War Media Analysis

Anhelina Babii (anhelina.babii@mail.utoronto.ca)
Exchange student at the University of Toronto, 27 King's College Cir,
Toronto, ON M5S Canada

Dariia Kolodiazhna (dariia.kolodiazhna@mail.utoronto.ca)
Exchange student at the University of Toronto, 27 King's College Cir,
Toronto, ON M5S Canada

Abstract

A couple of studies with similar topics have been released, however, most of them had a very shallow analysis that lacked some depth in terms of how text should be processed, especially if it is media.
Being Ukrainian, we do have the ability to understand and use Russian, therefore, it was decided that this is something we can and should do, considering everything that is currently going on.
The main goal for this project is to go deeper in Natural Language Processing, so that certain patterns and trends can be derived from all the obtained information. The results are also used to prove the hypothesis of the project. Several methods like TF-IDF and sentiment analysis were used in this research and are described in this paper.

Keywords: text analysis; sentiment analysis; TF-IDF; Ukraine; Russia; media analysis; propaganda; war.

Introduction

Since the main focus of this project is on analyzing text and finding some signs and patterns that can help define a certain sample as propagandistic or not (given that it is an article), we did our research and decided to go through three stages: general analysis that will show that the obtained data is representative, sentiment analysis that is primarly about neutral sentiment in the articles, and TF-IDF method to find trends that might be connected with certain words that might be not so easy to define as unique for the chosen period of time.
All of the methods used are a part of NLP – Natural Language Programming. NLP is the way to understand the meaning or emotion of texts written in human language without having a real person read and analyze them, thus making the process more automatic and less biased and dependent on the human who performs it.

Hypothesis

If not all, then most of Russian media, especially state-affiliated, are propagandistic, meaning the news articles posted constist of facts that are either exaggerated or non-existent.

1. Retrieving data

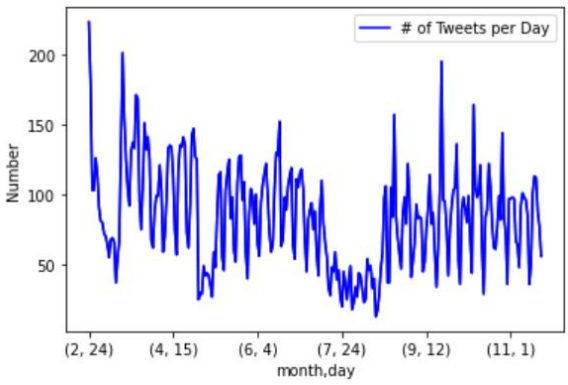
It was decided to use Twitter data as the news articles there are the perfect size – not too short and not too long. The time for the project was limited, so the normal data retrieval did not work – Twitter Developer Account requires at least 2 weeks to get.
Instead the data was parsed – parsed as from the web-page. Every single Tweet, whether it is a normal post or a reply, is still a single web-page with its unique link. This caveat was helped us use a Python library snscrape that lets the users parse the data (in this case – text) from the web-page, including lots of different social media and Twitter being one of them.
The dates were set – starting with February 24th, 2022; as well as two dictionaries were created for data parsing: the list of Twitter accounts and the list of keywords, so that not all news are downloaded, but only the needed ones – the ones that are somehow connected to the war.
The final dataset:

0	Date	User	Tweet	New Date
0	2022-09-23 16:55:00+00:00	tvrain	Включайте Дождь в 20.00. Тихон Дзядко подведет...	2022-09-23
1	2022-09-21 13:21:04+00:00	tvrain	Мы продолжаем знакомить вас с антивоенной поэз...	2022-09-21
2	2022-09-21 10:53:55+00:00	tvrain	Спецэфир на Дожде в 14:00 (мск). Будем обсужда...	2022-09-21

Table 1: Downloaded Twitter data in a dataset

2. Initial analysis (pre-analysis)

Retrieved data is not always good quality or representative, especially with big "DIY-ed" datasets.
The first couple tries downloading and creating the dataset were no good: the limits were either too big or too small, the data would not load properly. After finally getting a decent dataset, initial analysis was performed for general stats:



Plot 1: Number of Tweets posted per day

The data seems just fine – there are a few extremely suspiciously high points like 02/24 or 09/21. But both of them

are reasonable and simply connected to certain events that happened on these days: February 24th – the very start of the war, September 21st – the day of mobilization announcement in Russia.
The general activity of accounts was also checked along with their background:

User		
ukraina_ru	10000	government-owned, founded in 2014, belongs to the "Russia today" agency
bbcrussian	4753	Russian department of British broadcaster; in exile (now in Riga, Latvia)
SvobodaRadio	4466	Russian department of international broadcaster; in exile (now in Riga, Latvia)
ForbesRussia	1583	under pressure of the government
M_Simonyan	767	Russian propagandist
1prime_ru	512	65% owned by "Russia today"
navalny	309	Russian "oppositional" politician
tvrain	88	Russian "oppositional" broadcasting service in exile (now in Riga, Latvia)

Table 2: Activity and background check on accounts studied

There is an interesting correlation between the state-affiliated and oppositional accounts – both public media and personal. The number of news connected to war posted in every state-affiliated/propagandist source is more than twice bigger than in independent/oppositional (public media: ukraina_ru versus bbcrussian or SvobodaRadio; personal: M_Simonyan versus navalny)

3. Sentiment Analysis

Sentiment analysis is basically a way to understand an emotion (or lack of it) in the given message. Speaking about news, the main focus is on neutral sentiment in them – the readers should not be affected by the way the article is written. It took us a couple of preparation steps to be able to perform sentiment analysis.

3.1. Preprocess

The text under study should be preprocessed in certain way, starting with punctuation and upper-case removal.

```
# preprocessing - lower case, punctuation removal
def prep(tweet):
    tweet = tweet.lower()
    a=[w for w in tweet if w not in st.punctuation]
    return ''.join(a)
```

Code chunk 1: Code for text preprocessing

	Tweet	New Date	month	day
0	включайте дождь 2000 тихон дзядко подведет ито...	2022-09-23	9	23
1	продолжаем знакомить антивоенной поэзией напис...	2022-09-21	9	21

Table 2: Text after preprocessing

3.2. Stopwords

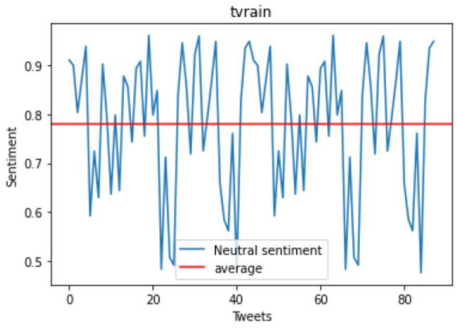
Stopwords are the words that exist in every language and do not affect the sentiment of the message. Simply put, they are not important for sentiment analysis as there is no emotion behind them. Therefore, they can be considered as noise and should be removed, too.

```
def remove_sw(tweet):
    words = nltk.word_tokenize(tweet)
    stopword=nltk.corpus.stopwords.words('russian')
    a=[w for w in words if w not in stopword]
    return ' '.join(a)
```

Code chunk 2: Removing stopwords

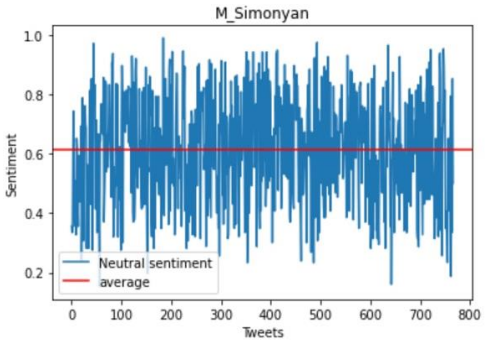
3.3. Sentiment Analysis & Results

There are dictionaries for Russian language, so the 'dostoevsky' library was used to perform the sentiment analysis on the Tweets from the dataset, grouping them by users and searching necessarily for neutral sentiment as it is the most important one in checking the news articles.



Plot 2: Neutral sentiment for the 'tvrain' media

The results here are not bad – the average of neutral sentiment is almost 80%. But this is an oppositional media source. What about government-owned ones?



Plot 3: Neutral sentiment for the 'M_Simonyan' media

As pictured above, neutral sentiment for propagandist M_Simonyan is much lower than for independent tv_rain. And what is interesting, is that it is lower because there are lots of extremely low results that are less than 20%.

4. TF-IDF Analysis

4.1 Methodology

TF-IDF (term frequency–inverse document frequency) is a measure that is widely used in machine learning that determines how relevant are given words to each document. It can be broken into two parts – term frequency(TF) and inverse document frequency (IDF). The first part basically measures how often a term shows up in each document. It is defined by the number of times a term appeared in a

document and can be adjusted for the length of the document. IDF measures how important is a term amongst the whole collection. And can be calculated as follows:

$$idf\,(t, D) = log\,(\frac{N}{count\,(d \in D{:}t \in d)})$$

IDF helps us to get rid of common words in each language like prepositions and auxiliary verbs since they appear frequently. Thus by taking inverse document frequency, we can minimize the weighting of frequent terms while making infrequent terms have a higher impact.

By multiplying these two values together we can get our final TF-IDF formula:

$$tf\,idf\,(t, d, D) = tf\,(t, d)\,.\,idf\,(t, D)$$

The higher is a TF-IDF value, the more uncommon and at the same time more frequent the term is.

4.2 Preprocess

Before using the TF-IDF method, we lemmatized our text to convert the different forms of the word to its initial form. In this way, we got rid of getting different cases of the same word in obtained results. We used the 'pymystem3' library for this. To make a corpus of documents from our dataset, we grouped our tweets by months, from February to November.

4.3 Results

month	term	tfidf	
2	русскоязычный	0.051614	russian speaking
2	завербовывать	0.051614	recruit
3	транснефть	0.086386	oil transfer
3	импорт	0.076161	Import
4	возбуждать	0.050254	open a criminal case
4	поправка	0.050254	amendments
5	националист	0.133372	nationalist
7	биолаборатория	0.056914	biolaboratory

month	term	tfidf	
8	отход	0.053976	leaving
9	отреагировать	0.066531	react
9	контрнаступление	0.106450	counteroffensive
10	аннексия	0.166960	annexation
10	бахмут	0.114059	open a criminal case
11	возбуждать	0.051407	Bahmut
11	неизвестный	0.051407	unknown
11	безнаказанность	0.046647	unpunished

Table 3: TF-IDF Analysis Results

Conclusion

There are indeed certain trends that can be used to define whether the news are propagandist. Using Russian media for this project, we discovered that neutral sentiment is much lower for state-affiliated and propagandist media, than for the independent ones; TF-IDF Analysis showed that the topics of Russian news changed nearly every month drastically, always trying to find a new thing to talk about – which is also suspicious.

In conclusion, we can state that government-owned media in Russia indeed tends to lie, exaggerate, hide facts or come up with new ones.

References

Shevtsov, A., Tzagkarakis & C. & Antonakaki, D. & Pratikakis, P & Ioannidis, S. (2022). *Twitter Dataset on the Russo-Ukrainian War*. Institute of Computer Science, Foundation for Research and Technology – Hellas, School of Electrical and Computer Engineering, Technical University of Crete.

Stecanella, B. (2019) *Understanding TF-ID: A Simple Introduction.* MonkeyLearn., https://monkeylearn.com/blog/what-is-tf-idf/

Evorov, E. (2020). *Dostoevsky - tonality analysis in Python in 5 minutes.* https://egorovegor.ru/analiz-tonalnosti-s-python-i-dostoevsky

JustAnotherActivist (2018) *snscrape. – Python library* https://github.com/JustAnotherArchivist/snscrape