**Distributed Data Processing and Sentiment Analysis of a Twitter Dataset**

Author: Jefferson William Teixeira

e-mail: [sba22201@student.cct.ie](mailto:sba22201@student.cct.ie)

Student ID: sba22201

Contents

[1 Introduction 3](#_Toc135925444)

[2 Materials and Methods 3](#_Toc135925445)

[3 Data Preparation and Visualization 3](#_Toc135925446)

[3.1 Text Processing for Sentiment Analysis 4](#_Toc135925447)

[3.2 Data Preparation for Time-Series Forecasting 5](#_Toc135925448)

[4 Data Storage Strategies 6](#_Toc135925449)

[5 Sentiment Time-Series Forecasting 8](#_Toc135925450)

[6 Conclusions 8](#_Toc135925451)

[7 References 8](#_Toc135925452)

# Introduction

In this project, distributed data processing and machine learning (ML) were applied to a large Twitter dataset that contains comments about the war in Ukraine. The dataset was collected from a Kaggle [repository](https://www.kaggle.com/code/ssaisuryateja/eda-and-sentiment-analysis/input) and contains about 364,875 tweets on this topic. For this project, the dataset was stored using distributed environment and a NoSQL database as well as processed for sentiment analysis. The tweets were analysed using time-series forecasting to predict the overall Twitter user’s sentiment in different time intervals as per the experiments presented in this report. Moreover, a benchmarking tool called Yahoo Cloud Serving Benchmark (YCSB) was used to evaluate and compare the performance of MySQL and MongoDB databases.

This report is organized as follows: Section 2 describes the tools used to complete this project. In Section 3, the dataset and the data preparation steps for the sentiment analysis and time-series forecasting are described. In Section 4, the distributed environment and benchmarking results are discussed, while Section 5 covers the sentiment analysis, time series forecasting experiments and the dynamic dashboard implemented. Finally, in Section 6, the conclusions are summarized.

# Materials and Methods

The source code and files used in this project were hosted on GitHub under a public organization called [CCT-MastersDA](https://github.com/CCT-MastersDA). The project’s repository is called [cct-sem2-ca2](https://github.com/CCT-MastersDA/cct-ca2), which can be accessed with the following command in any terminal: *git clone* [*https://github.com/CCT-MastersDA/cct-sem2-ca2.git*](https://github.com/CCT-MastersDA/cct-sem2-ca2.git). The source code is organized under the *jupyter* folder in the root directory which also contains a report folder for the documentation.

For the sentiment analysis and time-series forecasting, it was implemented a Jupyter file called ca2-jupyter.ipynb that uses the content of the following folders: datasets, which contains the csv files used in this project, the images folder, where all the generated images are stored, and the modules folder, which basically contains auxiliary classes for the text processing and data manipulation. The reason to separate these modules from the Jupyter notebooks was to keep the code organised and to follow the best programming practices with regards to reuse and code modularization.

The distributed environment used to store the datasets was Hadoop and the technology implemented to access and manipulate the data was PySpark, a Python API that enables large-scale data processing in Python. The dataset was also loaded into MongoDB, a NoSQL database. Regarding the benchmarking approach, MongoDB and MySQL databases were compared using YCSB, an open-source specification and software package for benchmarking NoSQL databases’ relative performance. For this purpose, it was implemented three bash scripts, which can be found under the *jupyter/benchmark* folder. The benchmarking-script.sh contains the code to run the experiments for the workloads in Mongo and MySQL. The parse-results.sh is responsible for collecting the metrics from the results files for each execution. Finally, the exec.sh script is the trigger script that receives as input the desired number of executions.

Regarding the setup, for the distributed environment and data storage experiments, a Linux VM was prepared with the following main required software: MongoDB, MySQL, YCSB, Hadoop, PySpark and Anaconda. The sentiment analysis and time-series forecasting were implemented in Windows and were executed in Anaconda environment.

# Data Preparation and Visualization

The dataset obtained has about 215mb and contains 364,875 tweets about the conflict in Ukraine that were published in just two days, between 01/04/22 and 02/04/22. Due to time and performance constraints, it was not feasible to collect data about this topic for the whole year.

This way, to overcome this problem and perform the analysis proposed in this project, the date column was created artificially, so that it covers the period between 01/01/2022 and 31/12/2022. Thus, each tweet was assigned a random timestamp, so that there are about 999 tweets a day in the processed dataset.

Since the focus of the analysis is the sentiment of the users over time, this approach does not affect that result, however, it makes the analysis less accurate in terms of the timing as the timestamps were modified. Therefore, the results presented in this report are meant to be taken as an exercise only.

The data preparation steps discussed in this section were implemented in the section 4 of the Jupyter notebook accompanying this report.

## Text Processing for Sentiment Analysis

The tweets were processed using Text Processing algorithms for sentiment analysis. This way, four versions of the tweets were generated from the raw text: (1) cleaned with stop words, (2) cleaned without stop words, (3) *lemmatized,* and (4) *stemmerized* tweets. Stop words do not add much information to the text, so their frequency could bias the models, therefore, they are usually removed from the dataset. *Lemmatization* and *Porter Stemmer* are common ways to extract the core meaning from the words, this way these techniques were also applied to the datasets (Müller and Guido, 2016). For each version of the cleaned tweets, it was calculated the sentiment using Python TextBlob library, where the possible polarities are positive, neutral, and negative. The count of sentiment for each processed version of the tweets is shown in Table 1, where the column names are explained as follows:

* tweet\_raw, sent\_raw: Original tweets and their sentiment.
* tweet\_str, sent\_str: Tweets after the removal of special characters and their sentiment.
* tweet\_clr, sent\_clr: Tweets after the removal of special characters and stop words and their sentiment.
* tweet\_st, sent\_st: Tweets after the removal of special characters, stop words and application of Porter Stemmer followed by their sentiment.
* tweet\_lm, sent\_lm: Tweets after the removal of special characters, stop words and application of Lemmatizer followed by their sentiment.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ***sent\_str*** | ***sent\_clr*** | ***sent\_st*** | ***sent\_lm*** |
| ***negative*** | 58,350 | 59,894 | 48,908 | 63,110 |
| ***neutral*** | 187,658 | 203,118 | 230,836 | 200,654 |
| ***positive*** | 118,625 | 101,621 | 84,889 | 100,869 |

Table 1- Sentiment count per tweet version

Overall, the different text processing techniques applied to the raw tweets did not change the big picture, which shows more neutral comments than negatives or positives. It is important to notice that this sentiment analysis alone is not able to explain what exactly Twitter users were neutral about in this dataset, as it is not possible to guarantee that all tweets are making judgements about the war, for example. While this dataset only contains tweets about the conflict, several subjects can be talked about it in social media, which makes it hard to extract more precise information about the users’ feelings or opinions towards such a complex subject based on this analysis. This can also be verified by the word cloud that was generated for each version of the processed tweets, as shown in the figures below, where the word clouds show different subjects on the topic, such as the role of media, misinformation, among others.

|  |  |
| --- | --- |
| Figure 1- Word cloud for tweet\_str | Figure 2- Word cloud for tweet\_clr |
| Figure 3- Word cloud for tweet\_st | Figure 4- Word cloud for tweet\_lm |

From the technical perspective, the different text processing approaches have an impact on the result of the lexical sentiment extraction, as the numbers show some tweets being classified differently, but in this dataset the overall feeling was kept consistent based on the sentiment counts.

## Data Preparation for Time-Series Forecasting

The final dataset contains the processed tweets and the synthetic timestamp. At this stage, the dataset was also verified for null and missing values, before being configured for a time-series. For the experiments in this report, only the *sent\_clr* sentiment values were used, which means that only the sentiment associated with the tweets after the removal of stop words and special characters were considered for the time-series forecasting. The same experiment could be replicated for the other versions, but since their polarity does not change overall, as explained in the previous section, only one version was evaluated.

As part of this step, the approach taken was to calculate the average polarity of the tweets per day, so that the time-series had a daily frequency. This strategy was adopted because of the purpose of the analysis, which is to forecast the sentiment of users in the next 7, 30 and 90 days, hence the data had to be aggregated. The generated time-series is shown in Figure 5, where the highlighted last 30 days of the year were used as predictors for the forecaster, which is explained in more detail in Section 5.

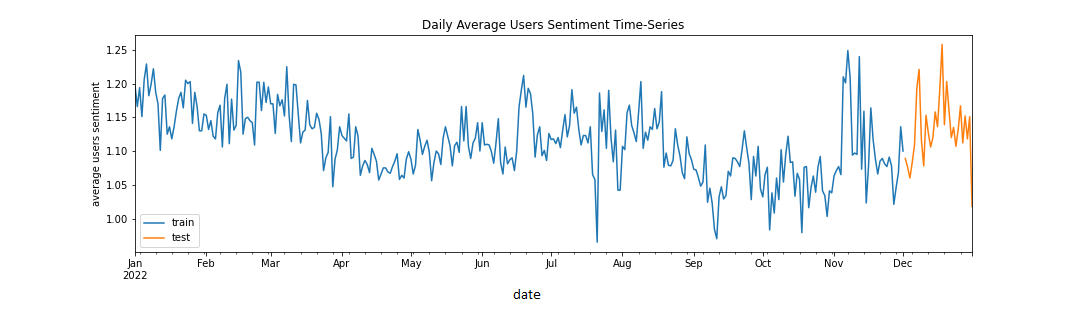


Figure 5- Average user's sentiment in 2022 (Daily)

# Data Storage Strategies

As part of the big data analysis, in this project, the dataset was stored in a Hadoop distributed environment and then transferred to MongoDB to demonstrate the application of a NoSQL database. Moreover, YCSB benchmarking tool was used to compare the performance of MySQL and MongoDB. This experiment was executed in the VM, and the steps were implemented in section 3 of the Jupyter notebook accompanying this report. The steps executed are explained bellow.

1. **Hadoop Setup**
2. In this step, the original dataset was uploaded to a Hadoop environment configured in the VM. The following screenshot shows how the distributed filesystem was initialized.

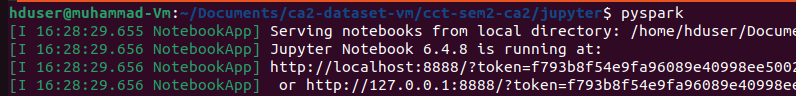
A screenshot of a computer program

Description automatically generated with medium confidence

1. Using the Hadoop command line interface, the datasets were uploaded to the filesystem. The screenshot below shows the files under user1.



1. In section 3.1 of the Jupyter notebook, it is demonstrated how the dataset was manipulated using PySpark API, where the dataset was loaded into main memory to be transferred to MongoDB as part of the next experiment, and some queries were executed to demonstrate how data is handled from Hadoop filesystem. The screenshot below shows PySpark being initialized in the VM.



1. **MongoDB**
2. In section 3.2 of the Jupyter notebook, the code to transfer the dataset from Hadoop to MongoDB was implemented. For this demonstration, Mongo database was first initialized as shown in the screenshot below.

A screen shot of a computer screen

Description automatically generated with low confidence

1. The screenshot below shows the databases and collections stored initially in the local environment. For this experiment, the database name is cct-ca2 and the collection where the tweets are stored is called twitter.

A screenshot of a computer

Description automatically generated

1. The database was removed so the experiment can be restarted without issues.



1. As part of the commands implemented in section 3.2 of the Jupyter notebook, a sample of the original dataset that contains 1000 rows were selected from the data frame collected from Hadoop. The reason for this was performance issues, as more rows would take too much time to process. For this operation, the MongoClient library was used, which received a dictionary from the data frame obtained using PySpark. As a result, a collection called twitter was recreated in the database as shown in the screenshots below.



A computer screen shot of white text

Description automatically generated with low confidence

1. **YCSB Benchmarking**

In this step, Mongo and MySQL databases were compared using a benchmarking tool called YCSB. For this experiment, three workloads were prepared with the following basic configuration:

|  |  |  |  |
| --- | --- | --- | --- |
| **Workload** | **#Operations** | **Read Proportion** | **Update Proportion** |
| workloada | 1000 | 50% | 50% |
| workloadb | 1000 | 95% | 5% |
| workloadc | 1000 | 100% | 0% |

The experiment was performed in the VM with MongoDB and MySQL up and running. The experiment was executed 5 times using the bash scripts implemented to facilitate and automate the process. The reason to run the tests several times was to avoid outliers in the metrics. In fact, for a more elaborated benchmarking strategy, factors like cache and other processes running in the machine at the same time should be considered. However, for the sake of simplicity, the approach taken in this project was calculating the average metrics of the 5 executions.

In the section 3.3 of the Jupyter notebook accompanying this report, it is shown how the results csv file was processed to get the average metrics of the executions, which is illustrated in.

# Sentiment Time-Series Forecasting

# Conclusions

# References

Müller, A.C. and Guido, S. (2016) *Introduction to machine learning with Python: a guide for data scientists*. First edition. Sebastopol, CA: O’Reilly Media, Inc.