Data Science Project

Mary Glantz

Linesh Dave

Using Sentiment Analysis to Understand the Relationship between Russian Presidents' Foreign Policy Rhetoric and Actions

December 3, 2023

Executive Summary

Analyzing the transcripts from the Kremlin website holds the potential to provide some useful insight into Kremlin foreign policy. At the very least, the analysis should show if there is some relationship between Kremlin words and actions. Using Natural Language Processing, it is possible to apply state-of-the-art transformer models to assess the sentiment of texts and to see if there exists any correlation between that sentiment and Russian foreign policy activities. First, the text needs to be prepared by cleaning the dataset and tokenizing the transcripts. The resultant dataset is then broken down into sets based upon topic to facilitate comparative modeling. Exploratory data analysis techniques, including frequency plots, term frequency analysis, term frequency-inverse document frequency analysis, topic modeling, named entity recognition, and VADER and TextBlob sentiment analysis show if there are any interesting patterns that should be explored more deeply with sentiment/emotion labeling. Finally, the datasets are analyzed using zero-shot, pre-trained sentiment analysis BERT models and using the VADER/TextBlob labeled datasets for sentiment analysis. The result shows a statistically significant correlation between the distribution of emotion labels of the transcripts and the years in which the transcripts are created.

Table of Contents

Project Scope

Problem Description

Russian - American relations have reached a new post-Cold War low. In 2022, Russia followed up their 2014 invasion of Ukraine's Crimea and Donbas regions with a full-scale invasion of the entire country, causing death and destruction for hundreds of thousands of people. The United States and NATO responded with military and economic support to Ukraine and strangling sanctions on Russia. People around the world worry about the potential for escalation as Russian President Vladimir Putin has resorted to nuclear saber-rattling. At the same time, as the battlefield situation stagnates, it is unclear how this conflict will end.

Resolving this crisis will depend to a large part on understanding how we got here in the first place. There are a large number of theories explaining the conflict's origins, with political scientists covering the spectrum from laying all responsibility on the Kremlin to blaming Washington for ignoring Russian security concerns. All these explanations seek, to a certain extent, to pull evidence from the speeches of political leaders. For example, most western analysts mark Vladimir Putin's 2007 speech at the Munich Security Forum (Red, 2022) as a turning point in Russia's relations with the west---the point at which Putin declared open hostility to the United States. Yet, picking and choosing evidence from leaders' speeches runs the risk of selection and/or confirmation bias. Humans tend to find evidence that supports their theories; they likely overlook evidence that contradicts their prior beliefs (Jervis, 2017). Fortunately, natural language processing (NLP) and machine learning offer a potential solution to this problem by enabling us to use machines to analyze an entire corpus of speeches to see if there are any discernible patterns. To that end, the data analytics problem that I am analyzing is whether there are any discernible patterns in the speeches and writings of the Russian presidents (Putin and Dmitry Medvedev) that would indicate aggression, defensiveness, or something else.

Project Importance

To resolve or prevent conflict, you need to identify and address the root causes of conflict. The problem lies in the difficulty of doing so. With the Russian large-scale invasion of Ukraine in February 2022, years of brewing hostility between Russia and NATO/the United States came to a violent head. But scholars and policymakers are unable to agree upon what the root causes of Russian-Western (a loose term meaning NATO, the EU, the United States, and Europe writ-large) conflict are. The main lines of dispute are between those who believe that Western unwillingness

to consider Russian security concerns led to this war (Mearsheimer, 2014) and those who believe the war is a result of Russian imperial ambitions (Shigeki, 2022).

Both sides are persuasive, but it is currently impossible to determine which is correct (or if both are). Their use of evidence may be part of the problem. In arguing their cases, political analysts select examples from Putin's (and to a lesser extent, Medvedev's) speeches to indicate whether Russia is aggressive or defensive. Thus, Putin's 2007 Munich speech is cited as evidence of his intention to actively challenge the United States and NATO. At the same time, the context of the speech was a NATO that had expanded into the former Soviet states of Estonia, Latvia, and Lithuania; that was debating expanding into Georgia and Ukraine; and that was actively planning to put missile defense systems in former Warsaw Pact countries. Thus, some analysts saw Russia becoming increasingly threatened and aggressive as a result.

Picking and choosing evidence is a human problem. Not only are humans inclined to confirmation bias (Jervis, 2017), but reading through everything the Russian presidents have said or written in the past 23 years is a daunting challenge. Fortunately, it is much easier for a computer to do, and, depending upon the methodology used, the computer is much less likely to suffer from confirmation bias. Thus, conducting an NLP sentiment analysis of all the Russian presidents' speeches and writings since 1999 may provide some interesting and useful insights to those interested in the root causes of the Russian-Western conflict.

Background

Some scholars have attempted to tackle the problem of analyzing Russian leaders' words using data science methodologies. Snegovaya (2020) used a bag-of-words approach to analyze whether the aggression toward the west in the Russian president's sentences correlated with key events---in particular, high oil prices, the year before NATO expansion, or Russian economic decline. She concluded that aggressive rhetoric correlated with high oil prices. There were some weaknesses in her approach, however. First, there was potential bias in the way she coded sentences as "aggressive" or "not aggressive." Second, she limited her analysis to just the three hypotheses listed. There are other events she did not consider, such as "color revolutions," that could have correlated even more strongly with aggressive rhetoric.

Sasse (2020) used a word-frequency analysis to explore the relationship between Russian rhetoric and policy. By looking at the Russian leaders' use of words like "Ukraine," "Crimea," and "Donbas," Sasse concludes that the relationship between rhetoric and policy is tenuous---it is

episodic, indicating, she argues, a discrepancy between rhetoric and policy. She argues that Putin only uses these terms during moments of crises. Unfortunately, her analysis is short and does not explore other relationships between words and policy. Nor does it address the obvious conclusion that a President may only talk about something when it is in the news, so the relationship may not be significant.

The number of data science analyses of Russian leaders' speech is surprisingly limited. There are, however, a number of other attempts to analyze sentiment of media and social media to understand the impact of international events, which provide useful methodological insights. Some Russian scholars have analyzed Western media using a bag-of-words approach to assess sentiment towards Russia (Khrustova L.E., 2020). To analyze the sentiment, they used the Loughran and McDonald base dictionary and labeled the words as "positive" – "negative", "uncertain" – "controversial", or "limiting" – "redundant." They conclude that with the introduction of sanctions against Russia in 2014, the Western media adopted an increasingly negative sentiment towards Russia. (Of course, the obvious problem with their analysis is the question of causality---was it the introduction of sanctions that led to negative sentiment, or the Russian invasion of Ukraine that did so?) Similar methodological insights can be gained by looking at the work of Nisch (2023) who examines sentiment in Ukrainian President Volodomyr Zelensky's official twitter statements.

Data Set Description

To analyze whether there are any discernible patterns in the speeches and writings of the Russian presidents (Putin and Dmitry Medvedev) that would indicate aggression, defensiveness, or something else, I will be using the dataset "The President's Words: A Term Frequency Analysis of Putin's and Medvedev's Statements in Official Kremlin Transcripts" found on this webpage: https://discuss-data.net/dataset/7e016752-2438-46c9-9526-d35d41c823a2/. The dataset consists of two parts: (1) A database of more than 10,000 transcripts published on the official website of the President of Russia since December 31, 1999 (last update 2023, January 16th); and (2) A frequency analysis of the terms used by president Putin and president Medvedev in these documents. For the purpose of my research, I will use only the file of English language transcripts of the full speeches/writings.

That file is a zipped json-format document. There are 9,349 rows and 10 columns. The columns are date, persons, transcript_unfiltered , kremlin_id, title, teaser, tags, transcript_filtered, and wordlist. Thus, there are a number of ways that this data can be explored and analyzed. All of the columns contain object data types except kremlin_id, which is an int64. There are no null

values. The most significant potential problem with this dataset is that it is an official Kremlin dataset, curated by the Kremlin to reflect their views and perspectives. A potentially more useful dataset would include news reports and social media.

Data Analytics Tools

The dataset includes both full texts of the transcripts and a field that is broken down into words from each transcript. A simple word-frequency analysis would require only the use of the wordlist. Sasse's (2023) analysis, however, shows the limitations of such an approach—word counts do not necessarily reveal sentiment or significant correlations between words and events. Importantly, words can be taken out of context. For that reason, Snegovaya (2020) conducted her bag-of-words analysis using full sentences. Ideally, I would like to contrast both approaches, first using a bag-of-words approach; subsequently I would like to try a more modern BERT analysis. Nisch (2023) used a BERT model for his analysis, though he used it in R and I will be using Python.

In order to use a BERT model, I will first have to pre-process my text using a suite of NLP-methodologies. This includes removing symbols and stop words, tokenizing the words, and other steps. Nisch (2023) argued that he did not use stemming because his research raised questions about the impact of stemming on model accuracy. Time-permitting, I would like to assess the model with both stemmed and unstemmed data. I will also need to conduct research to determine which pre-trained BERT model is optimal for this project. Nisch (2023) used "FastBERT" because it was used in comparable research to his.

Project Milestones

The steps in this project include first, researching and understanding the problem to be studied and the most useful approach to studying it. Background research indicates that understanding the sentiment of political leaders is useful to preventing and resolving international conflict. It also indicates that machine-learning techniques may be useful but have not been widely applied to the particular problem of Russian relations with the West. In this step it is also important to identify a dataset that will be relevant for analyzing the problem.

The next step will be to clean and prepare the data. Since I will be using sentiment analysis and natural language processing, this requires cleaning the data and applying NLP-methodologies so that the model can understand the data. In other words, I will essentially be turning the words into numbers in a matrix so the model can understand them and their relationships to each other.

Once the data is prepared, I will conduct exploratory data analysis. This will include deriving descriptive statistics from the data, and creating visualizations that may reveal patterns and other questions that I should explore and address.

The next step is modeling the data. This step will require that I identify the most appropriate model for answering my questions with the data. In addition, I will have to run the model and fine-tune it to get the best results. This step is really an iterative process, as running the model will lead to additional questions that will require running the model with adjustments. If time permits, I may run more than one type of model in order to compare their results.

Finally, with the results of the models and the exploratory data analysis, I will reach some conclusions and prepare them for presentation. This will involve not only assessing the performance and results of my model, but also suggesting future work on the topic.

Completion History

Assignments 1 - 5

Weeks 1 & 2	These past two weeks, I completed the course readings on the different types of data science project lifecycles and storytelling with data so that I could better understand how to approach and carry out a project. I identified a general topic/question I was interested in. Then I conducted background research on that question in order to see what methodologies had been tried, what worked and what didn't, and what data sets would be most useful. With that in mind, I identified a dataset that appears to contain the information I need to answer the question/problem. Happily, that dataset included links to other work that had been done on it. I saw that no one had tackled my specific question, but the other scholarly articles I found showed how others had used data science to address similar questions. Finally, I answered both the question for discussion (which fed into the written assignment 1) and completed assignment 1 on project scoping.
Weeks 3 & 4	For week 3, I completed the readings on data preprocessing, but found most did not address NLP tasks. So I found articles online on NLP preprocessing and used some of the O'Reilly books (Hands On Python Natural Language Processing; Applied Natural Language Processing with Python). I then preprocessed my data by cleaning it for NLP (lower casing, expanding contractions, removing punctuation and stopwords, etc.). I did some initial EDA to find out what other preprocessing I needed to do, and I researched what sorts of preprocessing was necessary for BERT transformers (which I will use because my data is unlabeled). Thanks to the advice of one of my classmates, I knew RoBERTa would be a good option. My research indicated that for unsupervised NLP of political language, that is the best option. For week 4, I conducted a comprehensive exploratory data analysis. This included basic statistics like word count and sentence/transcript length, and also more NLP-oriented tasks. For example, as I plotted the number of transcripts per year, I realized that it would be useful to filter the dataset by keyword. (In this case, I used keywords that are most important for understanding Russian foreign policy.) Then I plotted just those and looked at the plots by month and/or year. In addition, I conducted term frequency analysis, n-gram analysis, and topic modeling. I did those on both the general dataset and the datasets filtered by keyword. I then discovered it was easy to do some basic sentiment analysis, to label my data with positive, neutral, and negative labels. I used both VADER and TextBlob so I could compare the results. (With a column of labels now, I can do both the zero-shot modeling I planned and do a model with a labeled train and unlabeled test set. Finally, I performed Named Entity Recognition (just to see what the most common entities there were).
Weeks 5 & 6	In these two weeks, I explored using multiple variations of a RoBERTa model on my data. I also discovered what additional pre-processing needed to be done to the data to make it work with the models. First, I explored what to do about the unlabeled data---that included reading the article on using label-description training for zero-shot text classification, but I was really stumped on how to determine what text reflected what emotion. I suspect that

	sort of labeling would really have to be done by a linguist or expert in emotion and language to be accurate. I thus decided to let the model that had been trained on much more text than I have label the data itself, but to check that labeling by running the data through the model again in a shuffled format. Dealing with the three-class model was easier because I had labeled my data using TextBlob (and VADER, but VADER labeled everything "neutral"). In the process of exploring the text, however, it became clear that my data was unbalanced and so the model performed very poorly on precision and recall. I explored different ways to balance the data, including random oversampling, but that resulted in too large a dataset. So I randomly undersampled the data. I also had to segment the data because the model could only process samples of size 512. This required cutting each transcript into smaller chunks (while maintaining the integrity of the rows so the labels remained with the correct chunk). Finally, I ran the data through the models and tuned the hyperparameters until I got the optimal settings.
Weeks 7 & 8	In these two weeks I took the labeled dataset that my model produced and conducted another exploratory data analysis on the newly labeled dataset. This involved creating visualizations for the entire dataset by label, and then filtering the dataset by topic and creating more visualizations. These visualizations showed clear patterns. To confirm that these patterns really existed, I performed statistical tests on the data. I used chi-2 tests to establish the presence of a statistically significant relationship and then I used a Cramer's V to measure the strength of that relationship. I repeated this on each of the filtered datasets.

Lessons Learned

Assignments 1 - 5

Assignment 1	I learned how useful a lifecycle approach to tackling a data science challenge is. By investing time in background research and exploring different methodologies and data, I can take a more structured approach to using data to answer a question. This should enable me to hone in more quickly on answers rather than to just endlessly explore data and come up with more questions. In addition, I completed the readings on data storytelling. Those helped me understand what approaches I should take with the data in order to answer the question I posed.
Assignment 2	NLP data preprocessing is significantly different from traditional (at least to me) data preprocessing. The obvious points are the need to process the text by removing punctuation, tokenizing it, etc. The other thing I learned (unfortunately late in the process) was that how you preprocess the data is dependent upon what model you choose to run on your data. In other words, I knew I wanted to use a BERT model, but which one I choose affects whether or not I tokenize into words or sentences. In future, I will study the models first to determine which one I should use, and then do the preprocessing. I also learned that when importing a json file, a pandas dataframe will consider an empty string as a value, not a null value. The most important thing I learned is that data preprocessing is really a key stage in whatever modeling you do. There are a number of different ways to do it, all of which may lead to a different end result with your model. I opted

	to do it in two ways—one which I can use for EDA and another which will just make use of the Hugging Face tools that feed directly into the RoBERTa pre-trained model.
Assignment 3	As I tried to calculate statistics for my dataset, I learned that expanding the contractions converted my text into one large string, making the tokenization flawed. I also learned that really for sentiment analysis, you want to keep contractions. So I reprocessed the dataset dropping that code chunk. I also learned that exploring the data, especially with visualizations, was a really effective way of discovering and highlighting patterns and questions that can be explored with a model.
Assignment 4	In preparing this week's assignment, I was reminded of the importance of the data and data preparation for effective modeling. The dataset was unbalanced, with far more neutral examples than any others in both the three-class and seven-class cases. I had to balance the dataset, but in doing so, had to convert the dataset from a hugging face to pandas dataframe and back. I also was reminded about the importance of doing this balancing before splitting the dataset into train/test/and validate in order to avoid contaminating the data before the model saw it. Choosing the correct parameters was vital for ensuring not only that the model made the most accurate (and correct across all classes) predictions, but also for ensuring that I had enough memory in colab to complete the process. I gradually learned that I needed to perform part of the modeling in one notebook, save the model and the datasets, and create a new notebook to continue the process from a different starting point. Finally, because I was in the hugging face universe, I couldn't simply use a cross-validation from sci-kit learn, but had to import optuna and write a small script that would run the data through a reinitiated model ten times.
Assignment 5	In this assignment I learned how to take the data produced by my model and develop hypotheses and conclusions. I explored using both data visualizations and statistical tests to do this. I also did a bit of research on data security and integrity in order to understand the implications for my particular dataset and project.

Data Profiling and Preparation

Data Summary

As discussed above, the data I am using is taken from a project called "The President's Words: A Term Frequency Analysis of Putin's and Medvedev's Statements in Official Kremlin Transcripts" (Marcus, 2023). The creators of this data are dekoder.org, an organization at the Research Center for East European Studies at the University of Bremen. The license for using the data is an Open Data Commons Attribution License (ODC-By) v1.0. The dataset consists of almost 10,000 transcripts published on the official website of the President of Russia (http://en.kremlin.ru/events/president/transcripts) since December 31, 1999 (when Putin became president). It was last updated on January 16, 2023. This data, thus, reflects the official messages and views of Russian leadership under almost the entirety of the Putin era. It is, therefore, a potentially invaluable resource for understanding Putin's thinking about a variety of issues, including foreign policy. That said, a significant weakness of this dataset is that it is almost certainly biased, reflecting only what the Kremlin wants to say to the public. Real attitudes and sentiments may be hidden.

The data was collected beginning from March 1, 2020 to January 16, 2023 through web scraping from the Kremlin website. The dataset is in JSON format and consists of 9349 entries with 10 columns. The data file includes filtered versions of the original transcripts (with only Putin's and Medvedev's words left) and pre-processed wordlists (lemmatized versions of the filtered transcripts), but for my purposes (learning how to perform these tasks myself, I am dropping these columns and re-creating them myself. The data are not labeled for sentiment (or anything else), and there is no training data set. The columns in the dataset include the transcript, the date of the transcript, the person who wrote or spoke in the transcript, a kremlin id for the transcript, the place where the speech/writing occurred, the title of the transcript, a teaser for the transcript, tags, a filtered transcript, and a word list. The last two reflect the preprocessing the data collectors did to the initial data. In addition, while the pandas dataframe.info() indicates no null values, a reading of the data does show that some entries are left blank. That may indicate a challenge in converting json files to pandas. I will use code to fill those blanks with NAs.

Data Definition/Data Profile

Field Name	Definition	Data Type	Outliers	Frequency of Nulls	Potential Quality Issues
date	date and time at which	object	none	none	the time field seems inaccurate; the

	transcript was created				date will need to be converted to date/time object for time-series analysis
persons	supposed to be the person who created the transcript	object	none	9349	The data consists only of '[]', this is useless. I replaced the [] and now have no non-null values

transcript_unfiltered	This is the unfiltered transcript from the website.	object	none;
			after preprocessing I discovered there were 836 outliers in the word counts from these unfiltered transcripts.
			`Maximum Word Count: 39828`
			`Date of Transcript: 2013-04-25 16:50:00`
			`Title of Transcript: Direct Line with Vladimir Putin`
				4	Will include typos, punctuation, and other things irrelevant analysis.

kremlin_id	id number assigned by Kremlin webmaster	int64	none	none	these numbers are essentially meaningless, just something for the

					webmaster to keep track of
place	location of speech or publication	object	none	618	there may be inconsistencies in how locations are described
title	title of transcript	object	none	none	may be typos
teaser	short description of transcript	object	none	4281	may be typos
tags	words that describe the topic of the transcript	object	none	3933	may be mislabeled
transcript_filtered	transcripts filtered by punctuation, stopwords, etc	object	none (note: these do not have outliers, but as noted above, once processed, the word counts do have outliers.)	14	different number than unfiltered transcripts, so may have left out important transcripts
wordlist	lemmatized words from filtered transcripts	object	none (note: these do not have outliers, but as noted above, once processed, the word counts do have outliers.)	none	may be missing words from unfiltered transcripts (the difference between filtered and unfiltered is 10) that are important

Data Preparation

The preparation of data for Natural Language Processing depends a great deal upon what you intend to do with that data. I plan to use a transformer to conduct sentiment analysis on my data.

My data is not labeled, however, and does not have a target variable. Thus, I have two choices---manually label hundreds of transcripts or use an unsupervised learning methodology. Given the time involved in manually labeling data, I have opted for the latter. That enables me to use a library of preprocessing tools like tokenizers that come from the hugging face collection. These usually require very little text preprocessing.

That said, sentiment analysis is not the only useful information this dataset can reveal. Other exploratory data analysis tools for natural language processing can provide a great deal of context that can enhance and clarify the results of a sentiment analysis (Sharma, 2022). To that end, I dropped the columns in the dataset I did not consider relevant ('kremlin_id', 'teaser', 'tags', 'transcript_filtered', 'wordlist', and 'persons'). I then conducted text preprocessing: expanded contractions and created two columns of tokenized data from the transcripts: words and sentences. I then converted all the words to lower case. Some of the words I will be most interested in have more significant meaning as parts of phrases ("missile defense" as opposed to "missile" and "defense" for example). To identify these, I ran a search of all the words around key words (eg. america, states, europe, revolution), identified important words around them, and recreated some of these words into phrases. I then cleaned up the tokenized transcripts by removing punctuation and stop words (Mysiak, 2021). I also lemmatized the words using NLTK's wordnet lemmatizer ("NLTK :: Natural Language Toolkit,"). In addition, I converted the date/time column to a date/time object so that I could more easily do time series analysis. Finally, I saved that data frame as a csv file for exploratory data analysis and visualization.

Data Visualizations

Descriptive Statistics

It is possible to calculate descriptive statistics for some of the columns in the dataset. For example, the mean transcript length is 8350 characters. The maximum length is 232,337 characters and the minimum is only 59 characters. The median is 4709, the first quartile is 2677, and the third quartile is 8605 characters. The standard deviation is a large 14283.37. Breaking the transcripts into words, the mean word count is 1388. The maximum is 39828 and the minimum is 1. The median is 772, the first quartile is 445, the third quartile is 1415 words. The standard deviation is 2427. Removing the stop words, the mean word count is 722; the maximum is 20410; the minimum is one; and the standard deviation is 1228. The median is 409, the first quartile is 232, and the third quartile is 748. The mean word length is five characters. The maximum is eight characters, and the minimum is three. The median is five characters, the first quartile is 4.9 characters, and the third quartile is 5.2 characters. The standard deviation is 0.27.

The distribution of word count in the transcripts is interesting. Calculating the interquartile range and determining the upper and lower bound for outliers reveals that 836 of the 9349 transcripts have a word count that are outliers. The transcript with the highest word count (and, thus, the largest outlier) is the “Direct Line with Vladimir Putin” on 2013-04-25 with a word count of 232,337. This was a call-in conversation Putin had with the Russian people in the midst of a series of large protests against his presidency ("10 Years Since Bolotnaya, the Biggest Protests of the Putin Era," 2021).

Data Visualization Definitions

Overall, the descriptive statistics did not reveal a great deal about the relationship between Russian foreign policy and Russian leaders’ communications. The timing of the dramatic outlier in transcript length and the largest protests in Putin-era Russian history suggests that exploring the relationship between transcript length/frequency and the timing of significant foreign policy actions may be worthwhile. Schoenbach (2000) suggests that this is a useful approach. This can be accomplished by grouping the transcripts by a given date. In the visualizations below, the transcripts are grouped by year. The x-axis reflects the year and the y-axis indicates the count of transcripts issued that year. One of the most salient Russian foreign policy actions under Putin was the invasion of Ukraine and the annexation of the Crimean Peninsula. Filtering the datasets

by these keywords and repeating the process of plotting by date and count is another useful approach for visualizing the data.

In addition, one of the more common ways of visualizing a natural language processing project is through the use of word clouds. Sharma (2022) explains that "Word clouds are the visual representations of the frequency of different words present in a document. It gives importance to the more frequent words which are bigger in size compared to other less frequent words." Singh (2023) argues that while visually appealing, word clouds do not reveal that much information about the words in a text. Some suggest a simple bar chart showing which words are most common is more useful. In the case of the Kremlin website data, the simple bar chart revealed a few, common, words, so the word cloud actually gave a better sense of what the focus of the transcripts was.

Other exploratory data analysis approaches for natural language processing (NLP) do not lend themselves as well to static data visualizations. Topic modeling through Latent Dirichlet Allocation (LDA), for example, allows for the grouping of words in the dataset by topic, determining which topics have the most importance for the collection, visualizing what words are in each topic, and how much the topics overlap (Shahul, 2023). In this dataset, the LDA showed four topics, with the largest overlapping to some degree with the third largest. It is difficult to tell from the list of 30 most common terms in each topic what the topics are about. The words include 'courageous,' 'wrenching,' 'upheaval,' 'distinguished,' etc., so they sound like they could be some sort of inspiring call to action. A basic sentiment analysis of the corpus, however, using TextBlob suggests the sentiment of almost all the texts is positive (or neutral). VADER analysis concludes the texts are all neutral. This, combined with the LDA, suggests that Putin and Medvedev use positive or neutral language when communicating with the Russian population---regardless of the topic. These theories will be tested with more sophisticated sentiment analysis in the modelling section.

Data Visualization 1

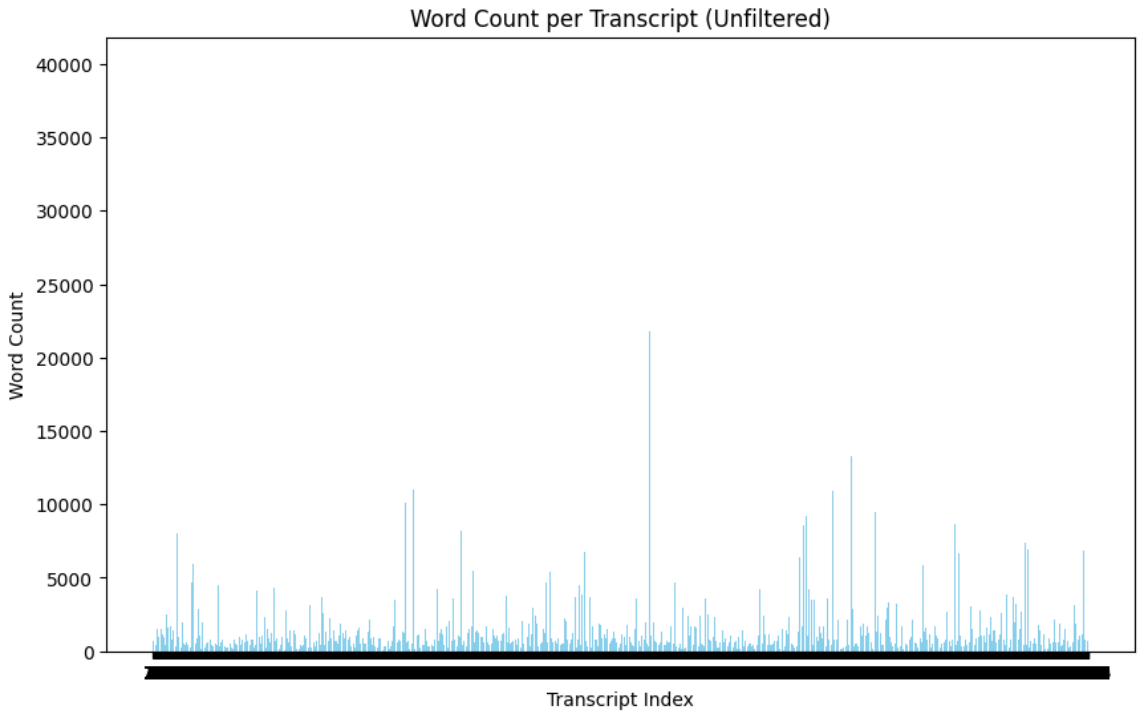

Figure 1: Word count of every transcript, note the outlier in the middle.

Figure 1 shows the word count (on the y-axis) of every transcript in the dataset ordered by transcript id number (which is in chronological order). In effect, this shows transcript word count over time. The most notable thing about this plot is the presence of a few spikes where the word count is noticeably higher. This includes the very noticeable outlier in the center of the plot that represents Putin's direct line call with the Russian public in April 2013. Overall there is no noticeable trend in the length of the transcripts. With the exception of the one outlier, the lengths appear fairly consistent over time.

The presence and timing of the huge outlier is important, however. Given its timing (around the time of the Bolotnaya protests) and its nature (a "direct line" call in show with the Russian public), it suggests that words may be important for understanding the Kremlin. At a minimum, it suggests that Putin responded with a great number of words to political unrest. Drilling down into the other high word count transcripts confirms this. The top ten transcripts by word count were either direct lines or press conferences. Those are the regularly placed spikes in the transcript by word count plot.

What this plot does not reveal, however, is what words or what type of words the Kremlin uses for any given purpose. It also does not reveal what the sentiment of those transcripts is. Nor does it show if that sentiment is different from other transcripts. By clearly showing outliers, however, it suggests that comparing the sentiment of those transcripts with a randomly selected equal

number of non-outlier transcripts may be useful. It may also be useful to plot the longest transcripts on an x-axis of time, to identify what events may have prompted (or been prompted by) the long transcripts. After running a sentiment analysis model, it would also be interesting to see this plot with the transcripts coded by color according to sentiment.

Data Visualization 2

Figure 2: Word cloud of all words in transcripts

Figure 2 is a word cloud of the most common words in the entire corpus of transcripts. The size of the words reflects the frequency of the words in the text. The larger the word, the more often it appears. "Vladimir" and "Putin" are the largest words, while other large words include "Russia" and "Russian" "Federation." The color of the words does not signify anything and is merely used to make the words easier to distinguish from each other.

The word cloud comports with a term-frequency (tf) analysis that found the following were the ten most common words (with the number of occurrences): 'vladimir': 48739, 'putin': 35783, 'dear': 11844, 'friends': 17842, 'new': 30501, 'year': 50089, 'eve': 15768, 'family': 16483, 'planned': 33930, 'gather': 18297. The words are all in lower case because the word cloud and term frequency analysis were conducted after the dataset was normalized by converting all text to lowercase. A term frequency-inverse document frequency (tf-idf) analysis reveals the most important terms. Applied to an entire collection of all transcripts, the most important words in each document are given a score and then ranked overall against all scores in the dataset. The most important words in this dataset are 'specific', 'nadeshda', 'okolnaya', 'would', 'russian', and

'russia.' ("Nadeshda" and "Okolnaya" are clearly russian words, so it is unclear what they are doing in the engligh-language transcripts (unless they are proper names).)

The words in the word cloud (and in the tf and tf-idf analysis) do not reveal much information. Given the source of the data, it is entirely unsurprising that 'vladimir,' 'putin,' and 'russia' appear most frequently in the texts. One noteworthy thing from the plot, however, is that the only countries other than Russia that appear by name are Ukraine and the United States. The word "respect" also shows up, which is interesting because many international relations theorists have surmised that the search for international respect is one of the drivers of Russian foreign policy (Elgin, 2020). N-gram analysis shows similar results to the tf analysis. The most common bi-gram is "vladimir putin" and the most common tri-gram is "russia vladimir putin."

Data Visualization 3

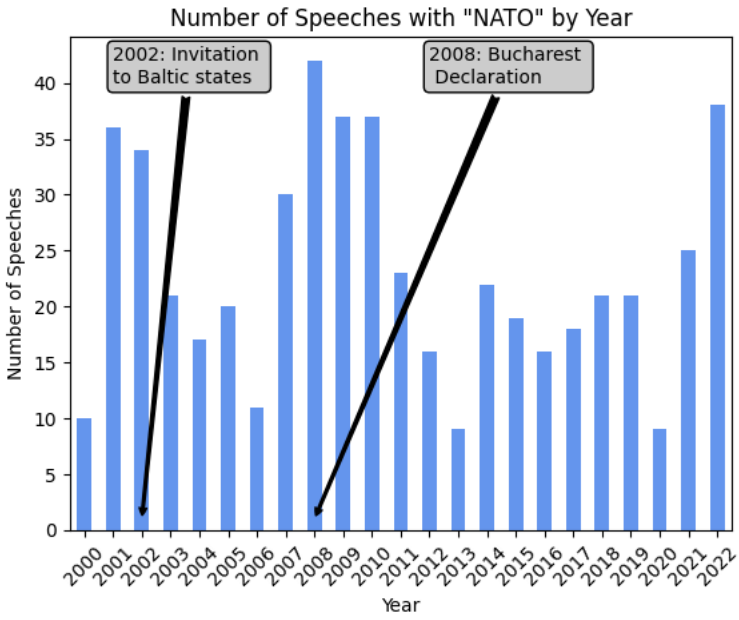

Figure 3: Transcript frequency by keyword "NATO"

Figure 3 shows the result of a frequency analysis on a smaller dataset. This dataset resulted from filtering the original dataset to include only those where transcripts mentioned the word "NATO." The speeches were then grouped by year, and the count of transcripts per year was plotted. The y-axis represents the number of speeches, the x-axis represents the year. The plot is annotated to show two important dates. The first is 2002, the date at which NATO issued an invitation for the former Soviet Baltic states to join. The second is 2008, the date of the Bucharest Declaration,

when NATO said the former Soviet republics Ukraine and Georgia would become NATO members in the future.

The plot suggests a correlation between Kremlin references to NATO and NATO expansion (or announcement of expansion) into the former Soviet Union. In the lead up to and immediate aftermath of the 2002 invitation to the first former Soviet states to join the Alliance, there is a noticeable increase in the number of transcripts mentioning NATO. Only ten transcripts mentioned NATO in 2001. In 2002, that number exceeded 35. In 2003, there were more than 30 transcripts mentioning NATO. By 2004, the number of transcripts mentioning NATO drops to about 20 and remains at or below that level for the next few years.

The number of transcripts mentioning NATO jumps again in 2007 in the lead up to the 2008 NATO Bucharest Summit at which the alliance was debating issuing an invitation to the former Soviet states of Ukraine and Georgia (Tharoor, 2023, July 10). In April 2008, the allies agreed to issue a statement saying that the two states would become NATO members---the Bucharest Declaration. The number of transcripts mentioning NATO in the dataset rises to its peak of over 40 in 2008, remaining higher than 35 until 2011 when it drops to 2004 levels again. The number remains below 25 until 2021 and 2022. In 2021 it increases to 25 and in 2002 it jumps to over 35. This coincides with Russia's February 2022 full-scale invasion of Ukraine, following a demand that NATO withdraw its infrastructure back to 1997 locations and that Ukraine never be allowed to enter the alliance (Roth, 2021, December 18).

Data Visualization 4

Figure 4: Transcript frequency by keyword "Ukraine"

Figure 4 is a plot of a frequency analysis of a dataset derived from filtering the overall dataset to include only transcripts that mention the word "Ukraine." The transcripts are grouped by year. The y-axis shows the count of transcripts in each group. The x-axis shows the year that the transcripts were issued. The plot is annotated to show two important dates in the history of Ukraine-Russian relations. The first date is 2014, when Russia invaded Ukraine, annexed Crimea, and began a long-term war in the eastern Ukrainian Donbass region. The second date is 2022, when Russia recognized the independence of the two Donbass regions (Donetsk and Luhansk) and launched a full-scale invasion of Ukraine.

Since Figure 2 showed Ukraine as one of only two countries besides Russia in the word cloud plot, a more detailed analysis of the appearance of the word "Ukraine" in the Kremlin transcripts seemed worthwhile. Figure 4 shows an interesting pattern. First, Ukraine does not appear in many Kremlin transcripts until 2004, the year of the "Orange Revolution, " when the Ukrainian people overthrew their Kremlin-backed president (Dickinson, 2020). That year there is a peak in the frequency of documents with "Ukraine." Similarly, there is another peak in 2008-2009, following the Bucharest Declaration. The most notable change, however, is in 2014, when the number of transcripts mentioning "Ukraine" skyrockets to over 100.

Following the dramatic increase in mentions of "Ukraine" in 2014, the plot shows a steady decline in the number of transcripts referencing "Ukraine." By 2020, there are barely over 20

transcripts that include the word. There is another increase in 2021, as the numbers increase above 30. The number of transcripts including "Ukraine" skyrocket yet again in 2022. That year the number jumps to above 80. It is also the year of Russia's full-scale invasion. A plot grouped by month and year, rather than just year, would be useful in showing if the 2021 increase occurs at the same time as the Russian massing of forces on the Ukraine border.

Data Visualization 5

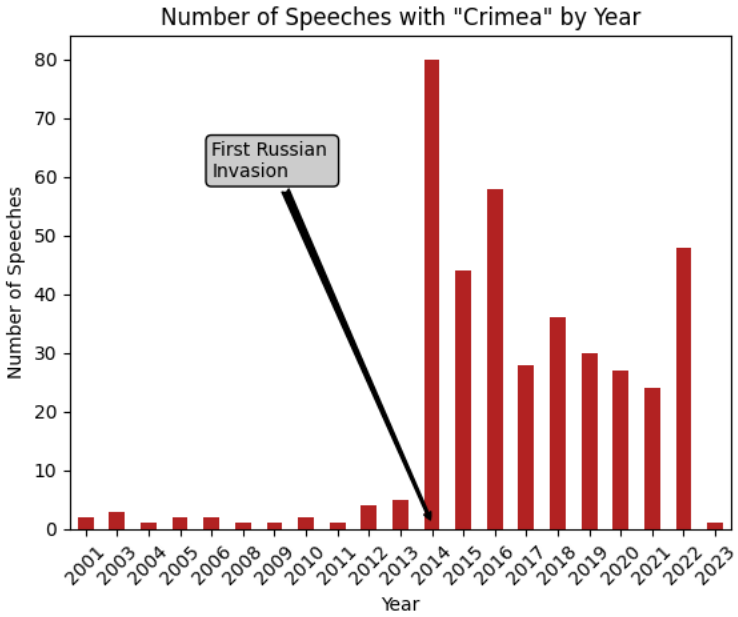

Figure 5: Transcript frequency by keyword "Crimea"

Figure 5, like figures 3 and 4, shows a frequency analysis of a subset of the overall dataset. This subset has been filtered by transcripts including the word "Crimea." The transcripts are grouped by year. The y-axis shows the number of speeches in each group. The x-axis shows the year. The plot is annotated to show the year of the first Russian invasion of Ukraine. That is also the year in which Russia occupied and annexed Crimea.

Crimea appears to have barely showed up on the Kremlin's radar screen prior to 2014. The number of transcripts including the word are consistently below five. This is a little surprising considering Crimea was home to Russia's Black Sea fleet, and Russia needed to negotiate with Ukraine to lease the fleet's Crimean port of Sevastopol. The Ukrainian president extended the lease in 2010, which may explain the tiny increase in transcripts referencing "Crimea" that year. There was a slight increase in mentions of "Crimea" in 2012 and 2013. Overall, however, it was not a topic for many of the transcripts on the Kremlin website.

The number of transcripts referencing “Crimea” increased dramatically in 2014. That year there were over 80 transcripts with the word “Crimea” in it. That is in sharp contrast to the preceding years in which the number never reached 10. This increase coincides with Russian military occupation and annexation of Crimea. In the ensuing years, the number of transcripts referencing Crimea remained high, never decreasing below 20. The number did gradually decrease, however. This coincided with a perceived diminution in the impact of the Crimean annexation on Putin’s popularity ratings (Goncharenko, 2019). It is unclear, though, if the Kremlin chose to reduce references to Crimea because it was becoming less popular, or if other factors caused that reduction. The number of references to Crimea increased dramatically again, however, with Russia’s full-scale invasion of Ukraine in 2022.

Data Modeling

Data Modeling Definitions

There are multiple possible approaches to perform sentiment analysis on texts. Snegovaya (2020) used an ordinary least squares (OLS) regression based upon a Word2Vec algorithm to identify which of Putin's speeches were "aggressive." In contrast, Nisch (2023) used a Bidirectional Encoder Representations from Transformers (BERT) model to label the emotions in Zelensky's tweets following the Russian full-scale invasion of that country. BERT is a pre-trained model (trained on English-language Wikipedia and the Brown Corpus (Lutkevich, 2020)), which many users have adapted to work on specific types of text. Liu (2019) adapted and significantly improved upon BERT by essentially training the model for longer, in larger batches, and on a larger dataset. Of note, they used the "CommonCrawl News Dataset," which should make their model, RoBERTa, perform better on the political texts on the Kremlin website.

Hugging Face (https://huggingface.co/) develops this model refinement further by leveraging a community of users who train models on other datasets and upload their results to the hub. The result is a "transfer learning" methodology that should save researchers considerable time and effort (Donges, 2022). This produces outstanding models for sentiment analysis. Hartmann et al. (2021) created a RoBERTa -based model fine-tuned on 5,304 manually annotated social media posts that classifies a text into 3 categories: positive, negative, and neutral. This model: "sentiment-roberta-large-english-3-classes" (Hartmann) is the first model used below. In Hartmann's pre-training, the model had a hold-out accuracy of 86.1%.

The second model is also created by Hartmann (2022). It is "Emotion English DistilRoBERTa-base." A DistilBERT or DistilRoBERTa model is a smaller, pre-trained model (Silva Barbon & Akabane, 2022), that should perform its task much more quickly than the full RoBERTa model. The Hartmann model is pre-trained on six datasets, which include representations from seven emotions (Hartmann). He selected a balanced subset of 2,811 examples for each emotion, resulting in approximately 20,000 total cases. In his pre-training, the model returned an evaluation accuracy of 66%.

Data Model 1

The first model used was the "sentiment-roberta-large-english-3-classes" (Hartmann). This is a RoBERTa model with a large number of parameters. Its architecture is "RobertaForSequenceClassification." The configuration file for this model includes the following initial hyperparameters:

A dropout probability for attention probabilities of 0.1; a beginning of sequence token id of 0 and an end of sequence token id of 2; a hidden layer activation function of GELU (Gaussian Error Linear Unit); a hidden layers dropout probability of 0.1; the dimensionality of the hidden layers is 1024; the weight initialization range is 0.2; the dimensionality of the intermediate (feed-forward) layers is 4096; the epsilon value for layer normalization is 1e-05; the maximum position embeddings are 514; the number of attention heads are 16; the number of hidden layers is 24; the pad token id is 1; the size of the token type vocabulary is 1; and the size of the model's total vocabulary is 50265.

The training arguments supplied to the model initially were: four training epochs; the learning rate was 2e-5; the batch sizes for training and evaluation were 4 (larger batch sizes overloaded the memory on Google Colab); weight decay was 0.01; the evaluation strategy was once at the end of each epoch; and the metric for best model was accuracy.

The initial dataset of Kremlin transcripts was unlabeled. In pre-processing, the data was labeled using TextBlob, creating three categories: negative, neutral, and positive. This enabled the use of this three-class sentiment model. The initial training/testing/and validation of the model produced an accuracy of over 99%. The dataset, however, was significantly unbalanced. As a result, the model achieved near perfect precision, recall, and F1 scores identifying the positive cases, but scores of 0 when identifying the negative class. In other words, the model almost always labeled the text as positive.

After balancing the dataset through random under sampling and through taking a percentage of a randomly over sampled dataset, the model achieved an accuracy of 71%, an F1 of 0.67 and 0.75, a precision of 0.57 and 0.86, and a recall of 0.8 and 0.67 on the validation data.

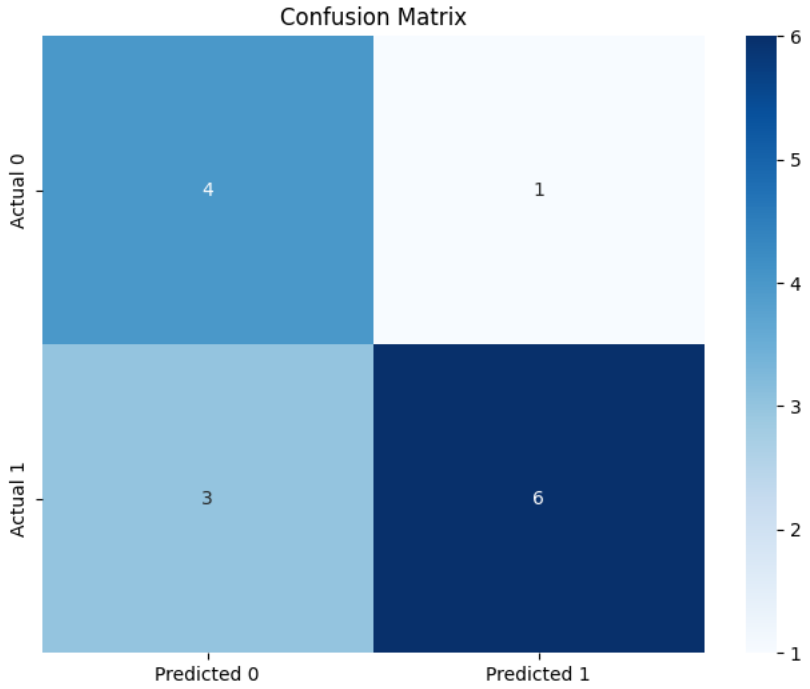

Figure 6: Confusion Matrix for Balanced Data and Model 1 prior to fine-tuning

Fine tuning the hyperparameters was conducted using an Optuna hyperparameter search. The initial hyperparameters were set as described above. The Optuna trainer was set to maximize the accuracy of the model through a run of ten trials. The trainer adjusted batch size, number of training epochs, and learning rate. It returned the best model as one with an accuracy of 88% and hyperparameters of: 'learning rate': 7.369e-06, 'number of training epochs': 3, 'seed': 29, and 'per device training batch size': 32.

The hyperparameter search process was repeated with the goal of maximizing the model's F1 score. The best model returned a macro average F1 score of 83% and had the parameters: 'learning rate': 4.65e-05, 'number of training epochs': 5, 'per device training batch size': 16.

Data Model 2

The second model used was Hartmann's "emotion-english-distilroberta-base" model (Hartmann). This model labels texts as one of seven emotions: "anger", "disgust", "fear", "joy", "neutral", "sadness", and "surprise." Like the previous model, it has a "RobertaForSequenceClassification" architecture. The configuration file for this model includes the following initial hyperparameters:

A dropout probability for attention probabilities of 0.1; a beginning of sequence token id of 0 and an end of sequence token id of 2; a hidden layer activation function of GELU (Gaussian Error Linear Unit); a hidden layers dropout probability of 0.1; the dimensionality of the hidden layers is 768; the weight initialization range is 0.02; the dimensionality of the intermediate (feed-forward) layers is 3072; the epsilon value for layer normalization is 1e-05; the maximum position embeddings are 514; the number of attention heads are 12; the number of hidden layers is 6; the pad token id is 1; the size of the token type vocabulary is 1; and the size of the model's total vocabulary is 50265.

The training arguments supplied to the model initially were: two training epochs; the learning rate was 2e-5; the batch sizes for training and evaluation were 16 (a smaller model with fewer training epochs meant it was possible to use larger batch sizes); weight decay was 0.01; the evaluation strategy was once at the end of each epoch; and the metric for best model was accuracy.

The dataset was labeled only for positive, neutral, and negative classes. This model, however, required seven different classes of emotions. With a dataset of over 90,000 rows, manually labeling a sample for training would have been the "gold-standard" for modeling but was not feasible for this project. The "emotion-english-distilroberta-base" model is, however, a pre-trained language model, which means it can label an unlabeled dataset with high accuracy. To test the consistency of that accuracy, however, a technique called "pseudo-labeling" was implemented. In this technique, the data was labeled by the model and then shuffled and split into three datasets: train, test, and validate. The training process was then repeated to see if the model was consistent in how it labeled the data.

Using this methodology, the model had an accuracy on the validation dataset of 86%. The F1 score for each class ranged from 80% to 91%. The precision for each class ranged from 77% to 91%. The recall for each class ranged from 75% to 94%.

The model's hyperparameters were fine-tuned using the Optuna hyperparameter search, with the goal of maximizing the model's accuracy. The training arguments: learning rate, batch size, and number of training epochs were adjusted through each of ten epochs. As a result, the best model had the parameters: 'learning rate': 4.48e-05, 'number of training epochs': 5, 'per device training batch size': 16. It returned an accuracy of 87%.

The hyperparameter search process was repeated with the goal of maximizing the model's F1 score, which was 84%. The best model had the parameters: 'learning rate': 1.25e-05, 'number of training epochs': 2, 'per device training batch size': 16.

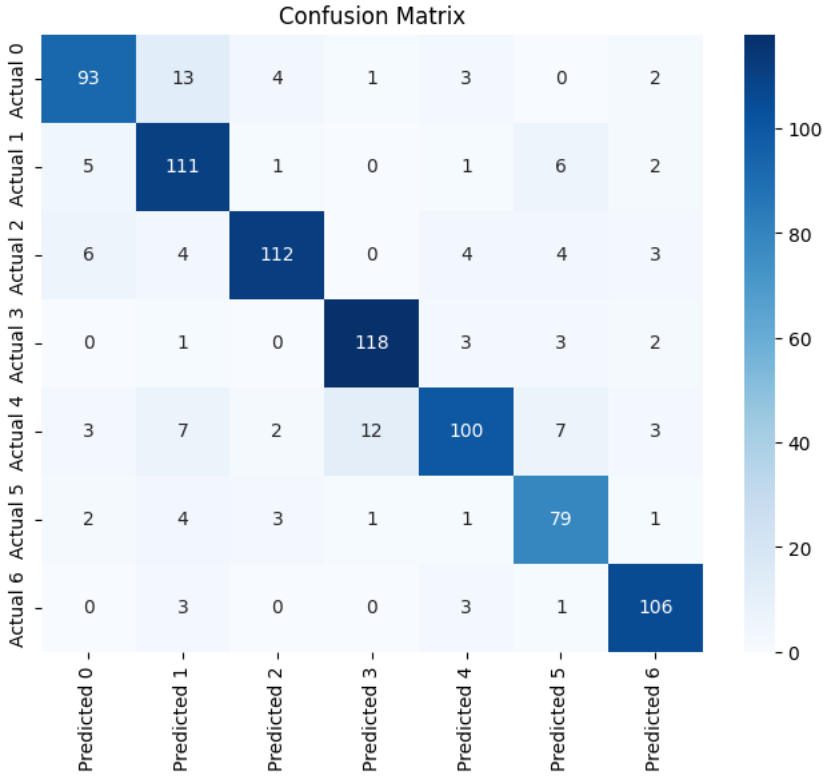

Figure 7: Confusion Matrix for "emotion-english-distilroberta-base" prior to fine-tuning

Review of Data Models

Both models performed well on their classification tasks, each achieving an accuracy on balanced data of 87-88%. Their tasks were slightly different, so it is difficult to compare them directly. The first model labeled the sentiment of the Kremlin transcripts as one of three classes: negative, neutral, and positive. With unbalanced data, it had very high accuracy, but lower F1 scores of 67% and 75%, which indicates a lower precision and recall. Once the data was balanced and the hyperparameters tuned, the model had an accuracy of 88% and an average F1 score of 83%. The second model had a presumably more difficult task of labeling the data as one of seven emotions. With fine-tuned hyperparameters, the model had an accuracy of 87% and an average F1 score of 84%.

The accuracy and F1 scores are only two measures of a model's utility, however. The size of the model is also a consideration when using finite computing resources. The second model was smaller, using a DistilBERT foundation. As a result, it took up less of the GPU memory, allowed

for larger batch sizes, and trained the data faster. This was a definite benefit, with only a slight trade-off in accuracy. In addition, it provided more nuanced assessments of the sentiment of each Kremlin transcript. In that regard, it better suited the goals of the initial task of this project.

Neither of these models was, however, a champion model in terms of the definition: "The champion model is the best predictive model that is chosen from a pool of candidate models (*SAS Help Center: Champion Models*, 2023)." Both were equally effective at their predictive classification task. The best model would, therefore, depend upon the goals of the task. In this case, the task is to identify the emotions in Kremlin transcripts to see if they correspond with any Russian foreign policy action. A model that provides more emotion categories is more useful for accomplishing this, so model number two is preferable.

Final Results

Assignment 5

Findings

There was little difference in the performance of the two models, but the "emotion-english-distilroberta-base" was able to label more categories of emotions. Those seven categories of emotion provided the potential for more insight into the tone of transcripts on the Kremlin website from 1999 to 2023. The first step in obtaining this insight was to create visualizations showing emotions over time (by year).

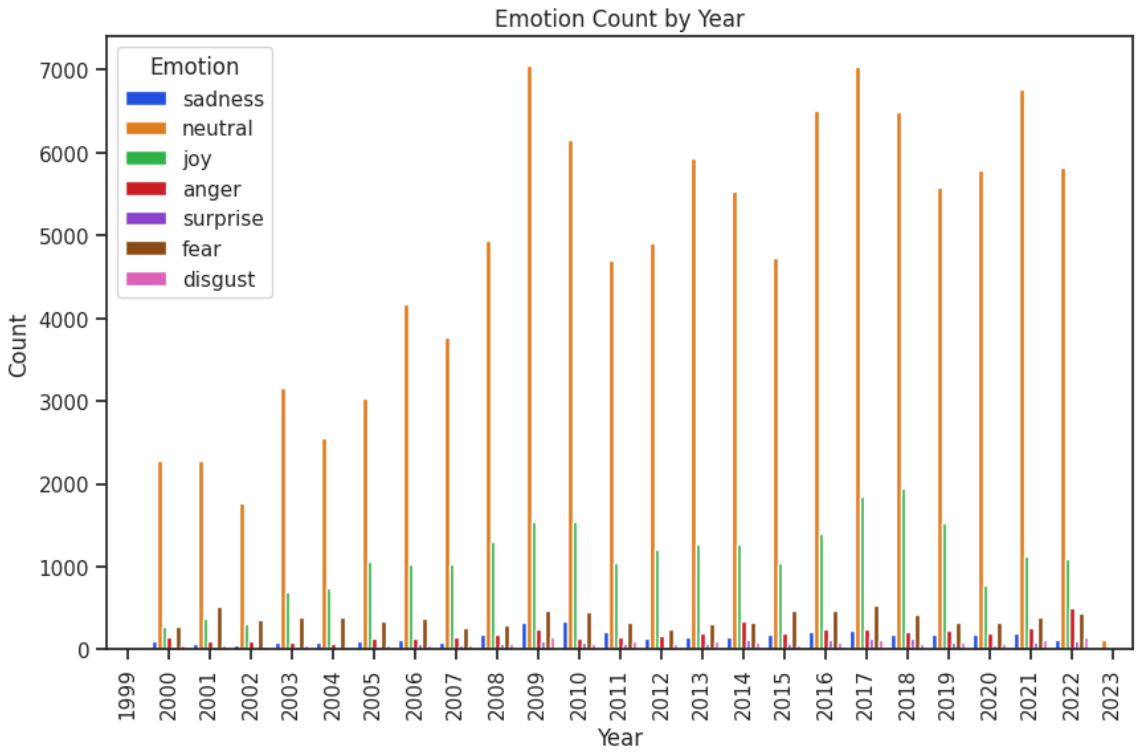

Figure 8: Emotion count by year in all transcripts

Figure 8 shows what appears to be a pattern, but the details are drowned out by the dominance of the "neutral" category. Plotting the individual emotions by year reveals more detail.

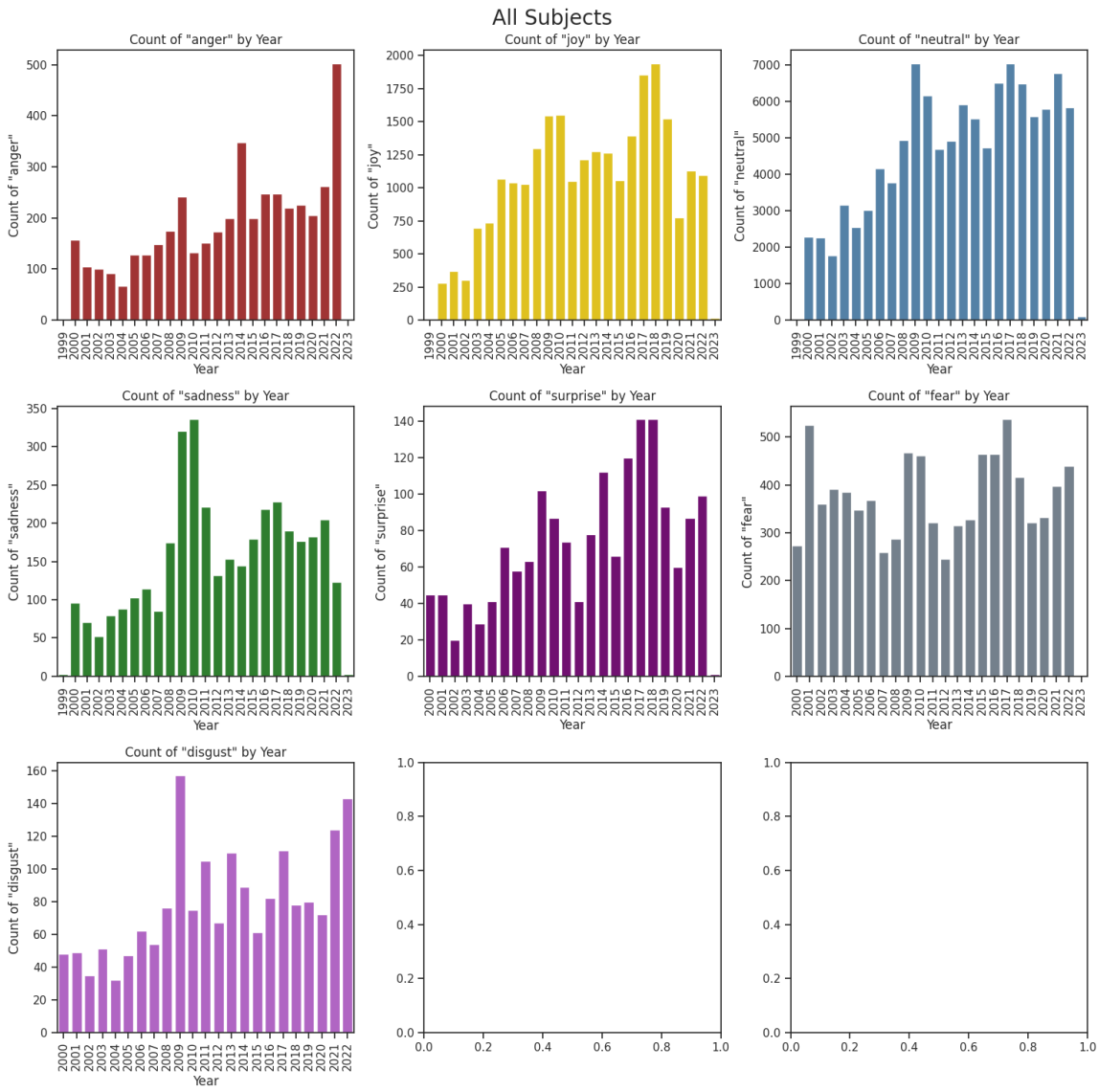

Figure 9: Gridplot of each emotion by year

Figure 9 shows a rough pattern in the appearance of each emotion, though it still does not provide too much specific insight. Certain peaks do suggest areas for further exploration. For example, a significant increase in "disgust" in 2009, 2021, and 2022, suggests those would be interesting years to look into further. In addition, 2009 and 2010 see a marked increase in "sadness" and "fear," while anger increases in 2014 and 2022.

Given the history of Russian foreign relations, the years 2009, 2010, 2014, 2021, and 2022 suggest certain phrases may be of interest. Specifically, NATO's Bucharest Declaration took place in 2008; the first Russian invasion of Ukraine took place in 2014; and the second invasion of Ukraine was in 2022 after a year of military buildup.

Filtering the datasets by the phrases “Ukraine,” “Crimea,” “NATO,” and the “United States” (and variants thereof), provided some useful information.

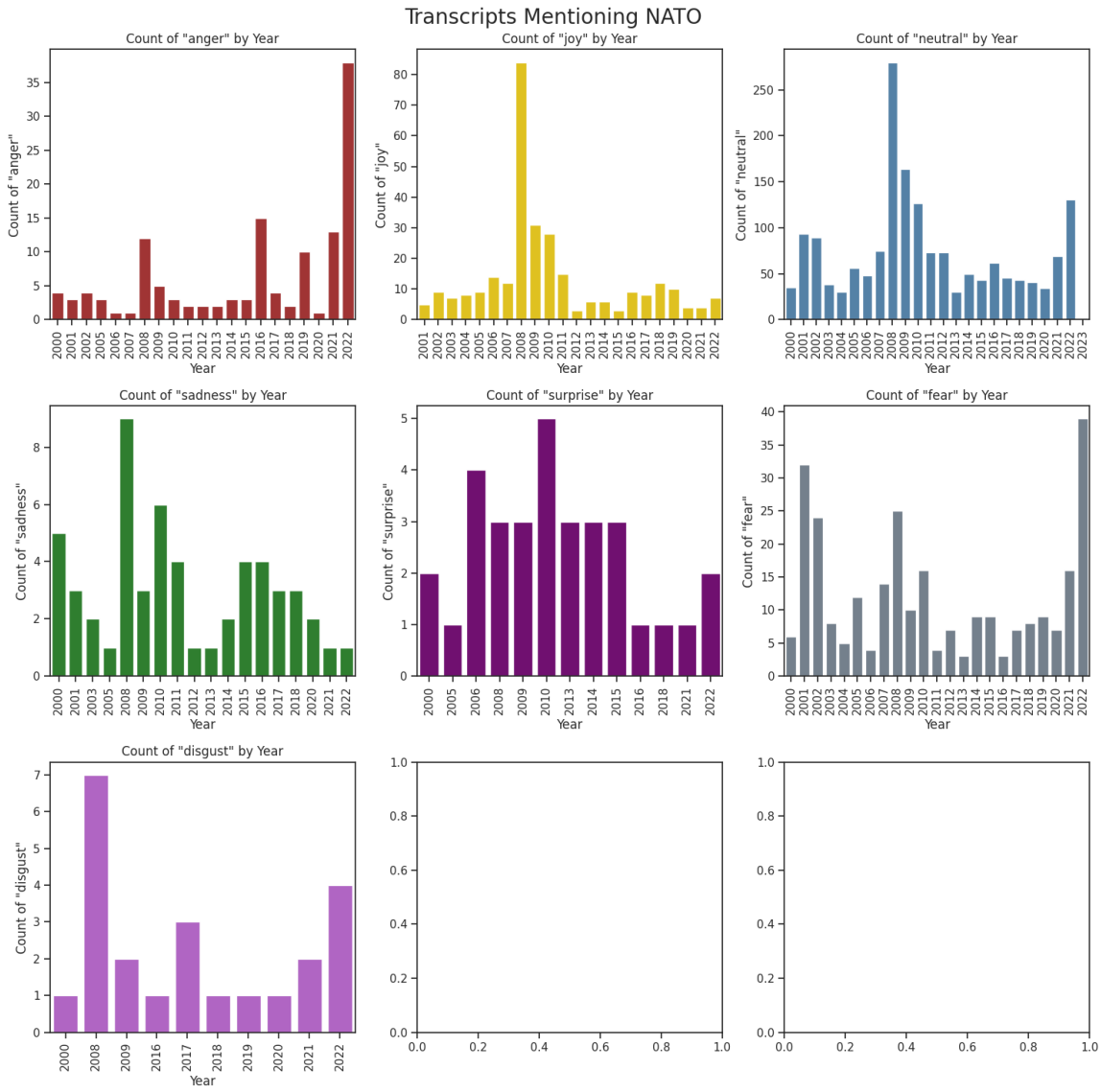

Figure 10: Emotions by year for transcripts containing "NATO"

Figure 10 shows some interesting results. Joy is the most common emotion in transcripts about NATO in 2008. At the same time, disgust and sadness also peak in transcripts about NATO that year. Unsurprisingly, anger and fear are the most common emotions in transcripts about NATO in 2022---the year of the Russian full-scale invasion of Ukraine.

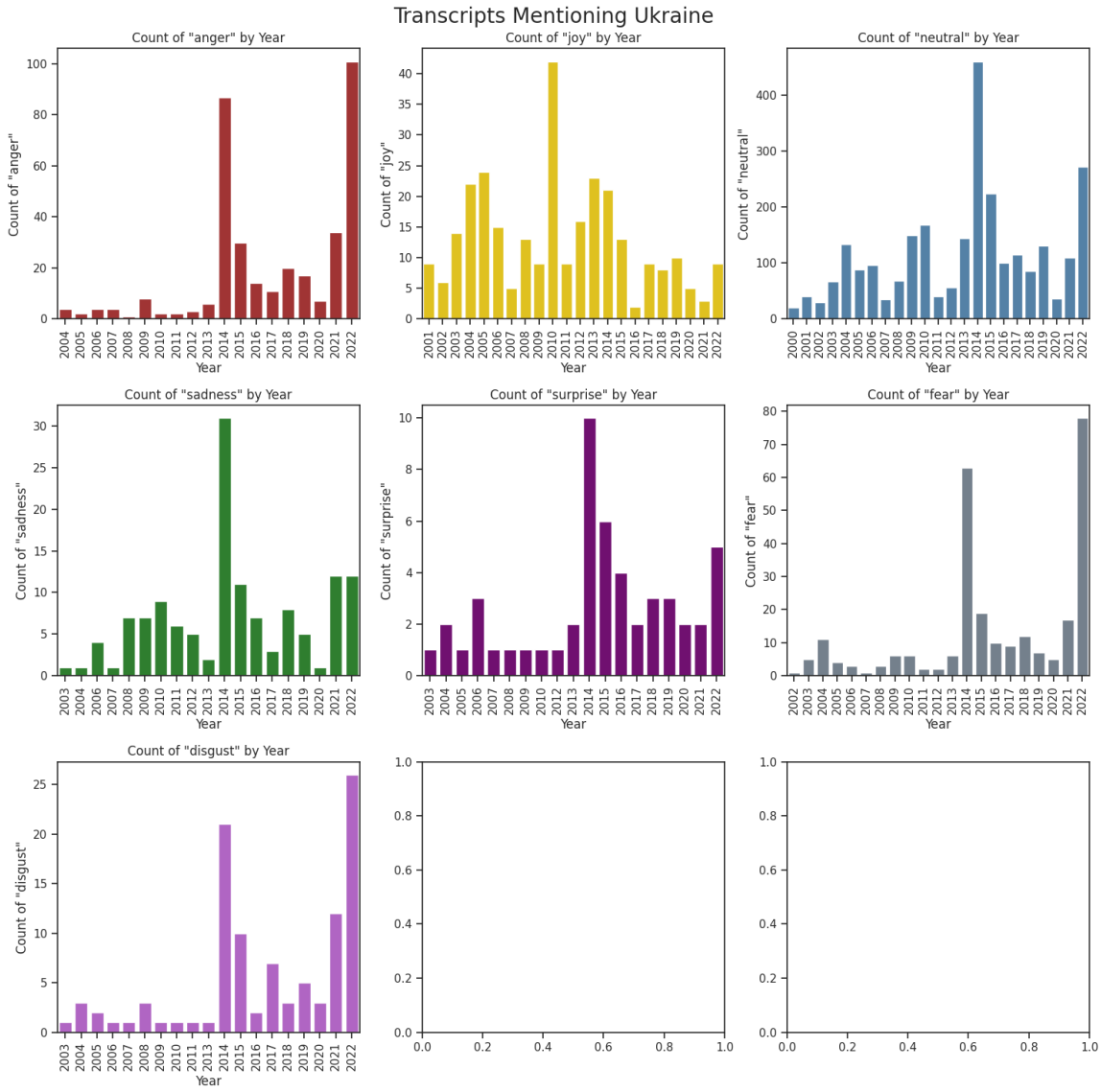

Figure 11: Emotions by year in transcripts mentioning Ukraine

When transcripts are filtered by the word "Ukraine," other interesting details are also shown. First, the year 2014 shows a peak in several emotional categories. Anger is the most common emotion that year, followed by fear and sadness. Disgust occurs fewer times than the other emotions, but appears more frequently in 2014 than in previous years. As with transcripts filtered by NATO, fear and anger (followed by disgust) are the most common emotions in 2022.

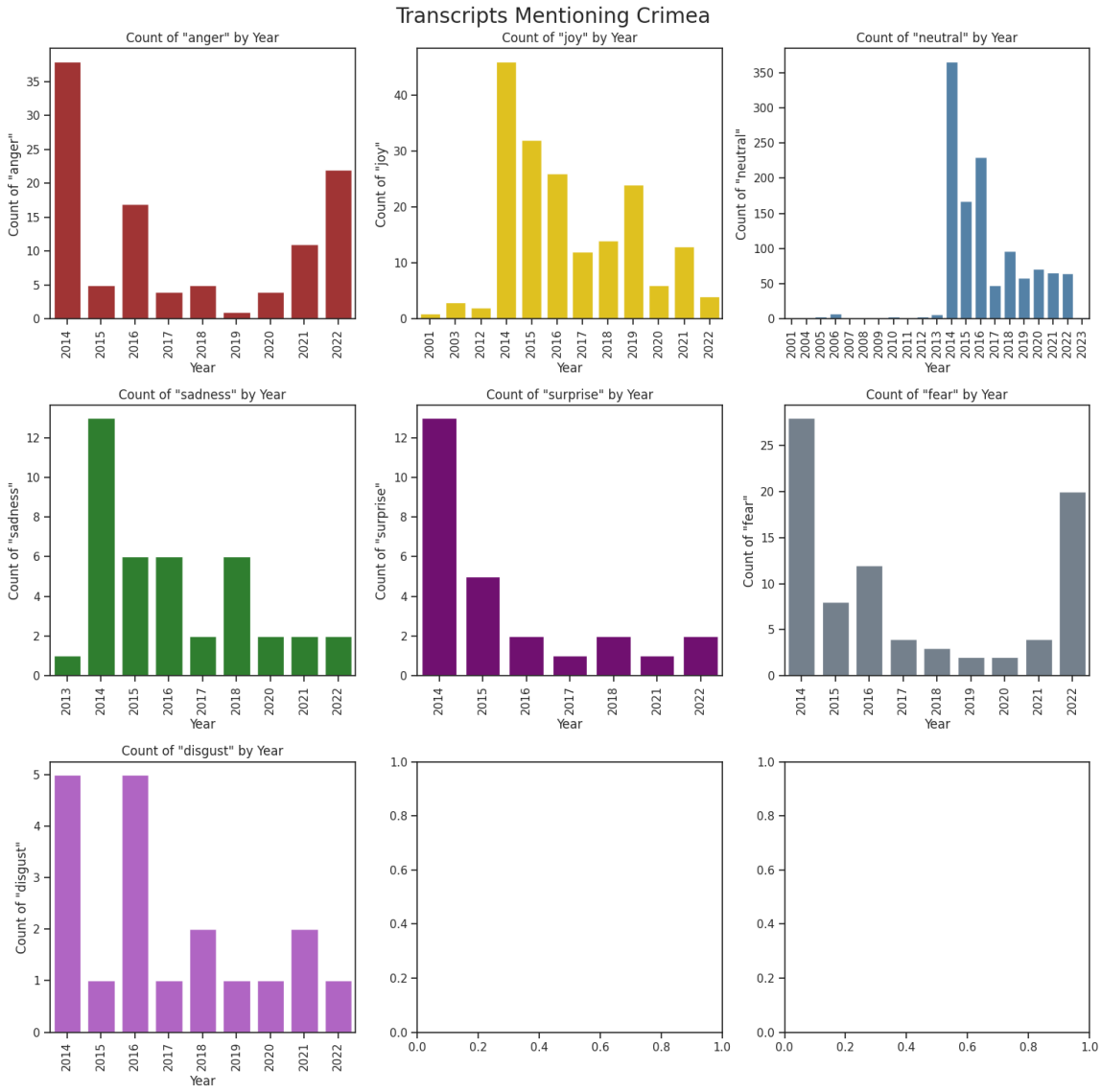

Figure 12: Emotions by year in transcripts mentioning Crimea

Figure 12 shows that Crimea does not appear often in transcripts prior to 2014 (the year Russia occupied and annexed it). In 2014, the dominant emotion in Kremlin transcripts mentioning Crimea was joy, followed by anger and fear. Anger and fear also appear more frequently in 2022.

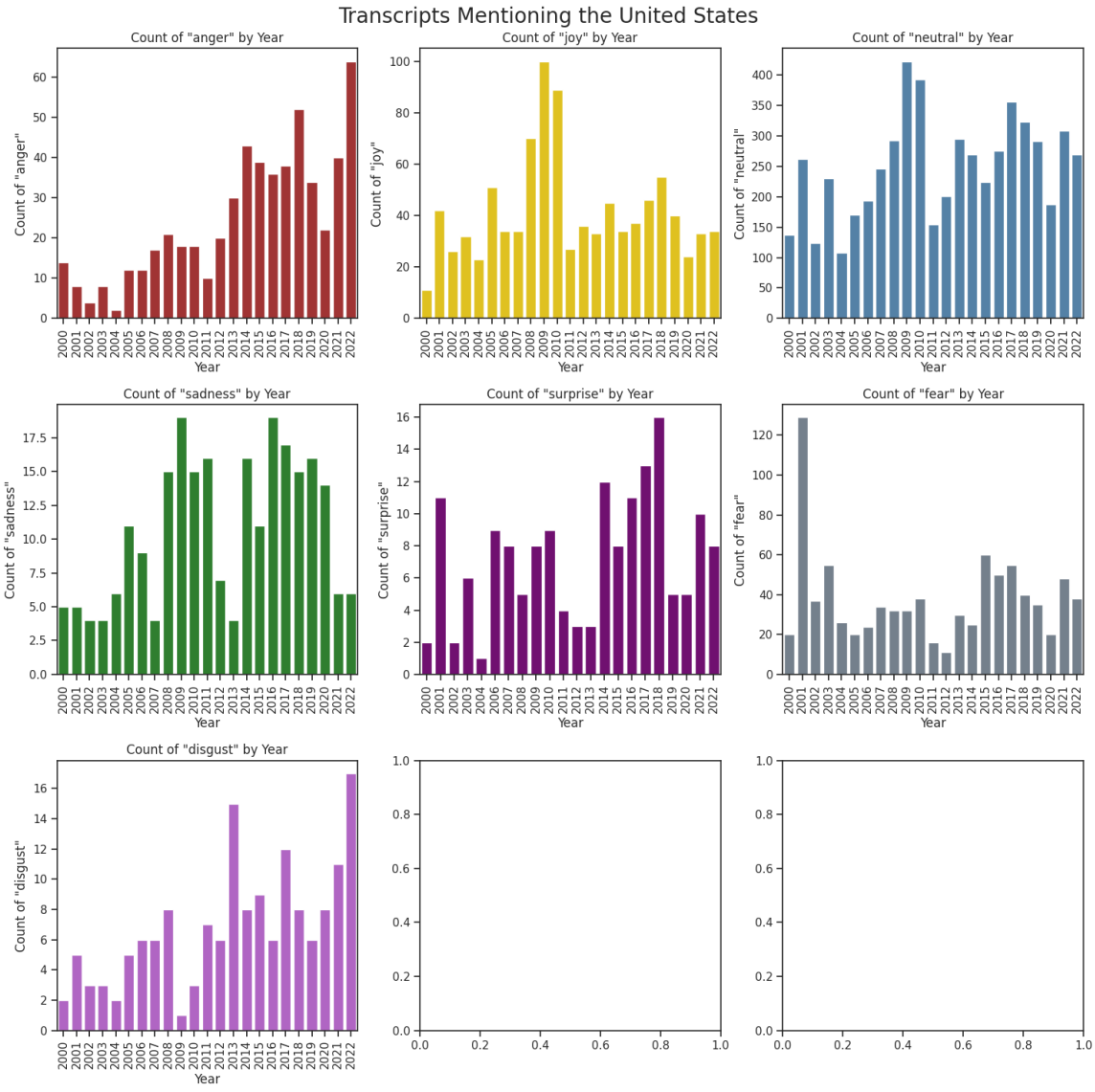

Figure 13: Emotion by year in transcripts mentioning the United States

Figure 13 reveals a more complex picture of emotions in Kremlin transcripts that reference the United States. In 2001, there is a peak of fear. This probably reflects the impact of the 2001 terror attacks against the United States and the resultant war in Afghanistan. This insight could be tested by looking at the distribution of emotions by month that year. Joy is the most predominant emotion in 2009—the year of Barack Obama's election and the subsequent "reset" or relations between the United States and Russia. There is, however, also evident a somewhat steady increase in disgust and anger in transcripts referencing the United States from 2012 to 2022. (In 2012, Putin's Kremlin was faced with strong public protests for the first time when Putin returned to the presidency after four years as prime minister. Putin argued the United States

was behind these protests.) The counts on the y-axes indicate that, like with NATO and Ukraine, anger and fear are the most common emotions in 2021 and 2022.

The relationships suggested at by the visualizations above are confirmed by statistical analysis. A chi-2 test of the relationship between emotion label and year for the entire dataset shows that there is a statistically significant relationship between the two variables (Chi-square statistic: 3529.72; P-value: 0.00). The Cramer's V score, however, was just 0.0599, indicating that the strength of that relationship is minimal.

The picture, changes, however, when the neutral values are filtered out and the datasets are filtered by key phrase as above. When the dataset includes only transcripts that mention NATO, the chi-2 test continues to show a statistically significant relationship between year and emotion (Chi-square statistic: 364.079; P-value: 0.00). The Cramer's V score, though, is higher at 0.2509, indicating the relationship is moderately strong. Similarly, with a dataset filtered for Ukraine, the chi-2 test shows a statistically significant relationship (Chi-square statistic: 481.49; P -value: 0.00) and the Cramer's V score shows the relationship is moderately strong (Cramer's V: 0.25). The relationship is still significantly significant for Crimea (Chi-square statistic: 113.649; P-value: 0.000), but weak (Cramer's V: 0.154). This is similar to transcripts referencing the United States (Chi-square statistic: 536.07; P-value: 0.00; Cramer's V: 0.17).

Heatmaps were created to show which emotions had the strongest relationship with year for each dataset.

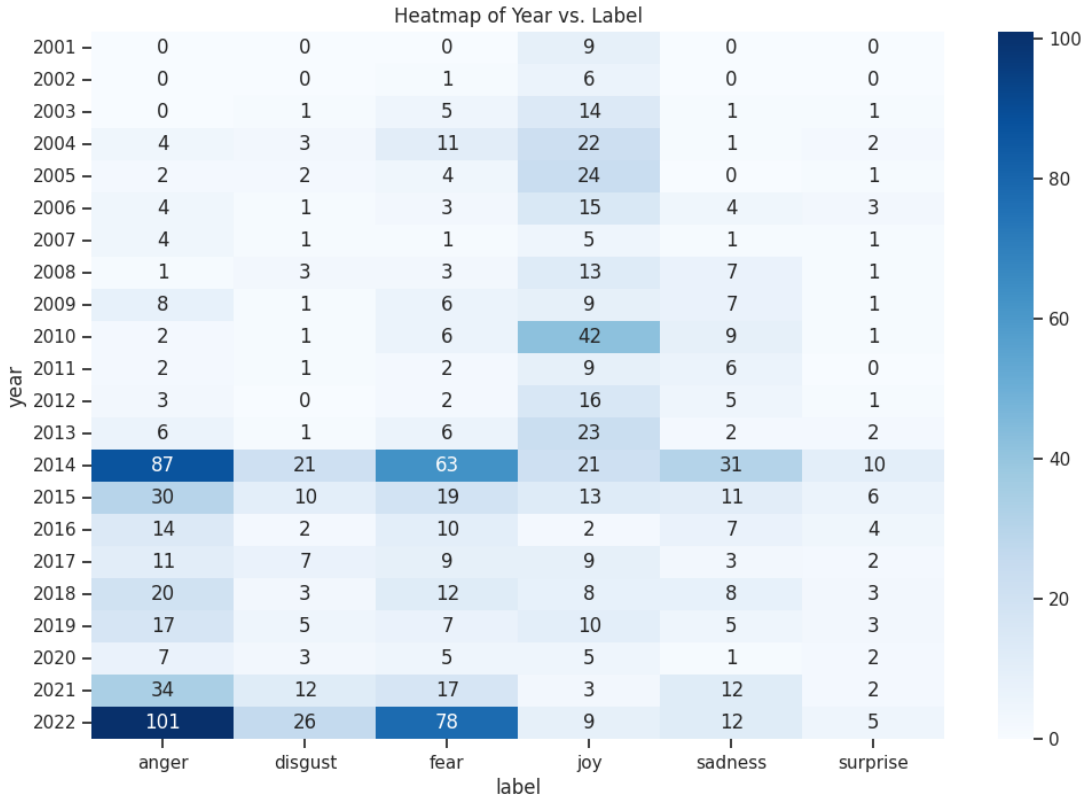

Figure 14: Relationship between year and emotion for Ukraine

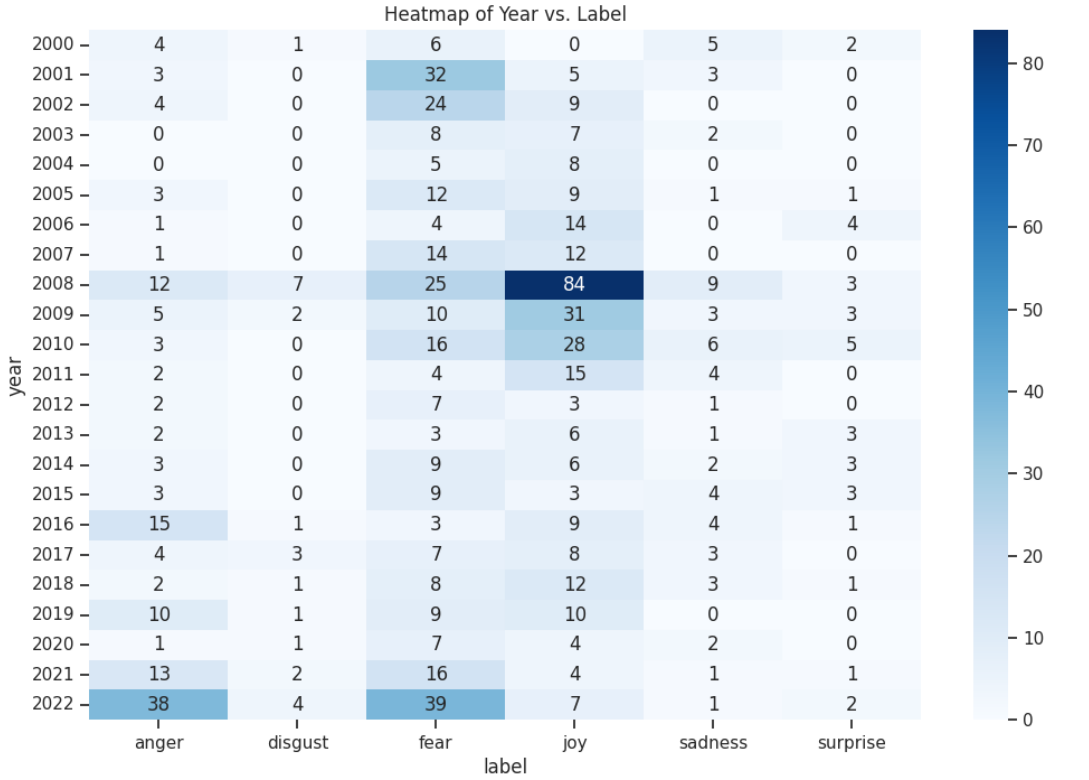

Figure 15: Relationship between year and emotion for NATO

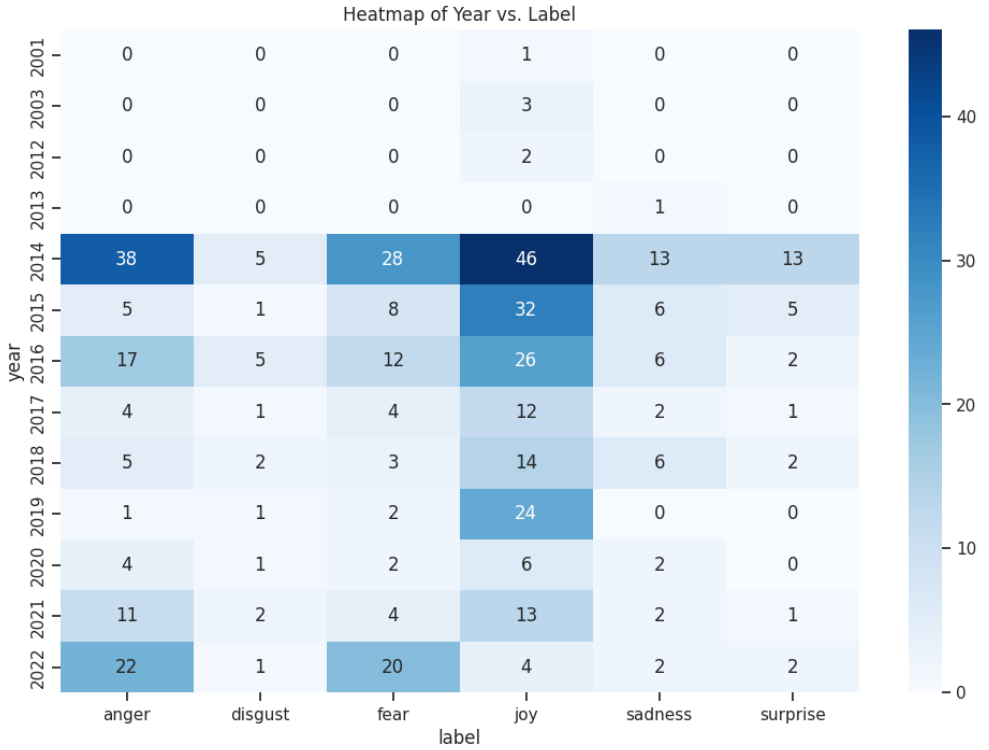

Figure 16: Relationship between year and emotion for Crimea

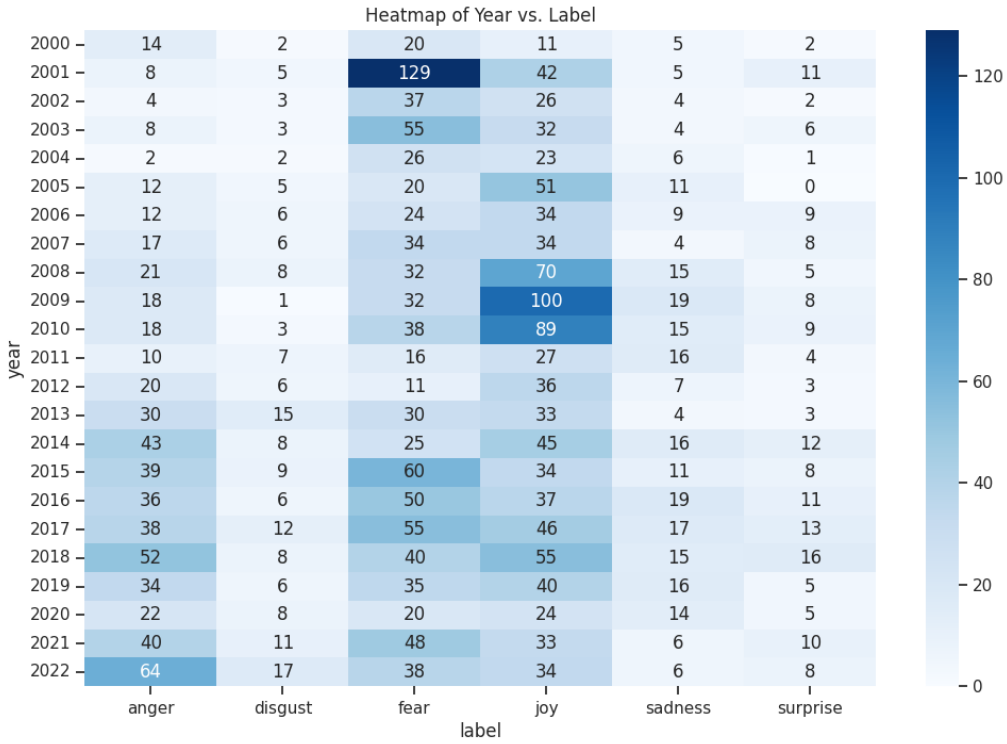

Figure 17: Relationship between year and emotion for United States

Review of Success or Completion

Modeling the datasets using the transformer "j-hartmann/emotion-english-distilroberta-base" resulted in useful emotion labels for each of the transcripts in the original dataset. These labels revealed that there is a statistically significant relationship between emotion and time and a potentially strong relationship between emotion labels and transcripts depending upon the time period and topic. Simple time and frequency analysis (using visualizations) showed patterns that suggested areas for further exploration. The initial visual exploration showed that there is a connection between year and the number of times the Kremlin uses certain words or phrases (suggesting the Kremlin uses words for a reason). After modeling the data to label it with emotions, the visualizations showed a relationship between emotion, time, and phrase. Statistical analysis showed this relationship was significant. Thus, the project showed that it is possible to extrapolate from years to actions (which is implied by the focus on 2014, 2022, and the invasions of Ukraine) and establish a relationship between words, emotions, and Russian foreign policy actions.

The results are tentative thus far, however. First, the dataset used was completely unlabeled. To conduct the modeling, it was labeled by machine. First, it was labeled as one of three sentiments using TextBlob. Second, it was labeled as one of seven emotions by the model that re-labeled it. This methodology could be improved by manually labeling the emotions of a percentage of the transcripts.

In addition, the interpretation of the results may reflect some cognitive biases. For instance, knowing what Russia did in Ukraine in 2014 and 2022, may have predisposed the researcher to see or exaggerate patterns that were not really in the data. Finally, this model was only 88% accurate and was pre-trained on social media. More accuracy could be gained by using a model that is trained on foreign-policy or other political transcripts.

Potential Data Privacy and Data Security Issues

The dataset used in this project was collected and maintained by dekoder.org. That organization scraped the transcripts from the Kremlin's public webpage. Per that website (kremlin.ru), all of the material on the website is available for public use as long as you cite the source of the data. Further, it is provided under Creative Commons 4.0 international license. Therefore, the only potential data privacy issue is if the Kremlin has posted information about individuals without their consent.

Data security is also of limited concern in this project due to the nature of the data and the analysis. All the data is publicly available. A potential concern, however, is that processing the

data, especially generating new data in order to balance the classes in the dataset, may risk corrupting it. This danger can be mitigated by maintaining the original dataset and comparing it to the results of the balanced data. Similarly, different methods can be used to balance the data that do not use random oversampling (and the generation of new data).

Recommendations for Future Analysis

As an initial exploration of the relationship between the emotion in Kremlin transcripts and the actions of Kremlin leaders, this project suggests numerous areas for future analysis.

First, this same data should be explored using different modeling methods. For example, it would be useful to fine tune a basic RoBERTa model from scratch, ensuring it is better able to assess the sentiment of large political transcripts as opposed to just social media. In addition, it would be interesting to compare the results of sentiment analysis using a transformer with sentiment analysis using a convolutional neural network (CNN). Using the original Russian-language dataset would also be an important contribution for determining if emotion is language agnostic.

In addition, a longer-term project could involve exploring the relationship between emotions and time period more closely. This could be done, for example, by highlighting the most used phrases in the texts and then analyzing the sentiment around them. It could also be done by exploring the sentiment around other key events in Russian domestic and foreign policy. The sentiment could also be analyzed by month or day if necessary.

Finally, it would be useful to compare how the Kremlin expresses emotion in transcripts on the official website and in communications in less official sources, like newspapers, television, and social media.

References

10 Years Since Bolotnaya, the Biggest Protests of the Putin Era. (2021). *The Moscow Times*. https://www.themoscowtimes.com/2021/12/09/10-years-since-bolotnaya-the-biggest-protests-of-the-putin-era-a75739

Dickinson, P. (2020). How Ukraine's Orange Revolution shaped twenty-first century geopolitics. https://www.atlanticcouncil.org/blogs/ukrainealert/how-ukraines-orange-revolution-shaped-twenty-first-century-geopolitics/

Donges, N. (2022, September 12, 2022). *What is Transfer Learning? Exploring the Popular Deep Learning Approach*. https://builtin.com/data-science/transfer-learning

Elgin, K. K. (2020). *Recognition and Respect: Understanding Russia's Defense of Its Great Power Status* [Academic dissertations (Ph.D.), Princeton University]. Princeton, NJ. http://arks.princeton.edu/ark:/88435/dsp01m039k797j

Goncharenko, R. (2019). What's left of the 'Crimea effect'? *dw.com*. https://www.dw.com/en/vladimir-putins-crimea-effect-ebbs-away-5-years-on/a-47941002

Hartmann, J. *j-hartmann/emotion-english-distilroberta-base · Hugging Face*. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

Hartmann, J. *j-hartmann/sentiment-roberta-large-english-3-classes · Hugging Face*. https://huggingface.co/j-hartmann/sentiment-roberta-large-english-3-classes

Hartmann, J. (2022). *Emotion English DistilRoBERTa-base*. In https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

Hartmann, J., Heitmann, M., Schamp, C., & Netzer, O. (2021). The Power of Brand Selfies. *Journal of Marketing Research*, *58*(6), 1159-1177. https://doi.org/10.1177/00222437211037258

Jervis, R. (2017). *How statesmen think : the psychology of international politics*. Princeton University Press.

Khrustova L.E., F. E. A., Fedorov F.Yu. . (2020). Tonality of Showing Russian Position in English Speaking Mass Media during Sanction Period. *Outlines of global transformations: politics, economics, law.* , *13*(4), 292-310. https://doi.org/10.23932/2542-0240-2020-13-4-14

Liu, Y. e. a. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/abs/1907.11692v1

Lutkevich, B. (2020). BERT language model. *Enterprise AI*. https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model

Marcus, D. (2023). *The President's Words: A Term Frequency Analysis of Putin's and Medvedev's Statements in Official Kremlin Transcripts*. dekoder.org.

Mearsheimer, J. J. (2014). Why the Ukraine Crisis Is the West's Fault: The Liberal Delusions That Provoked Putin [research-article]. *Foreign Affairs*, *93*(5), 77-89. http://ezproxy.umgc.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsjsr&AN=edsjsr.24483306&site=eds-live&scope=site

Mysiak, K. (2021). NLP Part 2| Pre-Processing Text Data Using Python - Towards Data Science. https://towardsdatascience.com/preprocessing-text-data-using-python-576206753c28

Nisch, S. (2023). Invasion of Ukraine: Frames and sentiments in Zelensky's Twitter communication. *Journal of contemporary European studies*, *ahead-of-print*(ahead-of-print), 1-15. https://doi.org/10.1080/14782804.2023.2198691

NLTK :: Natural Language Toolkit. https://www.nltk.org/

Red. (2022). *Munich Speech of Vladimir Putin*. https://www.docsonline.tv/munich-speech-of-vladimir-putin/

Roth, A. (2021, December 18). Russia issues list of demands it says must be met to lower tensions in Europe. *The Guardian*. https://www.theguardian.com/world/2021/dec/17/russia-issues-list-demands-tensions-europe-ukraine-nato

SAS Help Center: Champion Models. (2023). https://documentation.sas.com/doc/en/mdlmgrcdc/14.3/mdlmgrug/p18qonu6r3pmzgn1efc4n1fenb0v.htm

Sasse, G. (2020). Russia and Ukraine in Kremlin Rhetoric. *Russian Analytical Digest*(250), 14-15. https://doi.org/10.3929/ethz-b-000409840

Schoenbach, V. J. a. W. D. R. (2000). Data analysis and interpretation. In *Understanding the Fundamentals of Epidemiology: an evolving text*. http://www.epidemiolog.net/evolving/DataAnalysis-and-interpretation.pdf

Shahul, E. S. (2023). *Exploratory Data Analysis for Natural Language Processing: A Complete Guide to Python Tools*. Neptune.ai. https://neptune.ai/blog/exploratory-data-analysis-natural-language-processing-tools

Sharma, A. (2022). *A Beginner's Guide to Exploratory Data Analysis (EDA) on Text Data (Amazon Case Study)*. https://www.analyticsvidhya.com/blog/2020/04/beginners-guide-exploratory-data-analysis-text-data/

Shigeki, H. (2022). Historical Background of Putin's Invasion of Ukraine [Article]. *Asia-Pacific Review*, *29*(2), 19-34. https://doi.org/10.1080/13439006.2022.2105516

Silva Barbon, R., & Akabane, A. T. (2022). Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study. *Sensors (Basel, Switzerland)*, *22*(21), 8184. https://doi.org/10.3390/s22218184

Singh, H. (2023). *Complete Guide to EDA on Text Data*. Kaggle. https://www.kaggle.com/code/harshsingh2209/complete-guide-to-eda-on-text-data

Snegovaya, M. (2020). What Factors Contribute to the Aggressive Foreign Policy of Russian Leaders? *Problems of Post-Communism*, *67*(1), 93-110. https://doi.org/10.1080/10758216.2018.1554408

Tharoor, I. (2023, July 10). A Fateful Summit 15 Years Ago Hangs Over the NATO Meeting in Vilnius. *Washington Post*.

https://www.washingtonpost.com/world/2023/07/10/bucharest-2008-nato-summit-history-vilnius-putin-georgia-ukraine-membership/