnic cleansing" and "human shield" to describe the interactions between both parties. On the other hand, the discussions prior to the onset of the Ukraine–Russia conflict also exhibited growing signs of distrust with terms like "russia invade ukraine" and "want war russia" within its rhetoric. Due to the posts being limited to English, it appeared that many of the comments painted Russia in negative light, but we would have had more conflicting perspectives had we included posts in Russian and Ukrainian.

The incorporation of LDA topics into our regression model grants it the ability to consider not only individual words, but also overarching themes expressed, making for a more comprehensive approach that enhances prediction accuracy. Our model's ability to accurately predict post-conflict toxicity scores from pre-conflict toxicity scores indicate that these social media discussions contain early indications of unrest. While an increase in polarizing content and grievances surrounding a particular topic may not always lead directly to escalations, this toxic content can exacerbate tensions and make the conflict more likely. This would mean that governments and NGOs can monitor situations and topics that that signal growing unrest or societal division, and be immediately alerted when signs of escalation becomes prevalent and its associated toxicity levels reach a predefined critical point that could signify an increased likelihood of a conflict taking place. Furthermore, policymakers and social media platforms can use this predictive tool to gain an understanding into the language and behavioral patterns and language being used in response to events like elections or international crises in real-time.This would give policymakers and authorities the ability to address the grievances, trigger diplomatic interventions, and other peacekeeping measures to mitigate the ongoing tensions.

Further optimizations can be implemented both by governments and social media platforms to prevent a conflict from arising. This could mean that the model would be helpful in thematically and geographically pinpointing where online toxicity

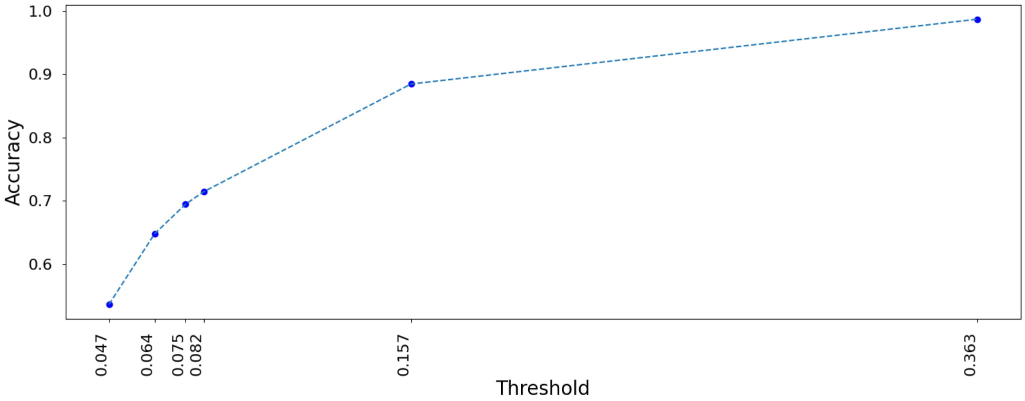


Figure 5: Accuracy thresholds for Ukraine-Russia conflict.

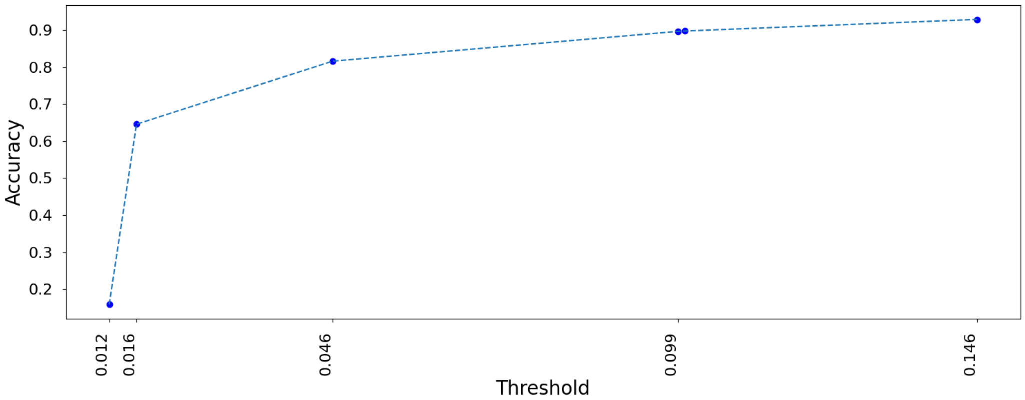


Figure 6: Accuracy thresholds for Hamas–Israel conflict.

is concentrated. For instance, if a certain region or group engages in more toxic content, the model would be able to pinpoint these areas as potential conflict zones and communities experiencing growing unrest. Social media platforms can also work to provide warning signs to users and strengthen moderation efforts in stances where a conflict is likely to occur. This could manifest in posts with high predicted toxicity scores to be flagged for review by human moderators and hidden from public view. In fact, developers may be able to tailor these interventions for individual users based on their predicted toxicity score in the form of warnings or temporary suspensions. To maintain engagement, these platforms can instead implement methods by elevating the voices of experts in a specified topic to prevent the spread of misinformation, and discourage instances of hate speech with customized interventions before it can incite violence.

6 Conclusion

Through the implementation of unsupervised and supervised machine-learning models, we have explored and observed how social media interactions can predict the escalation of two major conflicts. Particularly in times of crisis, negative sentiments and extremist perspectives are amplified on platforms like Twitter and Reddit. Furthermore, the limited regulation and addictive nature of these algorithms make these platforms effective tools for spreading misinformation and swaying public opinion, making them a catalyst for conflicts. With further fine-tuning and optimization, our models should have the ability to effectively predict a rise in toxicity in user interactions in real time. Such improvements will help policymakers and social media platforms obtain a better grasp of the dynamics of social media leading up to and during a conflict. What is more, they can help in developing frameworks to mitigate hostility with customized content moderation, and even predict disputes before they can occur. In particular, prior knowledge of a conflict is pivotal as it gives policymakers or other leaders the opportunity to act appropriately, and even formulate the proper measures to maintain peace and prevent the escalation of violence.

7 Limitations

Our results show that an uneven distribution of toxicity scores can heavily impact performance. In our experiments, this was most evident in the low MSE and MAE values for the Ukraine–Russia models despite being unable to properly distinguish the toxicity scores higher than 0.4, and would only be the case if the majority of data points were predicted to be low and their actual toxicity scores were low. This led to the Ukraine–Russia models having a tendency to bias towards lower toxicity scores in its predictions. Likewise, while the Hamas–Israel models performed better overall, they also experienced difficulty in the upper range, which further points to the too few high-toxicity examples. It is likely that all of the models' performances would improve if trained on a balanced training set to allow the models to effectively capture the nuances in the relationship between the text and their toxicity scores.

Additionally, the settings for minimum document frequency in the vectorization process may have negatively impacted the toxicity scores. The point of setting the minimum document frequency is to ensure that the vectorizer would extract important terms that will serve as predictors by filtering out excess noise. On the other hand, not sufficiently adjusting the maximum document frequency may have allowed overly frequent terms to dominate the feature set, further obscuring meaningful analysis. This was definitely the case as some of the terms in the topics were unrelated with the Ukraine-Russia content containing mentions of cryptocurrency and the Hamas–Israel content containing references to actions related to the platform. Correcting these thresholds could help eliminate this noise and enhance the model's ability to perform a more nuanced toxicity analysis.

Another potential reason for the models' performance was the variation in the number of samples in the training and testing sets. Since we were using pre-existing datasets, we were limited to what was available in only that dataset. The post-war datasets were significantly larger than the pre-war datasets, and likely may have compromised the models' ability to generalize based on their training set. This size mismatch likely affected the models' performance.

8 Acknowledgements

We would like to express our most sincerest gratitude for equipment and space contributions from Hofstra University. Additionally, we highly acknowledge Northeastern University for continued funding and support with other logistics. Lastly, website hosting and servers were provided by the owner and operators of https://www.naturallang.com.

Ethical Considerations

We have not used any human subjects for our experimentation. Nor do we express any opinion on the two conflicts studied.

References

Asaniczka. 2024. Daily public opinion on israel-palestine war.

Serpil Aslan. 2023. A deep learning-based sentiment analysis approach (mf-cnn-bilstm) and topic modeling of tweets related to the ukraine–russia conflict. *Applied Soft Computing*, 143:110404.

BwandoWando. 2024. (sunset) ukraine conflict twitter dataset.

Bledi Celiku and Aart Kraay. 2017. *Predicting Conflict*. The World Bank.

Philip Chang, Ying-Tzu Yu, Abraham Sanders, and Thilanka Munasinghe. 2023. Perceiving the ukraine-russia conflict: Topic modeling and clustering on twitter data. In *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 147–148.

Kai Chen, Zihao He, Keith Burghardt, Jingxin Zhang, and Kristina Lerman. 2024. Isamasred: A public dataset tracking reddit discussions on israel-hamas conflict. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):1900–1912.

Kenneth Church, Annika Schoene, John Ortega, Raman Chandrasekar, and Valia Kordoni. 2022. Emerging trends: Unfair, biased, addictive, dangerous, deadly, and insanely profitable. *Natural Language Engineering*, 29:1–26.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Ivan Dylko, Igor Dolgov, William Hoffman, Nicholas Eckhart, Maria Molina, and Omar Aaziz. 2018. Impact of customizability technology on political polarization. *Journal of Information Technology & Politics*, 15(1):19–33.

Madjid Erroukrma. 2023. Palestine tweet data 2022 september december.

Amisha Gangwar and Tanvi Mehta. 2023. Sentiment analysis of political tweets for israel using machine learning. In *Machine Learning and Big Data Analytics*, pages 191–201, Cham. Springer International Publishing.

Hudson Golino, Alexander P. Christensen, Robert Moulder, Seohyun Kim, and Steven M. Boker. 2021. Modeling latent topics in social media using dynamic exploratory graph analysis: The case of the right-wing and left-wing trolls in the 2016 us elections. *Psychometrika*, 87(1):156–187.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Xinlei He, Savvas Zannettou, Yun Shen, and Yang Zhang. 2024. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 770–787.

Jonas Kaiser and Adrian Rauchfleisch. 2020. Birds of a feather get recommended together: algorithmic homophily in youtube's channel recommendations in the united states and germany. *Social Media + Society*, 6:205630512096991.

Clara Maathuis and Iddo Kerkhof. 2023. The first two months in the war in ukraine through topic modeling and sentiment analysis. *Regional Science Policy & Practice*, 15(1):56–74.

Shubhanshu Mishra, Shivangi Prasad, and Shubhanshu Mishra. 2020a. Trained models for Multilingual Joint Fine-tuning of Transformer models for identifying Trolling, Aggression and Cyberbullying at TRAC 2020.

Sudhanshu Mishra, Shivangi Prasad, and Shubhanshu Mishra. 2020b. Multilingual joint fine-tuning of transformer models for identifying trolling, aggression and cyberbullying at TRAC 2020. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 120–125, Marseille, France. European Language Resources Association (ELRA).

Trisiladevi C. Nagavi and Aishwarya D. S. 2021. Detection and classification of toxic content for social media platforms. In *2021 4th International Conference on Recent Developments in Control, Automation & Power Engineering (RDCAPE)*, pages 368–373.

Kari Karppinen Natali Helberger and Lucia D'Acunto. 2018. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, 21(2):191–207.

Nurlela, Muhammad Ali Ramdhani, Dian Sa'adillah Maylawati, Undang Syaripudin, Eva Nurlatifah, and Rifqi Syamsul Fuadi. 2023. Sentiment analysis on the issue of the palestine-israel conflict on twitter using the convolutional neural network algorithm. In *2023 9th International Conference on Wireless and Telematics (ICWT)*, pages 1–6.

Daria Purtova. 2022. Russia-ukraine war - tweets dataset (65 days).

Salim Sazzed. 2022. The dynamics of ukraine-russian conflict through the lens of demographically diverse twitter data. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 6018–6024.

Amit Sheth, Valerie L. Shalin, and Ugur Kursuncu. 2022. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490:312–318.

Mohammed Taleb, Alami Hamza, Mohamed Zouitni, Nabil Burmani, Said Lafkiar, and Noureddine En-Nahnahi. 2022. Detection of toxicity in social media based on natural language processing methods. In *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pages 1–7.

Abhishek V Tatachar. 2021. Comparative assessment of regression models based on model evaluation metrics. *International Research Journal of Engineering and Technology (IRJET)*, 08:853–860.

Fenna Van Nes, Tineke Abma, Hans Jonsson, and Dorly Deeg. 2010. Language differences in qualitative research: is meaning lost in translation? *European journal of ageing*, 7:313–316.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 218–226, New York, NY, USA. Association for Computing Machinery.

Thomas Zeitzoff. 2017. How social media is changing conflict. *The Journal of Conflict Resolution*, 61(9):1970–1991.

Thomas Zeitzoff. 2018. Does social media influence conflict? evidence from the 2012 gaza conflict. *Journal of Conflict Resolution*, 62(1):29–63.