Mythbusters :
a Fake News detection platform

Big Data Analytics - MS1

Amir Ali, Jacek Czupyt, Javier Serrano, Jean-Baptiste Soubaras

October 2022

1 Description of the subject

1.1 Problem statement

We are sure you have heard about "Fake news" plenty of times recently. This is a topic which has gained an exponential fame with the enormous growth of the Internet and specially, social networks over the last few years.

This outburst of the social media flooded the internet with information, making it a competitor to traditional media. However, contrary to this latter, no accreditation nor proofreading process is required to post on social media. This means that everyone can post anything wanted, what makes it is very easy to post fake news. At the end, it results in the spread of misinformation.

1.2 Introduction

The purpose of our project is to treat posts written on different social media in order to check whether a statement on the actuality is false or not. This will imply to be able to stream data from social media sources, train a machine learning based model involving natural language processing methods to classify the different posts as "reliable" or "unreliable".

1.3 Significance of the study

This project could be useful in a context of fighting against misinformation, and would allow to automate the fact-checking process which is today performed manually by journalists in order to inform social media users more effectively of the veracity of the information that can be stumbled upon while browsing the internet.

1.4 Non-functional objectives

The non functional objectives of this project will be :

- efficiency, deploying a reliable classification model;

- maintainability, assuring the continuous functioning of the platform;
- ergonomy, creating an efficient front-end able to deal with a large range of queries;
- security, ensuring that the platform can't be deflected by someone willing to decredibilize real information or propagate fake news;
- ethics, ensuring that the errors made by the platform will not falsely harm the reputation of a journalist or politician or wrongly support and spread manipulated information.

2 Data sources

2.1 Streaming data

To process the news in real time, we will have to stream post sent on popular social media : Twitter, Reddit, and possibly Facebook. We will access the data using the APIs related to these networks.

The Twitter API allows us to stream or search for recent or historic tweets filtered by things such as keywords or hashtags. This should allow us to easily extract up-to-date news and events. It also allows us to receive up to 500k per month on its free tier, which should be more then sufficient for our purposes.

The Reddit API allows us to search for recent threads on a given subreddit, as well as stream updates on a particular thread. We plan to create a client which queries a single, or multiple news-related subreddits, and possibly converts them into a stream. The API allows for up to 60 calls/minute.

The Facebook API appears to give the ability to read the feed of selected pages, letting us query various news pages for new posts. However preliminary research suggests that this may require special permissions from said pages that we may not be able to get access to.

2.2 Batch data

To ensure the classification results of our machine learning algorithms, we will need to train them on already classified data sets giving the reliability of different social media statements or articles. Some fact checking organizations released such data sets containing short statements or title of articles manually labeled as real or fake. These data sets will feed the batch layer of our architecture (detailed in the next section) and will be used as training data sets.

We will use three of these data sets:

- The *Fakeddit* dataset
- The *LIAR* dataset
- The ISOT dataset

The *Fakeddit* data set is detailed in the article [3]. It contains a classification of statements posted on the social media Reddit in 6 classes (True, Satire, Misleading, Manipulated, False Connection and Imposter). The features used for this classification are the content of the different posts (text and image) and the comments. The data set (without taking into account the images that we won't use) is about 900 MB with the comments, and 200 MB with only the posts.

The *LIAR* data set consists of short statements made on the media (social or traditional) manually classified according to their degree of reliability (in a total of 6 classes : true, false, half-true, pants-fire, barely-true, mostly-true). The data set comes from the fact checking American organization *Politifact* and was introduced in [4]. It contains 12.8K of statements, which make about 3 MB of data.

The *ISOT* is a data set introduced in [1] and in [2], classifying articles from the press as true or fake. It contains about 45 K articles and is more a than 100 MB.

3 System Architecture

Figure 1: Shows our proposed architecture. It describes layers of the system, and technologies that we will used in each layer. The proposed architecture is based on the Lambda which is divided into batch, serving and speed layers.

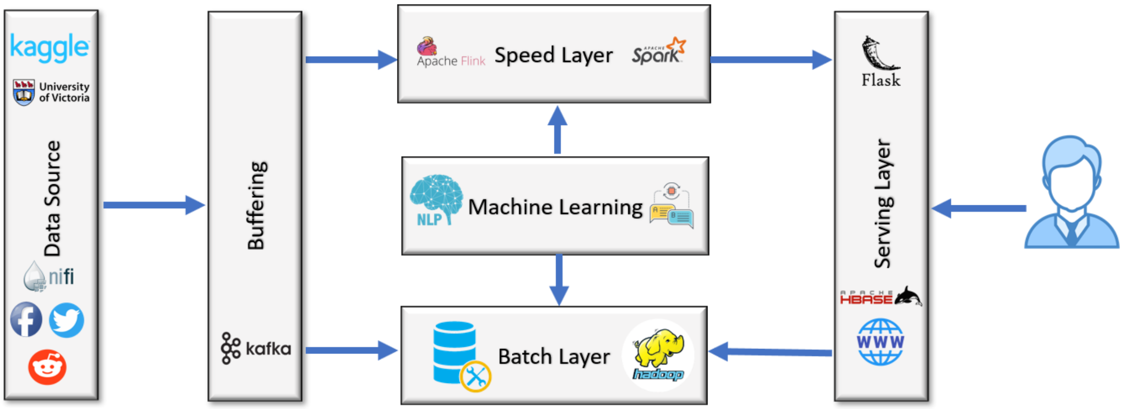

Figure 1: Our adoption of the lambda architecture

- Batch Layer: In this layer, we store all the raw data which come from Data Source and perform batch processing on the data. We plan to use Apache Hadoop for storage and Apache Spark for processing.
- Machine Learning: Fake news detection needs a text classification approach. To predict whether the given input is fake news or not we will use some machine learning classification methods.
- Speed Layer: This layer will analyze the real time data. We plan to use Apache Spark or Flink for the processing.
- Serving Layer: The serving layer will contain batch views and text data of fake news. We plan to use Apache Hbase for storage, Flask for the back-end server, and possibly some front-end framework such as React.

4 Required Technologies

These are the technologies that we will use in order to build a robust fake news detection system:

- Apache Hadoop
- Apache Spark
- Apache NiFi
- Apache Flink (optional)
- Scikit-Learn
- Natural Language Toolkit
- Flask Web Framework

5 Work distribution

Our group is composed of four students:

- Amir Ali : NLP, data processing
- Jacek Czupyt : stream data sources, speed layer
- Javier Serrano : data processing, serving layer
- Jean-Baptiste Soubaras : **team manager**, batch data sources, batch layer

References

[1] Saad S Ahmed H, Traore I. “ detection of online fake news using ngram analysis and machine learning techniques. 2017.

[2] Saad S Ahmed H, Traore I. Detecting opinion spams and fake news using text classification. Journal of Security and Privacy, 2018.

[3] Kai Nakamura, Sharon Levy, and William Yang Wang. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. arXiv preprint arXiv:1911.03854, 2019.

[4] William Yang Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648, 2017.