JACOBS UNIVERSITY

Data Mining Project Report

Emotion Detection Through Sentiment Analysis

Instructor:Prof.Adalbert F.X Wilhelm

Assosicate Professor

Jacobs University Bremen

Submitted by: Harshitha Konduru, Jeeva Sam, Elnaz Eslami

JACOBS UNIVERSITY

Table of Contents

JACOBS UNIVERSITY

Table of Figures

JACOBS UNIVERSITY

Executive Summary

Social media data mining is a common method for determining a group's preferences, biases, and overall feelings on a certain topic. Russia invaded Ukraine on February 24, 2022. The ground battle is also the starting point for a social media war. That social media conflict drew participants from all across the world. One of the most popular microblogging platforms are Twitter and Youtube.

Economic sanctions imposed on Russia have had a tremendous impact on residents' daily lives. Consumer confidence is eroding, and poverty is rising, thus analyzing these viewpoints can help us better understand how the people felt about various events leading up to and during the conflict.

People tend to update on ground truths, how they feel about it, and who they support in a lot of tweets and youtube comments about the Ukraine and Russia battle.

One of the most popular ways of analyzing these comments is through Natural Language Processing, especially using sentiment analysis.The procedure for sentiment analysis entails gathering data, evaluating it, pre-processing it, and then sentiment identification, feature selection, sentiment classification, and removal of polarity and subjectivity. If a comment's negative score is higher than its positive score, it is classified as a positive comment; if the converse is true, it is classified as a negative tweet; otherwise, it is classified as neutral.

We looked into the issue of sentiment analysis for Ukrainian and Russian comments from Twitter and Youtube which was obtained through web scraping, built a corpus of Ukrainian and Russian comments, and labeled each text with one of three categories: positive, negative, or neutral but not only the polarity but also we have specified the strength of each comments using Vader model.

Abstract

During a crisis, social media plays a critical role in information exchange. The data collected during this crisis can reflect large-scale public sentiment and attitudes. Furthermore, these data can be utilized to investigate various campaigns used by various entities to manipulate public opinion. We present a 2022 Russo-Ukrainian conflict "Twitter" and "YouTube " dataset. We share nearly 4000 comments made during the crisis, which were analyzed using sentiment analysis and ranked according to their strength.

Introduction:

Sentiment analysis or opinion mining is the computational study of people's opinions, sentiments, emotions, appraisals, and attitudes towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. The inception and rapid growth of the field coincide with those of the social media on the Web, for example, reviews, forum discussions, blogs, microblogs, Twitter, and social networks, because for the first time in human history, we have a huge volume of opinionated data recorded in digital forms. Since the early 2000, sentiment analysis has grown to be one of the most active research areas in natural language processing (NLP). It is also widely studied in data mining, Web mining, text mining, and information retrieval. In fact, it has spread from computer science to management sciences and social sciences such as marketing, finance, political science, communications, health science, and even history, due to its importance to business and society as a whole. This proliferation is due to the fact that opinions are central to almost all human activities and are key

influencers of our behaviors. Our beliefs and perceptions of reality, and the choices we make, are, to a considerable degree, conditioned upon how others see and evaluate the world. For this reason, whenever we need to make a decision, we often seek out the opinions of others. This is not only true for individuals but also true for organizations. Nowadays, if one wants to buy a consumer product, it is no longer limited to asking one's friends and family for opinions because there are many user reviews and discussions about the product in public forums on the Web. For an organization, it may no longer be necessary to conduct surveys, opinion polls, and focus groups in order to gather public opinions because there is an abundance of such information publicly available.

The war between Ukraine and Russia has given a way for the public to express their opinion in such social Media. This opinion conveys how the public feel about the war and their support towards each countries.Many countries supporting Ukraine by introducing economic sanctions on Russia.There are a lot of tweets about the Ukraine and Russia war where people tend to update about the ground truths, what they feel about it, and who they are supporting.

In this report we analyze the sentiments of people over the Ukraine and Russian War, we will take you through the task of the Ukraine and Russia war.[1]

Data Background:

We used Web scraping technique for collecting data in real-time from Twitter & YouTube. The Streaming API is the standard method used to collect Twitter data and snscrape technique for collecting data from YouTube We use a list of keywords associated with the crisis.

Web scraping:

Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. This information is collected and then exported into a format that is more useful for the user. Web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.Web scraping is used for contact scraping, and as a component of applications used for web indexing, web mining and data mining.[2]

Web scraping techniques used for collecting the data for our analysis as below:

Google-api-python-client

The Google API Client provides a common entry point to Google Play services and manages the network connection between the user's device and each Google service. However, the newer GoogleApi interface and its implementations are easier to use and are the preferred way to access Play services APIs.[3]

Snscrape

Snscrape is a Python library that can be used to scrape tweets through Twitter's API without any restrictions or request limits. Moreover, you don't even need a Twitter developer account to scrape tweets when you use snscrape[4].

Data Pre-processing:

Data preprocessing is a phase in data mining and data analysis process that turns raw data into a format that computers and machine learning can understand and evaluate. Text, photos, video, and other forms of real-world raw data are untidy. It may not only contain errors and inconsistencies, but it is frequently incomplete and unstructured.

As we are dealing with text in this project it is essential to clean the data before applying the machine learning algorithm on it. Text pre-processing in NLP includes Tokenization, lowercasing, stop word removal, stemming which is used to reduce dimensionality of the text [5].The python library used for the same in this project is Natural Language Toolkit (NLTK).

NLTK is a popular Python programming language for working with human language data. It includes a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum, as well as easy-to-use interfaces to over 50 corpora and lexical resources like WordNet [6].

The dataset was in CSV file , as it is raw data it had many other languages, emoji and punctuations. It also contained comments which were not related to Ukraine or the Russian war which were also included while cleaning the text.

Following steps were taken to clean the text.

1. Initially the text was converted into lowercase to bring all the words to one form.

2. Removed the punctuations for e.g.!”#$%&'()*+,-./:;<=>?@[\]^_`{|}

3. Apart from those punctuations other unwanted terms like 'https?://, www were also removed.

4. Unnecessary spaces between the word and among the sentences were removed.

5. Stop words were removed using the stop word package in NLTK.

Stop words: Stop words are the words which are commonly used in a language. The words such as ‘is’, ‘a’, ‘the’ are some of the examples of stop words.

6. Stemming was performed using NLTK. SnowballStemmer.

Stemming: It is the process of bringing down a word to its infinitive form for e.g: The word ‘eating’ after stemming is reduced to eat.

Martin Porter is the creator of Snowball Stemmer. The algorithm used here is known as "English Stemmer" or "Porter2 Stemmer," and it is more accurate. In terms of logic and speed, it is a small advance over the original Porter Stemmer [7].

Sentiment Analysis:

Sentiment analysis is a text analysis method that detects polarity (e.g. a *positive* or *negative* opinion) within the text, whether a whole document, paragraph, sentence,

or clause[8].Sentiment analysis aims to measure the attitude, sentiments, evaluations, attitudes, and emotions of a speaker/writer based on the computational treatment of subjectivity in a text[8].

A tweet may contain multiple sentiments all at once, which might get tricky for traditional Sentiment Analysis models.So we have used VADER as the sentiment Analysis Model which is sensitive to both polarity(positive/negative) and intensity(strength) of the emotion.

VADER

(Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. [9]

It is used for sentiment analysis of text which has both the polarities i.e. positive/negative. VADER is used to quantify how much positive or negative emotion the text has and also the intensity of emotion.

Advantages:

- It does not require any training data.
- It can very well understand the sentiment of a text containing emoticons, slangs, conjunctions, capital words, punctuations and much more.
- It works excellent on social media text.
- VADER can work with multiple domains.[9]

Using Vader to analyze the tweets

It is available in the NLTK package that can be applied directly to unlabeled text data.VADER's SentimentIntensityAnalyzer() takes in a string and returns a dictionary of scores in each of four categories:

- negative
- neutral
- positive
- compound *(computed by normalizing the scores above)*

```
nltk.download('vader_lexicon')
import ast
sentiments = SentimentIntensityAnalyzer()
data["Positive"] = [sentiments.polarity_scores(i)["pos"] for i in data["Comment"]]
data["Negative"] = [sentiments.polarity_scores(i)["neg"] for i in data["Comment"]]
data["Neutral"] = [sentiments.polarity_scores(i)["neu"] for i in data["Comment"]]
data["compound"] = [sentiments.polarity_scores(i)["compound"] for i in data["Comment"]]
data = data[["Comment", "Positive", "Negative", "Neutral", "compound"]]
data[:50]
```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     /Users/harshithakonduru/nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
```

	Comment	Positive	Negative	Neutral	compound
0		0.000	0.000	0.000	0.0000
1	cafinus thanks post graphic horrific gut wren...	0.158	0.240	0.601	-0.3612
2	horrific ending life russian soldiers lay dead...	0.000	0.474	0.526	-0.9274
3	dead zs istandwithukraine combatfootage	0.000	0.589	0.411	-0.6486
4	world watching coverage i've seen today mass c...	0.122	0.338	0.541	-0.7003
5	pennyflipper meidastouch istandwithukraine n...	0.000	0.000	1.000	0.0000
6	grandson wants make ukrainian foods recipes sa...	0.198	0.000	0.802	0.6369
7	chefjoseandres wckitchen thankyou much nistand...	0.000	0.000	1.000	0.0000
8	im going deleting artstation due fact theyve c...	0.195	0.000	0.805	0.4404
9	fresh batch drones landed chaps istandwithukr...	0.315	0.000	0.685	0.3182
10	used russian military vehicle ukrainians blew ...	0.000	0.000	1.000	0.0000
11	dead z lays mans land two destroyed russian ve...	0.000	0.385	0.615	-0.8176
12	don't look away shit happening speak istandwit...	0.000	0.375	0.625	-0.5574

Figure No.1: Snapshot of the results after applying the Vader Model

Here we used compound values to categorize each tweet into “Happy”, “Sad”, “Neutral”.

```
In [14]: for index, row in data.iterrows():
             if row["compound"] >= 0.05:
                 print(index, row["Comment"],"\n----->HAPPY")
             elif row["compound"] <= -0.05:
                 print(index,row["Comment"], "----->SAD")
             else:
                 print(index,row["Comment"],"----->NEUTRAL")
```

```
0   ----->NEUTRAL
1 cafinus thanks post  graphic horrific gut wrenching msgs ukraine today 💔 putinwarcrimes slavaukraini  istandwithuk
raine ----->SAD
2 horrific ending life russian soldiers lay dead wastelands ukraine never return home buried  istandwithukraine war r
ip genocideofukrainians  ----->SAD
3 dead zs  istandwithukraine combatfootage  ----->SAD
4 world watching coverage i’ve seen today mass civilians killed heartbreaking 💔  need nato help need helpnistandwith
ukraine🇺🇦  ----->SAD
5 pennyflipper meidastouch istandwithukraine n🇺🇦💙💛💔😢  ----->NEUTRAL
6 grandson wants make ukrainian foods recipes saw cabbage rolls golubtsi anyone recommend authentic recipe made cabba
ge rolls wants learn culture ukrainianrecipes istandwithukraine help
----->HAPPY
7 chefjoseandres wckitchen thankyou much nistandwithukraine nchefsforukrainenartistsforukraine  ----->NEUTRAL
8 im going deleting artstation due fact theyve censoring ukranian artists supporting russia istandwithukraine
----->HAPPY
9 fresh batch drones landed chaps  istandwithukraine
----->HAPPY
10 used russian military vehicle ukrainians blew smithereensnnthe crew vaporised impactnnistandwithukraine  combatfoo
tage  ----->NEUTRAL
```

Figure No.2: Snapshot after categorizing tweets into one of the annotated labels

Visualization - Word Cloud

A word cloud (also known as a tag cloud) is a visual representation of words. Cloud creators are used to highlight popular words and phrases based on frequency and relevance. They provide us with quick and simple visual insights that can lead to more in-depth analyses.[10]

Visualization - Twitter

```
positive =' '.join([i for i in data_positive['Comment'][data_positive['Positive'] > data_positive["Negative"]]])
stopwords = set(STOPWORDS)
wordcloud = WordCloud(stopwords=stopwords, background_color="white").generate(positive)
plt.figure( figsize=(15,10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

Figure No.3: Visual Representation of positive words in the Twitter Dataset

```
negative =' '.join([i for i in data_negative['Comment'][data_negative['Negative'] > data_negative["Positive"]]])
stopwords = set(STOPWORDS)
wordcloud = WordCloud(stopwords=stopwords, background_color="white").generate(negative)
wordcloud
plt.figure( figsize=(15,10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

Figure No.4: Visual Representation of negative words in the Twitter Dataset

Visualization - YouTube

```
positive =' '.join([i for i in data_positive['Comment'][data_positive['Positive'] > data_positive["Negative"]]])
stopwords = set(STOPWORDS)
wordcloud = WordCloud(stopwords=stopwords, background_color="white").generate(positive)
plt.figure( figsize=(15,10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

Figure No.5: Visual Representation of positive words in the YouTube Dataset

```
negative =' '.join([i for i in data_negative['Comment'][data_negative['Negative'] > data_negative["Positive"]]])
stopwords = set(STOPWORDS)
wordcloud = WordCloud(stopwords=stopwords, background_color="white").generate(negative)
wordcloud
plt.figure( figsize=(15,10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

Figure No.6: Visual Representation of negative words in the YouTube Dataset

Conclusion:

The increased tension between Ukraine-Russia results in the Russian invasion of Ukraine. Social Media plays a vital role in sharing information about the war and opinions from all over the world. Sentiment analysis helps to understand reactions and perspectives. The aim was to analyze the sentiments before and during the escalation. This report presentes the sentiment Analysis about the tweets on the Russian and Ukrainian war and discovers the polarity of public opinion of it. The tweets were annotated into three categories - positive, negative and neutral. We have used the VADER Model to carry out sentiment analysis. The compound value from the model results was used to categorize each tweet into one of the annotated labels. And Word cloud was used to visualize the most frequently used words in the positive and negative categories of tweets.

Future Work:

We want to further improve our analysis by carrying two aspects. Firstly we would like to include other languages for our analysis and categorize them accordingly. Second, we would like to carry more detailed analysis of the most ambiguous texts to detect the common cause of their ambiguity and also to introduce one more annotation category: ‘ambiguous’ for the texts the annotators are not certain about.

Appendix:

Complete code implementation for the above project is available on Github, which can be accessed through the link below:

https://github.com/HarshithaKonduru/Data-Mining-Report

References:

[1] Sentiment analysis
Available:
https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2
Accessed on 31 May 2022

[2] Data scraping
Available:
https://en.wikipedia.org/wiki/Data_scraping
Accessed on 31 May 2022

[3] Google API Client Library for Python Docs
Available:
https://googleapis.github.io/google-api-python-client/docs/
Accessed on 31 May 2022

[4] Snscrape
Available:https://pythonlang.dev/repo/justanotherarchivist-snscrape/
Accessed on 31 May 2022

[5] Text Preprocessing With NLTK (Toward Data Science).
Available:
https://towardsdatascience.com/nlp-preprocessing-with-nltk-3c04ee00edc0
Accessed on 31 May 2022

[6] Natural Language Tool Kit (Documentation)
Available: https://www.nltk.org/
Accessed on 31 May 2022

[7] Beginner's Guide to Stemming in Python NLTK (machinelearningknowledge.ai)
Available:
https://machinelearningknowledge.ai/beginners-guide-to-stemming-in-python-nltk/
Accessed on 31 May 2022

[8] SENTIMENTAL ANALYSIS USING VADER
Available:
https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664
Accessed on 01 June 2022

[9] Sentiment Analysis Made Easy Using VADER
Available:
Sentiment Analysis Made Easy Using VADER
Accessed on 01 June 2022

[10] What Is a Word Cloud?
Available:
https://monkeylearn.com/word-cloud/
Accessed on 01 June 2022

[11] Bobicev, Victoria & Kanishcheva, Olga & Cherednichenko, Olga. (2017). Sentiment analysis in the Ukrainian and Russian news.
Available:
https://www.researchgate.net/publication/321356659
Accessed on 10 May 2022

[12] Ehsan-Ul Haq, Gareth Tyson, Lik-Hang Lee, Tristan Braud, Pan Hui(2022) Twitter Dataset for 2022 Russo-Ukrainian Crisis.
Available:
Cornell University - Twitter Dataset for 2022 Russo-Ukrainian Crisis
Accessed on 10 May 2022