Sentiment Analysis on Twitter : Effects in a Social Network

Mohamed Ben Hamdoune and Yannis Tannier

Our purpose is to build a powerful platform for real-time data analysis of tweets on twitter trends. We also want to analyse all the tweets of 2017 based on a downloaded sample of data (average of 6To). All this data analysis will be accessible via a web interface that will be developed. We want to build a powerful system of sentiments analysis by making a database structure of tweets which is relevant about impacts and effects. The system should provide a faster way to execute Machine Learning methodologies behind data extracted from Twitter. Analysis news actuality by getting an analysis on actual trends with real stream data by building an efficient web interface to get results easily and build a system without false accounts and keep a control on data continuously.

[Cloud Computing] [Social Media Analytics]

Jason Scott Sadofsky acknowledges a Jason Scott, is an American archivist, historian of technology, and filmmaker. Archive Team is a group dedicated to preserving digital history that was founded by Jason Scott in 2009. Data was collected from the website of The Archive Team.

# Introduction

The main subject is Sentiment Analysis on Twitter, a microblogging platform where people can easily share their thought on anything and their habits too. We have a lot of publications on sentiment analysis but not so much research about impacts and their effect on society. The maximum characters are 140 which can be a good thing for the process of analysis because it will make it faster in a way to perform on small messages but in the other hand we should pay attention on accuracy of results. Event it’s an enormously continuous stream of data, Twitter is a good extra sentiment though an online community. Therefore, how to optimize all those streaming data and build a web interface for users who want to get data. Our project will use many methodologies from Machine Learning like unsupervised methods to make a classification of sentiments, and supervised method to predicate psychological profile. Finally, one big step will be and efficient system about control of massive data incoming, a check on false account and spam messages that will destroy our results for example.

Jeffrey Zeldman “The best way to engage honestly with the marketplace via Twitter is to never use the words “engage,” “honesty,” or “marketplace.”

In this study, we introduce readers to the problems of Data Processing and Cloud Computation, which have been rapidly developing over the last decade.

# Related Work

To begin, ye can refer to this article because we can relate that it is a point of start, it gives us a theoretical review on the development of Sentiment Analysis. The interested reader can also refer to previous surveys in the area, like [cite] and [cite], that helped us on machine learning techniques. As an active research field that has emerged for a long time now, sentiment analysis is now been greatly implement but it is also with a cost (for example, IBM Watson Tone Analyzer). Sentiment analysis is a discipline that extracts people’s feelings, opinions, thoughts and behaviors from user’s text data using Natural Language Processing (NLP) methods. For methodologies on preprocessing, and feature generation, we based our work on process. Then we can also cite for a reason, they removed numbers from theirs tweet thinking that in general, numbers are of no use when measuring sentiment and are removed from tweets to refine the tweet content but we wanted to keep them thinking the contrary. . We had over 800 million tweets in English and all that data to analyze where possible just because of the MapReduce Model in majority.

# Methodology

Machine learning approach relies on statistical algorithms in one hand to solve the Sentiment Analysis as a regular text classification problem that makes use of syntactic and/or linguistic features. Text Classification Problem Definition: We have a set of training records D where each record is Sentiment analysis algorithms and applications: A survey labeled to a class. The classification model was written in Python and we used TextBlop for sentiment analysis. Then for a given instance which mean an input of tweet coming via a StreamListener with Kinesis or in the Archive of unknown class, the model is used to predict in what label it should be. The hard classification problem about tweets is that we can’t really be sure about results only if tweet were confirmed by an human being or every user of Tweeter will have to add information like sentiment and emotion for his tweet before sending it .

## Pre-processing

In this paper we introduce two new resources for pre-processing twitter data, as you can see in the repository, an emoticon dictionary is available in Python and we can remove them for any tweets. It is really powerful because even it is relevant, we can’t take them for computing them after due to a problem of time. Positive, Negative, and Neutral are the three label for Sentiment Analysis and Joy, Fear, Anger, Surprise and Sadness are the fifth labels for Emotion Analysis. We pre-process all the tweets as follows:

1) We remove all the emoticons

b) We remove all URLs with a regular expression

c) replace targets (e.g. “@John”) by also removing them

d) removing “RT” for must of the tweet, because it doesn’t mean something for the application.

Tweets in a general way contain a lot of opinions about which are expressed in different ways by users .The twitters dataset used in this work is already labeled. The raw data is pre-processed. It deals with the preparation that removes the repeated words and punctuations and improves the efficiency the data. Forward other process are do like converting upper case to lower case: In case we are using case sensitive analysis. User names and URLs are not important from the perspective of future processing, hence their presence is futile. All usernames and URLs are removed for increase the real result.

## Feature Generation

The data for this task consists of tweets across various domains, classified into four emotions: joy, sadness, anger and fear and surprise. The Language problem: Powerful systems with good results are often for English sentences. Accordingly to the main source of data which is the Archive of tweets, most of data are in English.

# Cloud Computing

Cloud computing was used for the project with Aamazon Web Service (AWS) for helping developer making machine learning built systems quickly.

## Spark Framework

Here

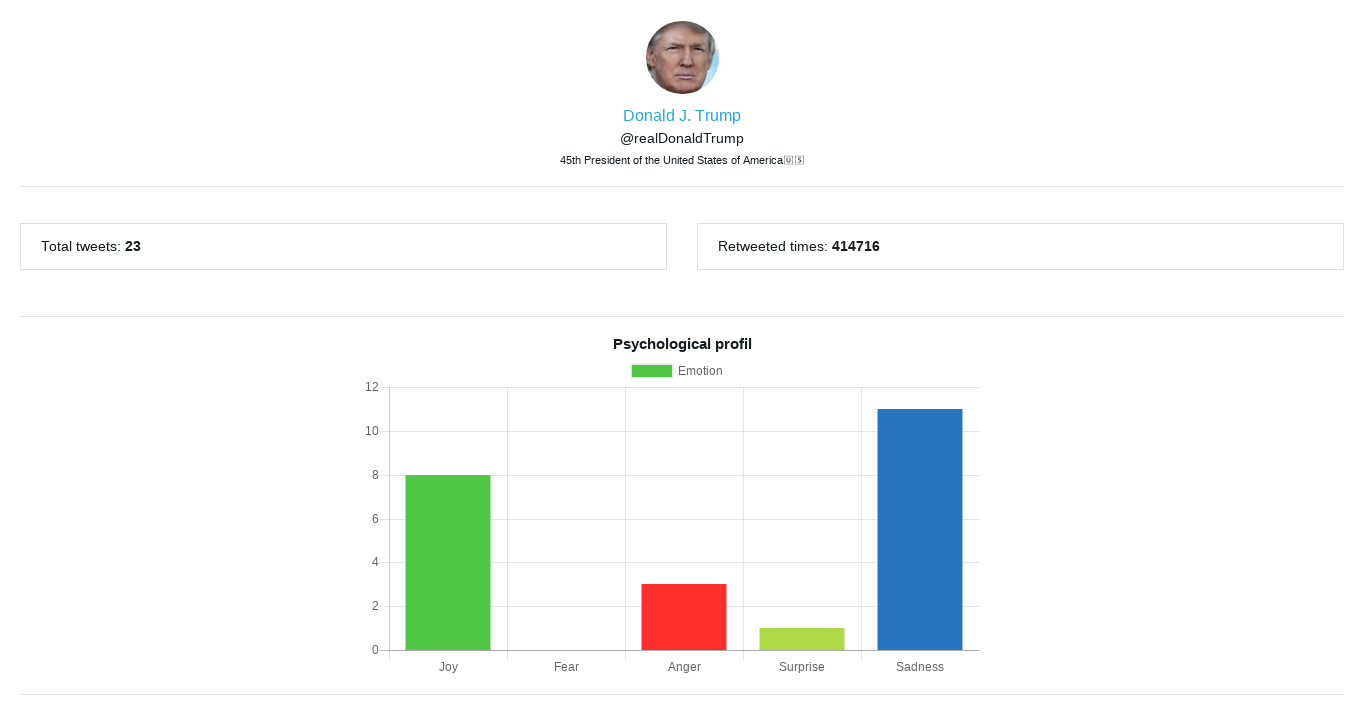
## Conception of the Database for the Web Interface

The World Wide Web (Web, for short), is a distributed information system based on hypertext. Web interfaces to databases have become very important. After outlining several reasons for interfacing databases with the Web, we provide an overview of Web Technology. We then applies techniques for building Web interfaces to databases.

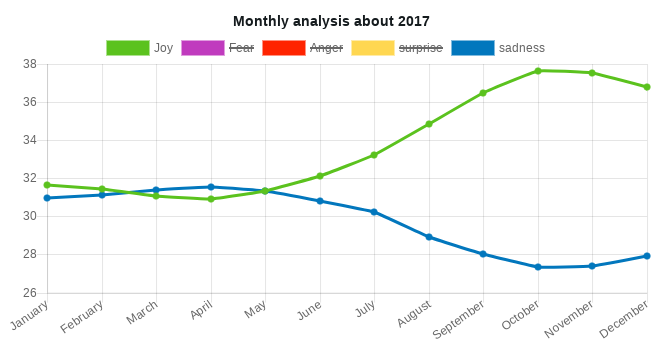
# Implementation

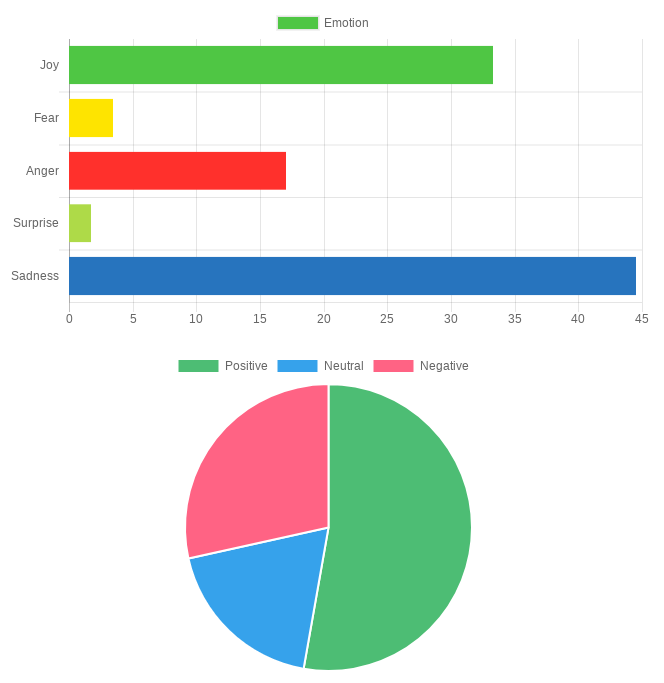
We have used the following technologies for many reasons. Hadoop assumes that conventional approaches (consisting of developing ever more powerful centralized systems) have technical and financial limitations. The development of distributed systems consisting of machines or nodes, relatively affordable (commodity hardware) and scaling out is an alternative from a technical and financial point of view A distributed system comprising tens, hundreds or thousands of nodes will regularly be confronted with hardware and / or software failures. Google has developed the Google File System (GFS), ancestor of the Hadoop Disrelated File System (HDFS) and The MapReduce Approach. MapReduce is a programming model designed specifically to read, milk and write very large volumes of data. A Hadoop program usually implements both map tasks and reduce tasks. Hadoop is particularly effective for dealing with problems that have one or more of the following characteristics: Volume of data to store or process very important. Need to perform processing on all data (batch rather than transactional, therefore). Heterogeneous data in terms of origin, structure, and format (JSON) Execute the tasks of a Hadoop job in parallel, without a pre-established order. A Hadoop cluster is made up of tens, hundreds, or thousands of nodes. It is the addition of the storage and processing capacities of each of these nodes which makes it possible to offer a storage space and a computing power yet to handle data volumes of several To or Po. To improve the performance of a read / write cluster, Hadoop’s file management system, HDFS, writes and reads files in blocks of 64 MB or 128 MB. Working on such large blocks maximizes data transfer rates by limiting search time on hard drives (seek time). // Input file graph and block MapReduce is a programming model designed specifically to read, process and write very large volumes of data. A Hadoop program usually implements both map tasks and reduce tasks. A Hadoop program is usually divided into three parts: The driver, which runs on a client machine, is responsible for configuring the job and submitting it for execution. The map is responsible for reading and processing data stored on disk. The reducer is responsible for consolidating the results from the map and write them on disk.

# Results



POMS (Profile of Mood States) is a psychological rating scale used for calculating the mood state score, the result depends on the values of 65 adjectives. For our project we reorganized the adjectives in two ways, first in 3 categories: Positive, Negative and Neutral, second way in 5 categories: Joy, Surprised, Fear, Angry, Sadness. More about the way to calculate POMS adjectives.





# Disucussion

We now turn our attention to the following interesting question: whether the subjective data that exist on the web carry useful information. Information can be thought of as data that reduce our uncertainty about some subject. According to this view, the diversity and pluralism of information on different topics can have a rather negative role. It is well understood, that true knowledge is being described by facts, rather than subjective opinions. However, this diversity in opinions, when analyzed, may deliver new information and contribute to the overall knowledge of a subject matter. This is especially true when the object of our study is the attitude of people. In this case, opinion native data can be useful to uncover the distribution of sentiments across time, or different groups of people. However, data mining differs from machine learning and statistics in that it deals with large volumes of data, stored primarily on disk. Some types of knowledge discovered from a database can be represented by a set of rules. The following is an example of a rule, stated informally: “Donald Trump with his totals of retweets incomes are greater than the average with the most sadly effects on users”. Of course, such riles are not universally true, and have degrees of “support” and “confidence”, as we shall see. Other types of knowledge are represented by equations relating different variables to each other, or by other mechanisms for predicting outcomes when the values of some variables are known. There are a variety of possible types of patterns that may be useful, and different techniques are used to find different types of patterns. Usually there is a manual component to data mining, consisting of preprocessing data to a form acceptable to the algorithms and post-processing of discovered patterns. For this reason, data mining is really a semiautomatic process in real life. The mode widely used applications are those that requires some sort of prediction. In our case, we want to predict emotions and sentiments, then a psychological profile. We outline what is classification, study techniques for building one type of classifiers, called decision-tree classifiers, and then study other predication techniques. Abstractly, the classification problem is this: Given that user belong to the archive, and given his tweet. We use a given instances (called training instances) of items along with the classes to which they belong, the problem is to predict the class (in our study it is a sentiment or an emotion) to which a new item belongs. This is an example with a sample of 383 623 424 tweets in English.

[tab:cross\_tab]

# Conclusion and Future Work

In this report we have presented a sentiment analysis tool on a Web interface, in one hand we used data from an archive end in the other we used real time stream analysis. Due to the absence of labelled data we couldn’t argue on reliability of data.

# References

We are grateful to the following people for resources, discussions and suggestions: Jason Scott (Archivist) and Diana Yuan (Co-Founder & Vice President, Talent & Operations from Indico).