Text Mining On Ukraine Russia War

Advanced-Data and Network Mining

Shalaka Waikar
Data Analytics
Dublin Business School
Dublin, Ireland
vishya.waikar@gmail.com

Introduction

The Ukraine-Russia war started in February 2022 and continues to date. Researchers have been using machine learning methods to analyze sentiments, extract trending topics, and understand global impacts on other economies. This study focuses on applying text mining techniques to Ukraine-Russia War Twitter data, with the primary aim of understanding the sentiments of people and extracting the most trending topics on Twitter.

Methodology

The study follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, which is a widely used framework for data mining projects. The CRISP-DM methodology consists of six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

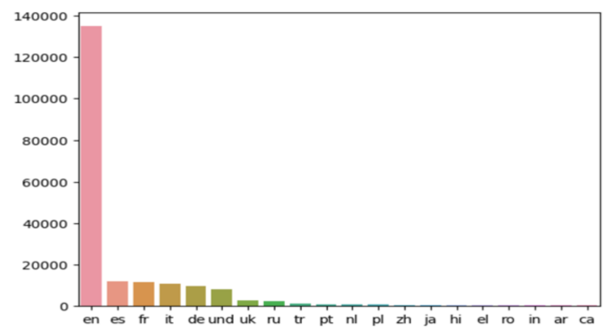
a. Business Understanding: The first stage of the CRISP-DM methodology involves defining the problem to be solved and understanding the project's goals. In this study, the goal is to apply Text Mining techniques to the Ukraine-Russia War Twitter Data to understand the sentiments of people a year after the war began and extract the most trending topics over Twitter regarding the war.

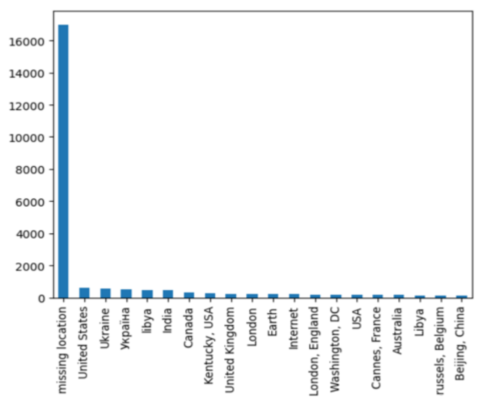
b. Data Source: The data source for this study is the Kaggle dataset on the Ukraine-Russia crisis, which is updated daily with the most recent tweets on the conflict. The dataset includes 1.2 million rows of tweets with information on the tweet's language, location, user, and text.

c. Data Understanding: The data understanding stage involves exploring and visualizing the data to gain insights into the data's characteristics. In this study, exploratory data analysis is performed using Python libraries such as Matplotlib and Seaborn. The analysis includes:

Tweets by Language: The number of tweets in different languages is analyzed, and the majority of tweets are in English. Therefore, the study focuses only on tweets in English.

Tweets by Location: As the majority of tweets are missing location information, this is not useful for the analysis.





Top Retweeted Tweets: The top ten retweeted tweets are displayed, and most of them are about the involvement of Putin, Ukraine, and China, or about Biden getting elected, and a few are of people criticizing the war.

Top 10 retweeted tweets

	text	retweetcount
39981	My daughter and I surviving the night in Ukraine. We are real people at war with crazy dictator and we need the world's support right now\n\n#StandWithUkraine https://t.co/FvdmY4GACj	99327
18239	24 kidnapped Ukrainian children from #Russia-occupied Donetsk region are passed to adoptive families in Novosibirsk, 3600 km away from home\n\nSecrecy of adoption will never let them learn were their parents killed, sent to concentration camps or deported\n\n#russiaisateroriststate https://t.co/nxT90VVU8y	9860
4294	Situation in #Ukraine: #ICC judges issue arrest warrants against Vladimir Vladimirovich Putin and Maria Alekseyevna Lvova-Belova\nRead more ⬇️\n https://t.co/5OMC7Xuuy5	8326
2012	I want to thank again every American who voted for Biden. When you were celebrating Trump is no longer a threat to humanity, I was drinking champagne for the historical moment too. But little did I know that moment would save soo many lives in #Ukraine. You & your @POTUS are 😎 https://t.co/DoappUX5jw	7777
2443	Sergei is a hero.\n\nDuring a battle in #Bakhmut, he dragged 9 wounded men to safety, applied 3 tourniquets, saved a dying soldier AND was firing his rifle in combat.\n\nSergei was trained by Sons of Liberty International (@OfficialSOLI) trainer Jason.\n\n#UkraineWillWin #Ukraine #NAFO https://t.co/R94CRWGRaD	7696
149323	Shocking! This is an actual freight train moving over a twisted, "bendy" railway track linking Ohio and Indiana in the USA!! \n\nIt's like a 4th world country.\n\nAmerica should fix its sh*thole infrastructure rather than waging endless wars.\n\n#Russia #Ukraine #China https://t.co/EPs9rca9pL	7075
179526	Pay attention to the IR leader's beard, it is full of execution ropes. This is the real face of the regime that has killed many people in #Iran, Syria, Iraq, Yemen, #Ukraine, etc. Help the people of Iran to overthrow it.\n#MahsaAmini \n@ZelenskyyUa \n@EUCouncil \n@vp\n@cnn\n@Ap\n@UN https://t.co/A35LiNQWfV	6148
7128	On this day in 2014, a #Russiajn Buk shot down a passenger aircraft #MH17 flying at 10 km from Amsterdam to Kuala Lumpur over #Donetsk region in eastern #Ukraine. \n\nAll 283 passengers and 15 crew were killed.\n\n https://t.co/O20x27pHJy	5825
20976	ua Glory to Ukraine and it's Soldiers ua\n#UkraineWar https://t.co/S4lWhnuxqu	5638
30429	THREAD: Proof that you are a Nazi if you don't #StandWithUkraine ua	5638

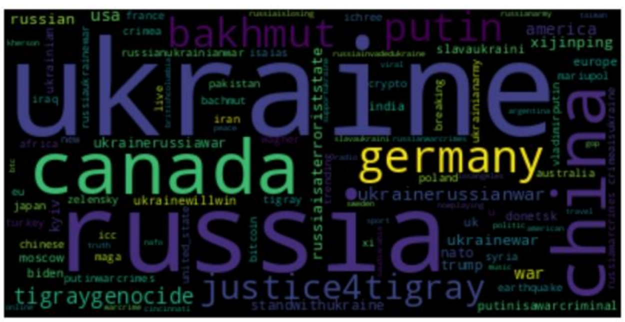
Top Tweets from Most Followed Account: The top most tweets from the most followed account are displayed, and most of them are about the involvement of Putin, Ukraine, and China, but some are not relevant to the war but are related to countries and their friendships or tourism.

Top most tweets from the most followed user accounts

	text	followers
143521	'Dear friends' Xi and Putin meet in #Moscow as #Ukraine war rages \n\nMoscow said the charge was one of several "clearly hostile displays" and opened a criminal case against the ICC prosecutor and judges. Beijing said the warrant reflected double standards\n\nhttps://t.co/UoRaxPG6U7	14685776
116774	Chinese leader #XiJinping told President #VladimirPutin that he was convinced that Putin enjoyed the #Russian people's support ahead of a presidential election scheduled for next year: Reuters	14685770
112821	#Russian President #VladimirPutin tells #Chinese President #XiJinping that they will discuss China's #Ukraine peace plan: Reuters\n\nChinese President Xi Jinping is in #Moscow for a three-day state visit to Russia. https://t.co/JZFIJZcvcj	14685748
73154	#China says #ICC should avoid 'double standards' after #Putin warrant \n\nThe International Criminal Court on Friday announced an arrest warrant for Putin on the accusation of unlawfully deporting #Ukrainian children.\nhttps://t.co/U0Nqzxr3b2	14685719
77039	#Chinese leader #XiJinping arrives in Moscow to meet with #Russian President #VladimirPutin \n\n#China and #Russia have described Xi's trip as part of efforts to further deepen their "no-limits friendship."\nhttps://t.co/jLQ5irSPfK	14685689
6948	Meet the giant #pandas Ru Yi and Ding Ding! The pair has been star attraction of the #Moscow Zoo since arrival in 2019. https://t.co/rICPkYCMUv	13097476
20670	As flowers bloom and buds sprout out from trees, an arboretum in #Sochi, #Russia has attracted tourists who are enjoying the beautiful springtime. https://t.co/n6yMZ5aXzq	13097462
26681	Giant pandas help advance friendship between #China, #Russia https://t.co/yDCDAsMF46	13097462
72184	Highlights and major milestones in bilateral ties between #China and Russia #CGTNGraphics https://t.co/26CfTErIFf	13097216
77635	Russia welcomes Chinese President #XiJinping for state visit https://t.co/g3F5QYL4n6	13097215

Word Cloud: A word cloud is formed to find the most used words, and words like Ukraine, Russia, China, Justice4Tigray, Tigray genocide, Ukraine-Russia war, etc. are highlighted.

Top most trending hashtags



d. Data Preprocessing: The data preprocessing stage involves cleaning and transforming the data into a format suitable for analysis. In this study, data cleanup is done in Python using BeautifulSoup and NLTK libraries. The tasks performed include:

Removal of HTML tags and links

- Transforming data into lowercase
- Converting contractions to words (e.g., can't've = cannot have; doin = doing)
- Removal of punctuations
- Filtering stop words
- Lemmatization to convert words to their root word (e.g., eat, ate, eating, eaten = eat)

	text	tweetcreated_date	tweetcreated_time	data_clean_list	tokenised_clean_list	tokenised_clean_text
185158	Biden admin's Kirby ridicules Putin visit to Mariupol: 'He has to know how badly he's doing inside Ukraine' https://t.co/ozma1AZzQ2 https://t.co/kuBXWsi79M #Foxnews #UkraineRussiaCrisis #russia #UkraineCrisis #RussiaInvadedUkraine #ukrainerussia #ukraine	2023-03-20	21:33:17	biden admin s kirby ridicules putin visit to mariupol he has to know how badly he s doing inside ukraine	[biden, admin, kirby, ridicule, putin, visit, mariupol, know, badly, inside, ukraine]	biden admin kirby ridicule putin visit mariupol know badly inside ukraine

Tokenization

e. Modeling (Sentiment Analysis): The modeling stage involves selecting and applying a model to the data to predict outcomes. In this study, sentiment analysis is performed using RapidMiner, and the VADER sentiment analysis algorithm is used for scoring and defining polarity. The steps performed include:

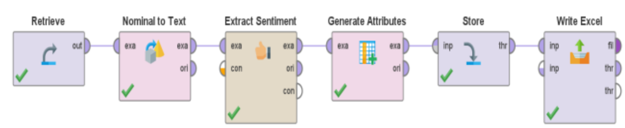
Loading cleaned data set

- Generating sentiments using VADER
- Aggregating polarity
- Balancing the dataset
- Predicting sentiments using the auto model in RapidMiner
- Evaluating the result using Fast Large Margin

Row No.	text	tweetcreate...	tweetcreate...	tokenised_clean_text
1	Success is gi...	2023-03-20	00:00:00	success give seekers video
2	#NSTleader ...	2023-03-20	00:00:00	nstleader international criminal court iraq war
3	Russia's co...	2023-03-20	00:00:00	russia combat losses ukrainewar front incl...
4	UAUkraine W...	2023-03-20	00:00:00	ukraine war footage follow new videos ukr...
5	British Intelli...	2023-03-20	00:00:00	british intelligence ukrainerussiawar map ...
6	Book summe...	2023-03-20	00:00:01	book summer camp canada national park go
7	#Russia's se...	2023-03-20	00:00:01	russia self delude victimization
8	#Ukraine: Ev...	2023-03-20	00:00:01	ukraine everyone hear recent mq incident r...
9	In preparatio...	2023-03-20	00:00:02	preparation vlads time prison want show t...

f. Modeling (Topic Modeling): The topic modeling stage involves extracting topics from the dataset. In this study, LDA (Latent Dirichlet Allocation) is used for topic modeling, and it is implemented in two ways: LDA modeling in Python using Gensim and LDA modeling in RapidMiner. The steps performed include:

* Running the model in Python
* Loading cleaned data set
* Preparing the data using Gensim
* Building the LDA model
* Extracting the topics
* Visualizing the topics using pyLDAvis



Row No.	Score	text	tweetcreate...	tweetcreate...	tokenised_...	sentiment
1	0.692	Success is gi...	2023-03-20	00:00:00	success give...	positive
2	-1.359	#NSTleader ...	2023-03-20	00:00:00	nstleader int...	negative
3	-0.795	Russia's co...	2023-03-20	00:00:00	russia comb...	negative
4	-0.744	UAUkraine W...	2023-03-20	00:00:00	ukraine war f...	negative
5	0.538	British Intelli...	2023-03-20	00:00:00	british intelli...	positive
6	0	Book summe...	2023-03-20	00:00:01	book summe...	neutral

b. Running the model in RapidMiner

Loading cleaned data set

Applying LDA operator

Extracting topics

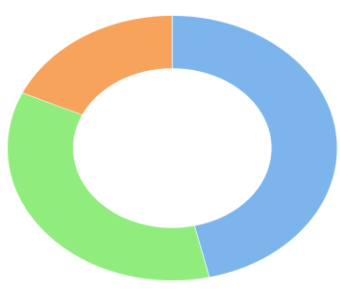
g. Evaluation: The evaluation stage involves assessing the quality of the models and determining whether they meet the project's goals. In this study, the quality of the sentiment analysis and topic modeling models is evaluated using performance metrics such as accuracy, precision, recall, and F1 score.

h. Deployment: The final stage of the CRISP-DM methodology involves deploying the models and implementing the insights gained from the data analysis. In this study, the insights gained from the sentiment analysis and topic modeling models can be used by policymakers and researchers to understand the sentiments of people towards the Ukraine-Russia War and extract the most trending topics over Twitter regarding the war. The models can also be integrated into a web application for real-time analysis and visualization of Twitter data related to the war.

Modeling- Sentiment Analysis

In the Modelling (Sentiment Analysis) stage of the study, the goal is to predict the sentiment of each tweet in the dataset. The sentiment analysis technique used in this study is VADER (Valence Aware Dictionary and Sentiment Reasoner). VADER is a rule-based sentiment analysis tool that is specifically designed for analyzing social media texts. It uses a lexicon of words with associated sentiment scores and rules to identify the sentiment of a given text.

After the sentiment analysis is performed using RapidMiner, the data is balanced by multiplying and dividing it into positive, negative, and neutral sentiments. This ensures that the model has an equal number of tweets for each sentiment category, which helps to avoid bias in the results.

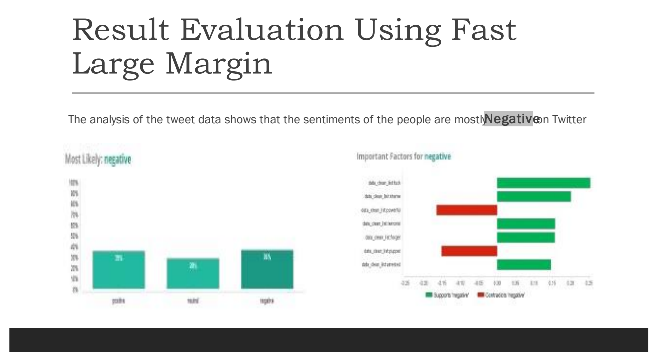


To evaluate the performance of the sentiment analysis model, two algorithms are used - Fast Large Margin and Generalized Linear Model. The Fast Large Margin algorithm shows the highest accuracy and performs the best, followed by the Generalized Linear Model.

Row No.	sentiment	count(sentiment)
1	negative	19466
2	neutral	7616
3	positive	14859

Based on the results, the sentiment of people on Twitter regarding the Ukraine-Russia war is mostly negative. This insight can help in understanding the overall public opinion about the war and can be useful for policymakers and researchers working in this area.

Models	Accuracy	Standard Deviation (+-)	Classification Error	Standard Deviation (+-)
Naïve Bayes	37.4%	0.5%	62.6%	0.5%
Generalized Linear Model	56.0%	0.8%	44.0%	0.8%
Fast large Margin	61.6%	0.7%	38.4%	0.7%
Decision Tree	41.7%	0.7%	58.5%	0.7%
Random Forest	39.1%	0.7%	60.9%	0.7%
Gradient Booster	35.2%	0.5%	64.8%	0.5%



Modeling- Topic Model

In the Modeling (Topic Model) stage of the study, the goal is to extract topics from the Ukraine-Russia War Twitter data. The approach used for this is Latent Dirichlet Allocation (LDA), which is a widely used method for topic modeling in natural language processing.

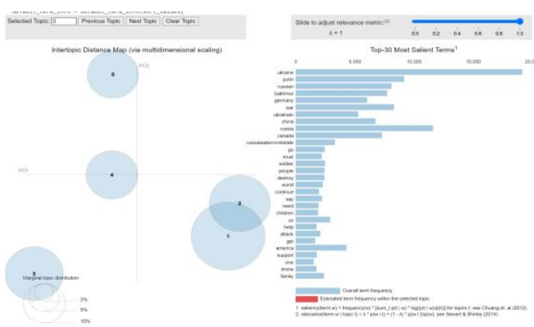
	number_of_topics	coherence_score_C_v	coherence_score_U_mass
0	1	0.281096	-3.121876
1	2	0.565669	-2.679459
2	3	0.371025	-5.075367
3	4	0.520381	-3.592864
4	5	0.479557	-4.640056
5	6	0.451019	-6.827097
6	7	0.414954	-5.044110
7	8	0.422059	-6.048332
8	9	0.467584	-6.339196

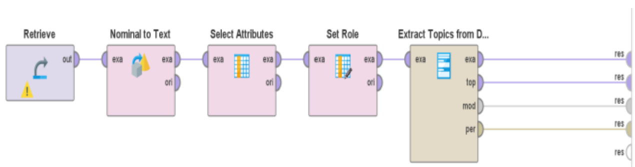
LDA assumes that each document is a mixture of topics and that each word in a document is associated with a particular topic. The goal of LDA is to identify the underlying topics that explain the co-occurrence of words in the corpus.

Topic: 0 Word: 0.097*"putin" + 0.026*"go" + 0.022*"canada" + 0.019*"help" + 0.018*"get" + 0.016*"one" + 0.015*"russiawarcrimes" + 0.013*"time" + 0.013*"stop" + 0.011*"mariupol"
Topic: 1 Word: 0.066*"russian" + 0.027*"russiaisaterroriststate" + 0.020*"soldier" + 0.019*"world" + 0.018*"say" + 0.015*"support" + 0.014*"drone" + 0.013*"crimea" + 0.011*"close" + 0.011*"use"
Topic: 2 Word: 0.039*"russia" + 0.035*"china" + 0.018*"ukraine" + 0.015*"us" + 0.013*"america" + 0.012*"family" + 0.011*"usa" + 0.010*"ukrainerussianwar" + 0.010*"putinisawarcriminal" + 0.007*"nato"
Topic: 3 Word: 0.119*"ukraine" + 0.058*"bakhmut" + 0.046*"germany" + 0.044*"war" + 0.040*"ukrainian" + 0.039*"canada" + 0.030*"russia" + 0.018*"destroy" + 0.016*"attack" + 0.013*"america"
Topic: 4 Word: 0.026*"war" + 0.023*"must" + 0.023*"people" + 0.020*"continue" + 0.020*"need" + 0.020*"children" + 0.014*"icc" + 0.014*"make" + 0.014*"right" + 0.013*"tell"

To implement LDA, two approaches are used in this study. The first approach is to run the LDA model in Python using the

Gensim library, which is a popular open-source library for topic modeling in Python. The Gensim library provides a simple interface for implementing LDA models and allows for the selection of a number of topics and other model parameters.





The second approach used in this study is to implement LDA modeling in Rapidminer. Rapidminer is a data mining and machine learning platform that provides a wide range of tools for data preparation, modeling, and evaluation. Rapidminer's LDA operator allows for the selection of the number of topics, alpha, and beta values and provides options for topic modeling based on different approaches such as Gibbs Sampling, Variational Inference, and Expectation Maximization.

Topic #	Topics (RAPIDMINER)
1	Tigraygenocide_war_committed_women_atrocities
2	russian_forces_destroyed_military_near_ukranianarmy
3	russia_china_president_meeting_leader_peace
4	need_peace_make america_see_iraq_never_stop
5	footage_attack_soldiers_horrible_near_kherson

Once the LDA models are implemented, the topics are extracted, and the results are analyzed to identify the most significant topics in the data. These topics can provide insights into the most discussed themes on Twitter related to the Ukraine-Russia war.

Topic #	Topics (Python with Gensim)
1	putinisawarcriminal_issue_home_arrest_warrant
2	russia_destroy_tank_kill_soldier_kherson
3	president_see_army_army_take_missiles_near country_video
4	people_women_children_still_crimes_genocide
5	putin_go_canada_help_russianwarcrimes

Conclusion

The study shows that the sentiments of people on Twitter regarding the Ukraine-Russia War are mostly negative. The most trending topics over Twitter regarding the war are related to Ukraine, Russia, and China, with hashtags like #Justice4Tigray, #Tigraygenocide, and #Ukraine-Russia war getting highlighted. The study provides valuable insights into public opinion on the war and can be used by policymakers and analysts to understand the sentiments of people better.

Future Scope

The study can be extended by analyzing the impact of the war on other countries and their economies. The study can also be extended to include other social media platforms like Facebook and Instagram, which could provide more insights into public opinion on the war.

Key Takeaway

The Key learnings and takeaways from the assignment on Text Mining in Ukraine-Russia War:

1. Text mining techniques can be used to analyze people's sentiments and extract trending topics related to a particular event or topic.
2. The CRISP-DM methodology is useful for organizing the text mining process, including understanding the business problem, selecting appropriate data sources, and preparing and modeling the data.
3. Exploratory data analysis is a crucial step in understanding the data better before preprocessing and modeling.
4. Preprocessing of social media data is necessary because it is noisy and requires data cleanup. This includes the removal of HTML tags, and links, converting contractions to words, removal of punctuations, filtering stop words, lemmatization, and tokenization.
5. Sentiment analysis is a popular text mining technique used to determine the attitude or emotion of a writer towards a specific topic or product. Vader Sentiment Analysis is widely used to determine the polarity of tweets.
6. Topic modeling is another text mining technique used to extract topics from a given set of documents (in this case, tweets). LDA (Latent Dirichlet Allocation) is one of the most popular methods used for this purpose.
7. The tweet data analysis shows that the people's sentiments are mostly negative on Twitter towards the Ukraine-Russia war.
8. Understanding the sentiments of people towards a particular topic can help in making informed decisions and formulating appropriate policies. It can also be useful for businesses to gain insights into consumer opinions and improve their products or services accordingly.

Bibliography

Bwandowando (2022), 'Ukraine Conflict Twitter Dataset'. Available at: *https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows*? (Accessed: 20 March 2023)

Pedro, J. (2022), 'Understanding Topic Coherence Measures', Towards Data Science. Available at: https://towardsdatascience.com/understanding-topic-coherence-measures-4aa41339634c (Accessed: 23 March 2023)

Tripathi (2019), 'What Is Balanced And Imbalanced Dataset?', Medium. Available at: https://medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5 (Accessed: 23 March 2023)

'Auto Model' (no date), Rapidminer. Available at: https://docs.rapidminer.com/latest/studio/guided/auto-model/ (Accessed: 6 April 2023)

Selvaraj, N. (2020), 'A Beginner's Guide to Sentiment Analysis with Python', Towards Data Science. Available at: https://towardsdatascience.com/a-beginners-guide-to-sentiment-analysis-in-python-95e354ea84f6 (Accessed: 6 April 2023)

'Sentiment Analysis: Concepts, Models, and Examples' (2021), turbolab. Available at: https://turbolab.in/sentiment-analysis-concepts-models-and-examples/ (Accessed: 7 April 2023)

'NLP Pipeline' (no date). Available at: https://alvinntnu.github.io/NTNU_ENC2045_LECTURES/nlp/nlp-pipeline.html (Accessed: 7 April 2023)

Kapadia, S. (2019), 'Evaluate Topic Models: Latent Dirichlet Allocation (LDA)', Towards Data Science. Available at: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0 (Accessed: 11 April 2023)

Kim, S. (2022), 'Let us Extract some Topics from Text Data — Part I: Latent Dirichlet Allocation (LDA)', Towards Data Science. Available at: https://towardsdatascience.com/let-us-extract-some-topics-from-text-data-part-i-latent-dirichlet-allocation-lda-e335ee3e5fa4 (Accessed: 16 April 2023)