Dynamic Topic Modelling of Online Discussions on the Russian War in Ukraine

Taras Ustyianovych[1], Nadiia Kasianchuk[2,3], Halina Falfushynska[4,6,7], Solomiia Fedushko[5,8] and Eduard Siemens[6]

[1]*Department of Artificial Intelligence Systems, Lviv Polytechnic National University, Kniazia Romana Str. 5, Lviv, Ukraine*
[2]*Faculty of Biology, Adam Mickiewicz University, Uniwersytetu Poznańskiego Str. 6, Poznań, Poland*
[3]*Faculty of Pharmacy, Bogomolets National Medical University, Taras Shevchenko Str. 13, Kyiv, Ukraine*
[4]*Department of Marine Biology, Institute for Biological Sciences, University of Rostock, Albert Einstein Str. 3, Rostock, Germany*
[5]*Department of Social Communication and Information Activities, Lviv Polytechnic National University, Kniazia Romana Str. 5, Lviv, Ukraine*
[6]*Department of Electrical, Mechanical and Industrial Engineering, Anhalt University for Applied Sciences, Bernburger Str. 55, Köthen, Germany*
[7]*Statistical Analysis Center, Wroclaw Medical University, K. Marcinkowski Str. 2-6, Wroclaw, Poland*
[8]*Department of Information Systems, Faculty of Management, Comenius University in Bratislava, Odbojárov Str. 10, Bratislava, Slovakia*

taras.o.ustyianovych@lpnu.ua, nadkas2@st.amu.edu.pl, halina.falfushynska@uni-rostock.de, solomiia.s.fedushko@lpnu.ua, eduard.siemens@hs-anhalt.de

Keywords: Machine Learning Operations (MLOps), Social Media Discussions, Russian War in Ukraine, Splunk Enterprise, Latent Dirichlet Allocation (LDA).

Abstract: The availability of robust end-to-end ML processes plays a crucial role in delivering an accurate and reliable system for real-time text data inference. In this paper, we present an approach to building machine learning operations (MLOps) and an observability application to perform topic modelling of online discussions in social media, here observed based on topics and threads related to the Russian war in Ukraine. Splunk Enterprise is the main tool and platform used throughout this research with its knowledge discovery, dashboarding, and alerting. 30GB of social media text data coming from a Russian social network VKontakte over the time line January 2022 to May 2023. Main inquiries included text mining and topic modelling, which we managed to perform over the observation period using Python frameworks, mainly gensim for text processing and MLflow for experiment management and logging. The Splunk architecture allowed us to ingest and analyse the results and prediction of ML experiments for dynamic topic modelling, and served as a MLOps solution. The designed set of five dashboards played a crucial role in determining the optimal model hyperparameters (number of topics, A-priori belief on document-topic distribution, number of total corpus passes) and drift detection which occurred almost every two-three weeks depending on the phase of the war. Our application assisted us with text analysis, discovering how events on the battlefield influenced social media discussions, and what post attributes contributed to a high user engagement. With our setup we were able to find out how antiwar hashtags have been used to promote misleading content actually supporting the war against Ukraine. The analysis of the researched discussions shows a trend where usage of adjectives decreased over time since the war has started, whereas an increase for nouns and verbs usage over time. Information distortion has steadily been present in the content leading to bias and misleading data in social media discussions.

1 INTRODUCTION

The rapid proliferation of social media platforms has transformed the way information circulates, impacting everything from daily conversations to global events. As these platforms become central hubs of information exchange, they also become fertile grounds for the spreading fakes, so called trolls activities and targeted propaganda campaigns. The

Russian war in Ukraine as a significant geopolitical event can be used as a major example where, amidst the vast amounts of discussions, opinions, and perspectives across various social media platforms, instances of misinformation, orchestrated facts manipulation and down to bare propaganda campaigns can bias the perception of social media discussion For this, several sources of Russian platforms have been observed [1]. This immense amount of real-time data, which is a mixture of genuine insights and manipulated narratives is more and more used as a decision-making base by policymakers, managers of businesses and even by researchers. Analyzing the data in real-time is required to find trends, categorize information into meaningful pieces, map inter-dependencies, and model social media discussions, which can be efficiently accomplished using machine learning capabilities. Therefore, processing raw data through computational algorithms would eliminate the toil, and let analysts and researchers focus on knowledge mining of organized information with a drilldown capability to dive into the original data.

However, harnessing this vast volume of data and extracting meaningful insights from it in real-time presents several challenges. Conventional data analysis tools and methodologies often fall short when faced with the dynamic nature of social media content, which is characterized by its heterogeneity, volume, and velocity [2]. Machine Learning Operations (MLOps) - a set of practices, guidelines, and tools - promises to bridge this gap [3]. MLOps, when combined with robust end-to-end machine learning processes, enables the development of systems that can efficiently process, analyse, and reveal insights from massive streams of real-time text data. The implemented MLOps-powered system improved the observability of topic models, assisted with data drifts detection, and determining the frequency for retraining. We found out that model training time increases with the number of topics for the text to be split into. The analyzed metrics used to evaluate the model performance are correlated in such a way: perplexity has a 79.5% correlation with UMass coherence score and 71.8% with CV coherence scores. The correlation between these two coherence scores is just 27.1%. This confirms that decision for the optimal model hyperparameters can be made in conjunction with these three metrics.

According to the CRISP-ML(Q) model [4], effective monitoring of machine learning operations is as vital as the development of the model itself [3]. This is particularly true in dynamic environments with frequent data change and drift is paramountto maintaining high standards for performance and knowledge mining in such settings.

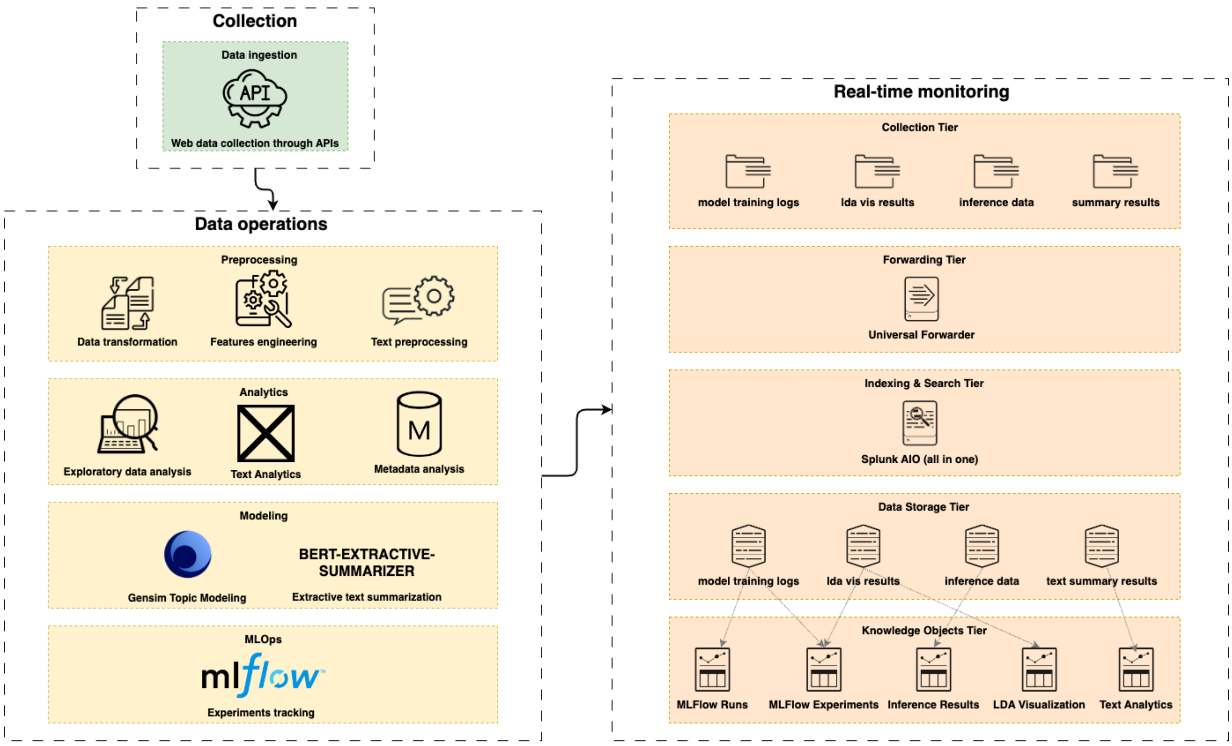


Figure 1: Workflow architecture of the machine learning operations process.

In this paper, we improve the efficiency and user-experience of processing and analyzing vast amounts of social media discussions. Our methodology, underpinned by advanced text mining techniques and topic modelling, not only aids in knowledge discovery but also equips us with the tools to detect model drifts, ensuring the continual relevance and accuracy of our insights. As we navigate through the intricacies of our solution, we also shed light on the importance of real-time monitoring, the challenges posed by the dynamic environment of social media, and the invaluable insights that can be drawn from such an endeavour.

2 MATERIALS AND METHODS

Machine Learning operations and their monitoring are intricately tied to the specific modelling use-case in question. CV and UMass coherence scores and perplexity were used in this research to evaluate the performance of the trained topic models [5]. Similarly, the optimization of hyperparameters varies across algorithms, contingent upon the model architecture. In this context, we delve into the monitoring and evaluation of the *Latent Dirichlet Allocation (LDA)* algorithm, trained in Python using the *gensim* library, to automatically model topics within social media data pertinent to geopolitical events [6, 7]. The LDA algorithm is favoured by numerous researchers due to its efficiency and adaptability. Recent studies underscore its prominence as a generative text model [6]. Given the data's nature, it's imperative to continuously update knowledge in real-time and remain vigilant to a significant number of outliers or any signs of performance degradation. Upon spotting any of these indicators, swift detection and proactive measures are essential for being able to observe and to identivy trends and biases in the subject under study [8]. The presence of robust MLOps becomes even more critical for high-load production systems. Disruptions in service level objectives or agreements (SLOs and SLAs) can have detrimental effects on the system reliability and performance [9].

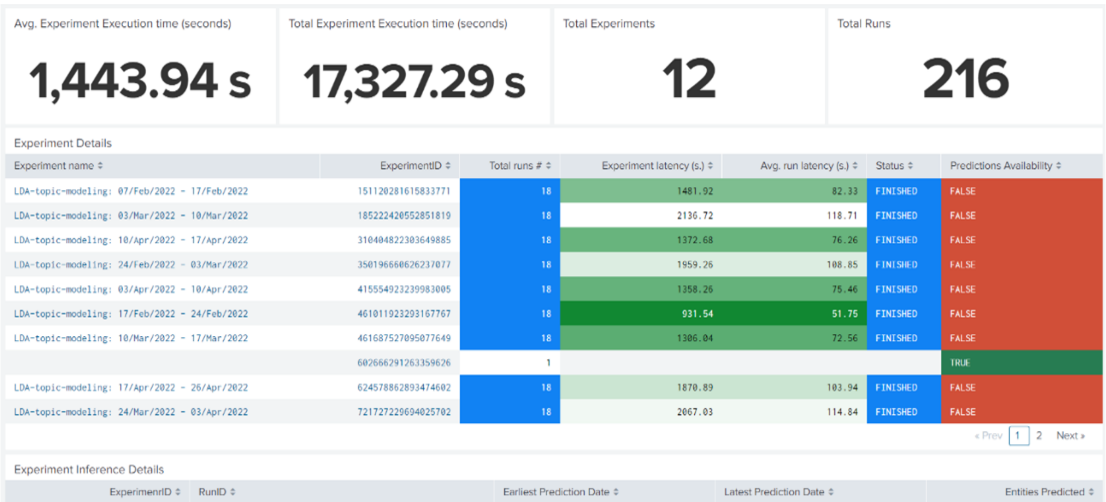
Having these prerequisites in mind, the architecture depicted in the Figure 1 has appeared conceiving.

According to the architecture diagram (Figure 1), there are three primary layers, namely:

- data collection;
- data operations;
- real-time monitoring.

To understand the use-case and implementation, a concise description of each layer is provided.

Data collection has been performed using officially available APIs from *VKontakte* social media platform to retrieve a discrete number of posts based on a specified query for the period since January 2022 until May 2023. The following keyword parameters were specified in the search query to retrieve the most relevant pieces of information related to the actual Russian war in Ukraine: *"war"*, *"specoperation"*, and *"Ukraine"*. The returned data undergoes a standard procedure involving cleaning, transformations, feature engineering, and text pre-processing. The latter is



Avg. Experiment Execution time (seconds)	Total Experiment Execution time (seconds)	Total Experiments	Total Runs
1,443.94 s	17,327.29 s	12	216

Experiment Details

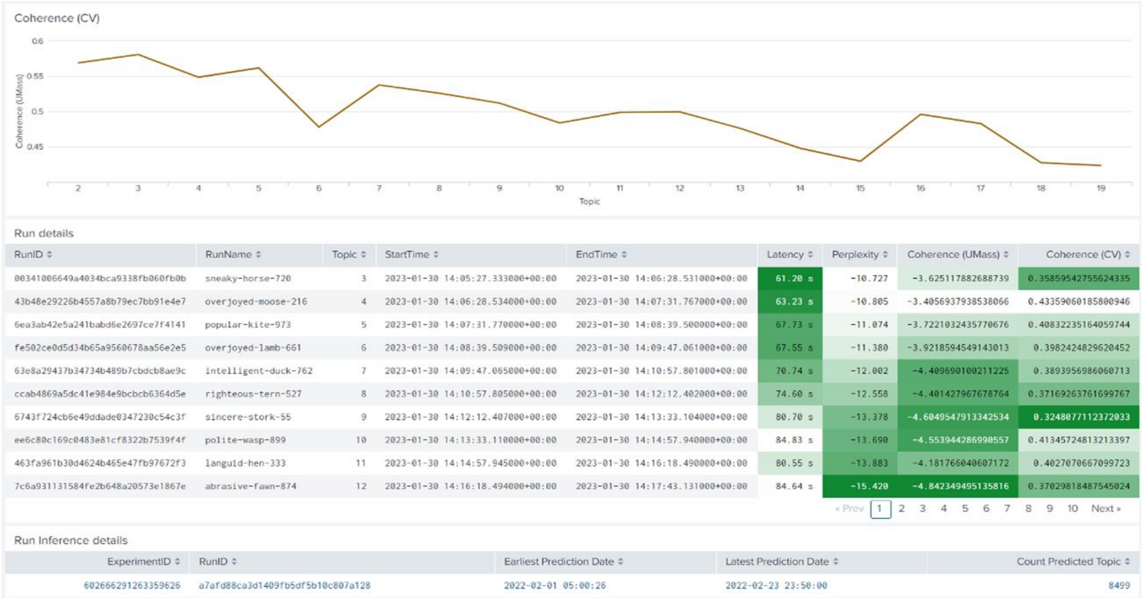
Experiment name	ExperimentID	Total runs #	Experiment latency (s.)	Avg. run latency (s.)	Status	Predictions Availability
LDA-topic-modeling: 07/Feb/2022 - 17/Feb/2022	151120281615833771	18	1481.92	82.33	FINISHED	FALSE
LDA-topic-modeling: 03/Mar/2022 - 10/Mar/2022	185222420552851819	18	2136.72	118.71	FINISHED	FALSE
LDA-topic-modeling: 10/Apr/2022 - 17/Apr/2022	310404822303649885	18	1372.68	76.26	FINISHED	FALSE
LDA-topic-modeling: 24/Feb/2022 - 03/Mar/2022	350196660626237077	18	1959.26	108.85	FINISHED	FALSE
LDA-topic-modeling: 03/Apr/2022 - 10/Apr/2022	415554923239983005	18	1358.26	75.46	FINISHED	FALSE
LDA-topic-modeling: 17/Feb/2022 - 24/Feb/2022	461011923293167767	18	931.54	51.75	FINISHED	FALSE
LDA-topic-modeling: 10/Mar/2022 - 17/Mar/2022	461687527095077649	18	1306.04	72.56	FINISHED	FALSE
	602666291263359626	1				TRUE
LDA-topic-modeling: 17/Apr/2022 - 26/Apr/2022	624578862893474602	18	1870.89	103.94	FINISHED	FALSE
LDA-topic-modeling: 24/Mar/2022 - 03/Apr/2022	721727229694025702	18	2067.03	114.84	FINISHED	FALSE

« Prev 1 2 Next »

Experiment Inference Details

ExperimenrID	RunID	Earliest Prediction Date	Latest Prediction Date	Entities Predicted

Figure 2: Visual representation of MLFlow experiments, detailing run metrics, execution times, and the relationship between.



Run details

RunID	RunName	Topic	StartTime	EndTime	Latency	Perplexity	Coherence (UMass)	Coherence (CV)
00341006649a4034bca9338fb060fb0b	sneaky-horse-720	3	2023-01-30 14:05:27.333000+00:00	2023-01-30 14:06:28.531000+00:00	61.20 s	-10.727	-3.625117882688739	0.35859542755624335
43b48e29226b4557a8b79ec7bb91e4e7	overjoyed-moose-216	4	2023-01-30 14:06:28.534000+00:00	2023-01-30 14:07:31.767000+00:00	63.23 s	-10.805	-3.4056937938538066	0.43359060185800946
6ea3ab42e5a241babd6e2697ce7f4141	popular-kite-973	5	2023-01-30 14:07:31.770000+00:00	2023-01-30 14:08:39.500000+00:00	67.73 s	-11.074	-3.7221032435770676	0.40832235164059744
fe502ce0d5d34b65a9560678aa56e2e5	overjoyed-lamb-661	6	2023-01-30 14:08:39.509000+00:00	2023-01-30 14:09:47.061000+00:00	67.55 s	-11.380	-3.9218594549143013	0.398242482962045 2
63e8a29437b34734b489b7cbdcb8ae9c	intelligent-duck-762	7	2023-01-30 14:09:47.065000+00:00	2023-01-30 14:10:57.801000+00:00	70.74 s	-12.002	-4.409690100211225	0.3893956986060713
ccab4869a5dc41e984e9bcbcb6364d5e	righteous-tern-527	8	2023-01-30 14:10:57.805000+00:00	2023-01-30 14:12:12.402000+00:00	74.60 s	-12.558	-4.401427967678764	0.37169263761699767
6743f724cb6e49ddade0347230c54c3f	sincere-stork-55	9	2023-01-30 14:12:12.407000+00:00	2023-01-30 14:13:33.104000+00:00	80.70 s	-13.378	-4.6049547913342534	0.3248077112372033
ee6c80c169c0483e81cf8322b7539f4f	polite-wasp-899	10	2023-01-30 14:13:33.110000+00:00	2023-01-30 14:14:57.940000+00:00	84.83 s	-13.690	-4.553944286990557	0.41345724813213397
463fa961b30d4624b465e47fb97672f3	languid-hen-333	11	2023-01-30 14:14:57.945000+00:00	2023-01-30 14:16:18.490000+00:00	80.55 s	-13.883	-4.181766040607172	0.4027070667099723
7c6a931131584fe2b648a20573e1867e	abrasive-fawn-874	12	2023-01-30 14:16:18.494000+00:00	2023-01-30 14:17:43.131000+00:00	84.64 s	-15.420	-4.842349495135816	0.37029818487545024

« Prev 1 2 3 4 5 6 7 8 9 10 Next »

Run Inference details

ExperimentID	RunID	Earliest Prediction Date	Latest Prediction Date	Count Predicted Topic
602666291263359626	a7afd88ca3d1409fb5df5b10c807a128	2022-02-01 05:00:26	2022-02-23 23:50:00	8499

Figure 3: A detailed visualisation of run metrics, emphasizing the correlation between the number of topics and run latency, performance scores, and the exploration of hyperparameters across different runs for optimal model training.

particularly crucial as it serves as input for the LDA model. Subsequently, metadata and consolidated statistics related to the social media posts are stored and passed to the next layer. Text vectorization and model training have been conducted using the *gensim*, *spaCy*, and *NLTK* Python-libraries. For hyper-parameter tuning, experiment management, and subsequent model deployment, the open-source library *MLFlow* has been used.

The monitoring artefacts in the third layer, storing the output from the previous step, are split into four categories, see Figure 1, right part:

- Model training logs that include information about hyperparameters, training duration, experiment and run IDs, metrics, and additional metadata.
- LDA visualization results, which comprise JSON files needed to reconstruct and depict the determined topics of the best-performing model on a 2D Cartesian plot. The outcomes from this category aid decision-making during the model selection process post hyperparameter tuning.
- The inference data category preserves post data, the label assigned by the model, probability scores, and the principal component values of the input documents' vectors.
- Summary results encapsulate text summaries for each identified topic.

Each category contains common keys, such as experiment and run IDs, derived from the preceding step when leveraging MLFlow's logging features.

Once the output data is systematically categorized, the Splunk ecosystem is employed. The monitoring layer's implementation predominantly utilizes Splunk Enterprise, a comprehensive data platform, designed for real-time information indexing, storage, and processing. Splunk was selected due to its developer-centric features and adaptable search processing language.

A universal forwarder, a streamlined version of Splunk Enterprise equipped with data forwarding functionalities, performs real-time data ingestion to a a dedicated Splu nk Enterprise instance tasked with indexing and searching operations. Each index retains data corresponding to its category. For example, model training logs are exclusively ingested into the model training index, while inference data is confined to the inference data index. Leveraging IDs consistent across indexes enables the execution of joined operations and other data-related tasks.

The concluding phase of the third layer encompasses data visualization via interactive dashboards equipped with a variety of dropdowns and filter options. Additionally, for enhanced monitoring, alerts can be configured to detect model performance degradation and identify anomalies.

In summary, five dashboards, termed as "knowledge objects," were formulated within Splunk Enterprise. These dashboards interrelate and collectively aim to offer real-time insights into MLOps and the intricate analysis of social media data.

3 RESULTS AND DISCUSSIONS

This section describes the derived artifacts, specifi findings their significance along with suggestions for enhancements in future research. We have created the following knowledge objects to aid MLOps and observability implementation:

- Experiments
- Runs
- Topic Modelling Visualization
- Inference Results
- Text Analytics

We will sequentially discuss each of these knowledge objects to elucidate their value, challenges during their development, and inherent limitations.

The Experiments dashboard (Figure 2) showcases information corresponding to a unit within the MLFlow architecture. An experiment comprises multiple runs, where a run is a single execution of a script with defined parameters. In our scenario, the experiment pertains to training a topic modelling model over a specified duration. It's essential to consider hyperparameters like the number of topics, alpha score, and iterations over the dataset to achieve optimal results. Each run encapsulates a specific set of hyperparameters and, upon completion, yields metrics for evaluation and comparison.

The aforementioned dashboard displays summarized statistics about various experiments. It provides filtering options based on time window, experiment name, and ID. The dashboard's single-value panels convey details about the average and total execution time, number of experiments, and runs. The table presents each experiment's comprehensive attributes: its name, unique identifier, number of runs, total and average execution latency per run, status, and an indicator for the availability of prediction results. For each experiment, the optimal run is selected, and its hyperparameters are employed for the final topic modelling. The dashboard contains a table with details about the best run, and if inference results exist, one can navigate to this dashboard to evaluate the topic modelling outcomes by clicking on the selected run.

Evaluating the set hyperparameters of each run within an experiment necessitates comparing performance metrics. The runs dashboard (Figure 3) empowers engineers and analysts to execute this task. Leveraging Splunk's drill-down features, transitioning from Experiments to Runs knowledge objects significantly enhances the user experience. The latter dashboard is prefiltered based on the selected experiment ID from the previous dashboard. Run latency is a crucial consideration during model training. The Runs dashboard displays a bar chart and summary statistics for runs' execution times. A discernible pattern emerges: a higher number of topics correlate with increased run latency. Perplexity and coherence scores are plotted against the number of topics using line charts, facilitating easy correlation of metrics for a given topic count. The same piece of information is replicated below these line charts in a table "Run details" with supplementary metadata about specific runs. The built-in capabilities for table formatting afford a comprehensive view of performance metrics, aiding analysts in their comparison. Furthermore, the dashboard enables the identification of optimal hyperparameters, which can be adjusted in subsequent runs.

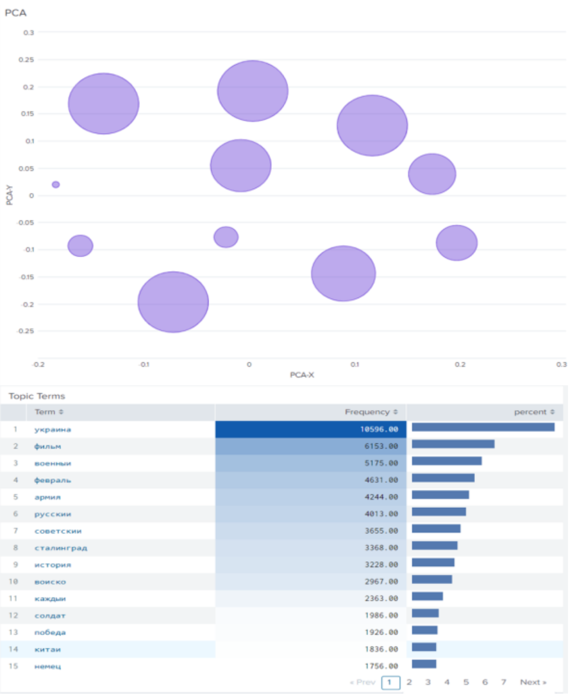


Figure 4: Interactive Topic Modelling Visualization. Displaying modelled topics on a Cartesian plot using the LDA algorithm. The bubble chart demonstrates topic distributions, with size indicating document frequency and proximity suggesting potential topic similarities. The layout aids in discerning term commonalities across topics and highlights the LDA's emphasis on word frequency over sentiment analysis.

After selecting the best runs, the topic modelling model with the most suitable hyperparameters can be trained for inference. Topic modelling objective is to categorize a text dataset into meaningful groups to extract knowledge.

The Topic Modelling Visualization (Figure 4) dashboard provides a suite of panels and filters to:

- Determine the placement of each modelled topic on a Cartesian plot and its relationship to other topics.
- Analyse a topic's keywords and terms.
- Identify common keywords and terms across the modelled topics.

We employed the LDA algorithm for topic modelling. The visual formatting of the panels aids decision-making. For instance, a bubble chart (Figure 4) depicts topics on a Cartesian plot, with larger bubbles indicating topics with more documents. Closely situated or overlapping bubbles might suggest topic similarity, possibly indicating that the topic count is too high. The table with topics terms below shows the most crucial keywords that describe the processed messages. The top terms shown on Figure 4 are “Ukraine”, “movie”, “military” (adjective), “February”, “army”, “Russian”, “soviet”, “Stalingrad”, “history”, “military” (noun), “soldier”, “victory”, “China”, “German”. The frequency and percent columns indicate the quantity of these keywords’ appearance in the input documents. We can conclude most of the collected data based on the input search parameters is related to 1) the war in Ukraine, 2) historical battles during the Second World War, 3) geopolitical situation in the world.

Hence, such visualizations are invaluable, facilitating efficient task completion. Similarly, if specific terms frequently appear in multiple topics, it's a clear sign of their relatedness. This observation highlights an LDA limitation: it doesn't capture sentiment and relies on word counts in the training text data.

The Inference Results dashboard (Figure 5), the fourth one, is designed to showcase predictions made by the model with the best hyperparameters from a given experiment. The data for this dashboard shares common keys, such as experiment and run IDs, enabling linkage with other datasets. Consequently, filters for these IDs are present on the Inference Results dashboard. Primary panels include:

- Document count by topics.
- Average, minimum, and maximum distances to the topics' centroids for the documents.
- Probability box plots to display the likelihood of documents belonging to the predicted topic.
- t-SNE visualization of the predicted document-to-topic associations.
- A table with summary statistics about each topic's distance to its centroid, aiding in identifying topics with numerous outliers.

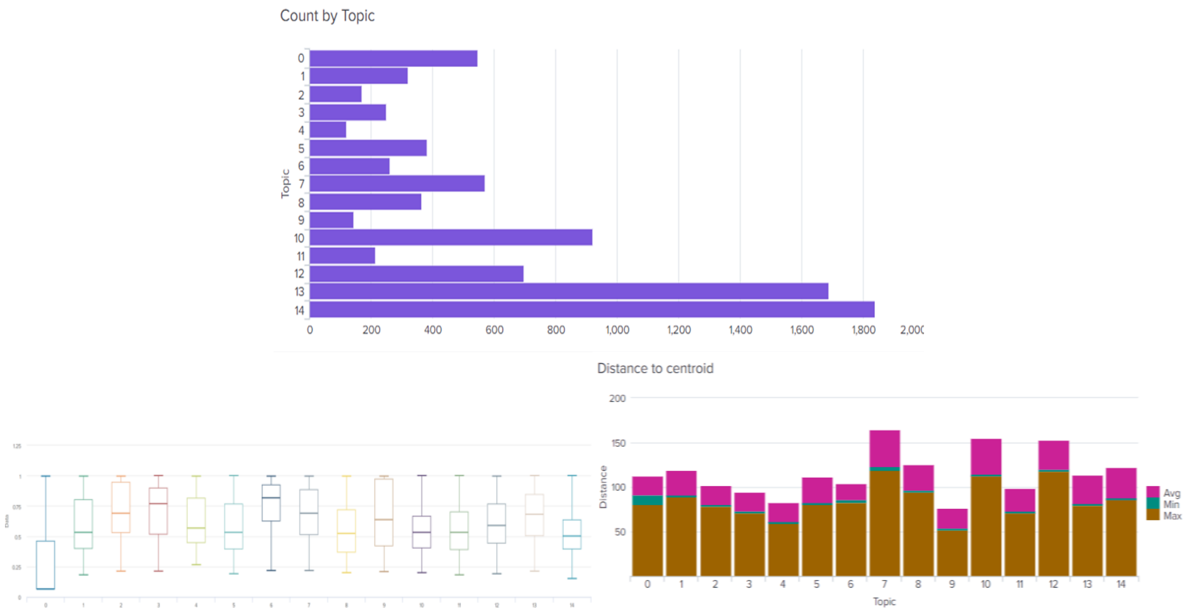


Figure 5: Inference Results Overview. An in-depth display of model predictions based on optimal hyperparameters. This dashboard visualizes topic distributions, document-to-topic distances, and prediction likelihoods. It aids in the real-time evaluation of model performance, detection of outliers, and understanding topic dynamics, offering insights into both broad and specific textual representations within the dataset.

Another objective of the Inference Results dashboard (Figure 5) is to detect the trained model's performance degradation. Indications such as a significant number of outliers or declining probabilities of a document belonging to the predicted topic over time are potent signals of such scenarios. Based on our empirical evaluation, if a topic comprises more than 10% outliers among its documents, the model's performance has diminished, rendering it less adept at predicting the correct topic for unseen documents. Such observations were spotted usually on a regular basis ranging between two to four weeks, and are known as the topic survival time.

We present a categorization of topics from the Inference Results dashboard in Table 1.

We observed that topics with more considerable centroid distances tend to share common terms and keywords. Such topics offer a broader representation of the entire text dataset rather than focusing on specific facets. However, this doesn't render them invaluable. These instances exemplify generic meanings and aid in extracting such text entities. Similarly, the Inference Results panels can assist researchers in determining irrelevant topics and related documents, which might be excluded for other exploration tasks.

Text Analytics, the final dashboard of the developed Splunk application, consists of word cloud visualizations, analyses of extracted hashtags, and named entities. Users can interactively select documents related to particular topics, sorted by engagement metrics like views, likes, shares, and comments.

In this case, the designed dashboards were populated with data from social media discussions about the Russian war in Ukraine. The solution's primary benefit lies in its capability for dynamic topic modelling, swift outlier identification, and signs for model retraining. We monitored how discussion topics and used keywords evolved over time and determined the main patterns. Discussions before the invasion (January-February 2022) can be characterized by such keywords used: “accusations” “threats”, “tension”, “uncertain”. Once the full-scale invasion began, we noticed a drastic increase in the number of generated posts. The main terms related to the Russian war in Ukraine were “betrayal", "surrender", "loss of opportunities". Also, historical figures were frequently mentioned. The third period, after more than a year of war, has seen a slight decrease in adjectives usage. This finding means the sentiment of the messages has tended to become neutral rather than emotionally appellative. We observe more and more nouns in the formed topics during this period. Overall, the comparison of the three periods shows that the war has had a significantly shaped the way social media discussions evolve. The war has also led to a sharp increase in the generated and shared content related to this subject. However, there are some signs that the tone of discussions about the war may be changing, as the conflict drags on. This solution has proven for real-time social media analysis and exploration. The incorporation of Splunk's inherent capabilities and add-ons broadens the visualization range and tools for data interaction and visualization.

Table 1. Categorization of the topics indices introduced at the Inference Results dashboard.

Category	Topics	Mined knowledge
Russian War in Ukraine	4, 9, 11, 14.	These topics have the highest number of documents mainly because of the queries specified during data collection. The distance to centroids is low indicating minimum outliers. This simplifies further detection of related documents.
Local subjects	3, 8, 12	The category posts describe internal news happening within the country and region. The documents belonging to these topics are classified with high probability, which is explained by usage of certain words and phrases related to local sites and their events.
Internal affairs	1, 7, 10	A considerably high number of documents is typical for this category. They represent discussions about internal political and economic issues.
Global geopolitics	0, 6, 13	The topics contain posts with pieces of news and opinions about events happening throughout the world. The main terms include named entities (countries, states, and their leaders).
Other	2, 5	The documents of this category have mainly low probabilities since their subjects of discussion are mixed and can be classified into many small topics.

Despite the efficacy of knowledge extraction and the well-designed dashboard set, this solution faces specific challenges and limitations. Primarily, the Splunk Enterprise is resource-intensive, and the software demands powerful computational capabilities. If these constraints can't be met, open-source visualization and data manipulation tools might be a viable alternative.

The selected LDA algorithm also has inherent drawbacks. It doesn't represent topic evolution over time, necessitating our custom implementation. The algorithm doesn't model sentence structure and disregards sentiment. When analysing opinion data from social media, similar keyword sets might convey opposing sentiments across different documents. Ideally, such observations would be classified as distinct topics. However, the LDA algorithm would likely consider only word usage within documents, resulting in either a single topic or multiple closely related topics. Comprehensive reviews of other topic modelling algorithms validate their successful deployment in various domains [10, 11], and BERTopic emerges as a robust transformer-based neural topic modelling algorithm. The model supports diverse topic modelling activities, including dynamic ones. Another advantage is the integration of GPT models to summarize returned topics [12]. Consequently, topic summarization emerges as a promising research area to glean insights from vast document volumes [13, 14, 15].

4 CONCLUSIONS

This endeavour facilitated real-time monitoring of social media discussions surrounding the Russian war in Ukraine. We ingested and analysed approximately 30GB of data from February 2022 to May 2023 in Splunk using the developed dashboards. We not only modelled a plethora of topics during the study period but also discerned the distinctions among them. The MLOps-specific dashboards for experiment and run evaluations enabled swift determination of the optimal topic count for data segmentation. We postulate that web discussions mirror actual battlefield events and portray the conflict's gravity. Notably, the discovery dashboard and text analysis provide invaluable insights into whether specific text entities and overall topics contain destructive content, which is paramount in the current era of copious opinion data on social media platforms. Another solution advantage is the discovery of top hashtags. On social media, hashtags amplify information's searchability and visibility. We pinpointed frequently used hashtags within specific topics, which is also invaluable for identifying misleading or harmful content. Swift identification of popular text pieces, based on user engagement metrics and their predicted topics, facilitated comprehensive knowledge mining of web discussions about geopolitical events. The MLOps facet of the solution is vital for precise modelling, making it an invaluable tool for media analysis and content moderation.

ACKNOWLEDGMENTS

We want to extend our gratitude to all the people who stand for the freedom of Ukraine and the entire civilized world.

REFERENCES

[1] V. Solopova, O. Popescu, C. Benzmüller, and T. Landgraf, "Automated multilingual detection of pro-kremlin propaganda in newspapers and telegram posts", Datenbank-Spektrum, vol. 23, no. 1, p. 5-14, 2023, doi: 10.1007/s13222-023-00437-2.

[2] M. Popova, E. Siemens, and K. Karpov, "The concept of text processing in an ontological approach to spatio-temporal social network analysis", 2023 30th International Conference on Systems, Signals and Image Processing (IWSSIP), 2023, doi: 10.1109/iwssip58668.2023.10180274.

[3] D. Kreuzberger, N. Kuehl, and S. Hirschl, "Machine learning operations (mlops): overview, definition, and architecture", IEEE Access, vol. 11, p. 31866-31879, 2022, doi: 10.1109/access.2023.3262138.

[4] Studer, S. Bui, T.B. Drescher, C. Hanuschkin, A. Winkler, L., Peters, and et. al, "A Machine Learning Process Model with Quality Assurance Methodology." Mach. Learn. Knowl. Extr. 2021, 3, 392-413, doi: 10.3390/make3020020.

[5] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15). Association for Computing Machinery, New York, NY, USA, 399-408, [Online]. Available: https://doi.org/10.1145/2684822.2685324.

[6] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation", Journal of Machine Learning Research 3 (2003) 993-1022, doi: 10.5555/944919.944937.

[7] R. Churchill and L. Singh, (2022). The evolution of topic modeling. ACM Computing Surveys, 54(10s),(2022) 1-35, doi: 10.1145/3507900.

[8] S. Fedushko, T. Ustyianovych, and M. Gregus, "Real-time high-load infrastructure transaction status output prediction using operational intelligence and big data technologies", Electronics, vol. 9, no. 4, p. 668, 2020, doi: 10.3390/electronics9040668.

[9] S. Fedushko, T. Ustyianovych, Y. Syerov, and T. Peracek, "User-engagement score and slis/slos/slas measurements correlation of e-business projects

through big data analysis", Applied Sciences, vol. 10, no. 24, p. 9112, 2020, doi: 10.3390/app10249112.

[10] I. Vayansky and S. Kumar, "A review of topic modeling methods", Information Systems, vol. 94, p. 101582, 2020. doi: 10.1016/j.is.2020.101582.

[11] A. Abdelrazek, Y. Eid, E. Gawish, H. Mohamed, and A. Hassan, "Topic modeling algorithms and applications: a survey", Information Systems, vol. 112, p. 102131, 2023, doi: 10.1016/j.is.2022.102131.

[12] "BERTopic". [Online]. Available: https://maartengr.github.io/BERTopic/index.html [Accessed: Jul. 27, 2023].

[13] R. Belwal, S. Rai, and A. Gupta, "Extractive text summarization using clustering-based topic modeling", Soft Computing, vol. 27, no. 7, p. 3965-3982, 2022, doi: 10.1007/s00500-022-07534-6.

[14] A. Alambo et al., "Topic-Centric Unsupervised Multi-Document Summarization of Scientific and News Articles," 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 2020, pp. 591-596, doi: 10.1109/BigData50022.2020.9378403.

[15] A. Joshi, E. Fidalgo, E. Alegre, and L. Fernández-Robles, "Deepsumm: exploiting topic models and sequence to sequence networks for extractive text summarization", Expert Systems With Applications, vol. 211, p. 118442, 2023, doi: 10.1016/j.eswa.2022.118442.