**Project of Advanced Databases**

Define a problem you want to solve:

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 to 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.

Build an efficient web interface to get results easily.

Build a system without false accounts and keep a control on data continuously.

Propose and describe a solution:

An Apache Spark Implementation for Sentiment Analysis on Twitter Data

Alexandros Baltas, Andreas Kanavos, Athanasios K. Tsakalidis

Computer Engineering and Informatics Department, University of Patras, Patras, Greece

[Link of the publication](https://www.google.fr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwiOrIer0PDYAhXSh7QKHU1vD9MQFggoMAA&url=http%3A%2F%2Fwww.springer.com%2Fcda%2Fcontent%2Fdocument%2Fcda_downloaddocument%2F9783319570440-c2.pdf%3FSGWID%3D0-0-45-1604969-p180774546&usg=AOvVaw0OgLruU-ci8kQGXeGaV6aV)

Determining Word–Emotion Associations from Tweets by Multi-Label Classification

Felipe Bravo-Marquez, Eibe Frank and Bernhard Pfahringer

Department of Computer Science, The University of Waikato Hamilton, New Zealand

Saif M. Mohammad, National Research Council Canada Ottawa, ON, Canada

[Link of the publication](https://drive.google.com/open?id=1DB0f16xP6ds_5DnFKvnWfV05FA2Cd6or)

Emotions from text machine learning for text-based emotion prediction

Cecilia Ovesdotter Alm Dept. of Linguistics UIUC Illinois, USA ebbaalm@uiuc.edu

Dan Roth Dept. of Computer Science UIUC Illinois, USA danr@uiuc.edu

Richard Sproat Dept. of Linguistics Dept. of Electrical Eng. UIUC Illinois, USA rws@uiuc.edu

[Link of the publication](https://drive.google.com/open?id=1H13DAb73bA6uLXUn8t9-ryYtRDJG73gs)

An Extensive study of Sentiment Analysis tools and Binary

Vishal Vyasa, V.Umab

Department Of Computer Science, Pondicherry University, Puducherry- 605014, India

7-8 Dec. 2017

[Link of the publication](https://drive.google.com/open?id=1Hno_S_TTMSyAWrr7NwlpB_AvpaeVAN62)

Sentiment analysis on Twitter A text mining approach to the Syrian refugee crisis

Nazan Öztürka, Serkan Ayvazb

a Department of Computer Engineering, Bahcesehir University, Besiktas 34349, Istanbul, Turkey

23 Oct. 2017

[Link of the publication](https://drive.google.com/open?id=1h4jh1U4X2f9yLgRuk8jhze7YyxeJ_GNy)

Sentiment Analysis for Twitter: Going Beyond Tweet Text

Lahari Poddar, Kishaloy Halder, Xianyan Jia

School of Computing - National University of Singapore

26 Nov. 2016

[Link of the publication](https://drive.google.com/open?id=1PGYxdyubY-aHSN8A5tKWQjqZD1lgpYko)

Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis

ZHAO JIANQIANG, GUI XIAOLIN

1School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China

2Xi'an Politics Institute, Xi'an, Shaanxi Province, 710068, China

3Key laboratory of Computer Network of Shaanxi Province, Xi'an 710049, China

28 Mar. 2017

[Link of the publication](https://drive.google.com/open?id=1ilrcloKepqEPYqp4rUEMa46sSPzA4jse)

A Deep Learning Approach to Sentiment Intensity Scoring of English Tweets

Hardik Meisheri, Rupsa Saha, Priyanka Sinha, Lipika Dey

Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis

Copenhagen, Denmark, September 7–11, 2017

[Link of the publication](https://drive.google.com/file/d/1iZSWeuqbC6vC58As9A2S3YXVszo62Ipu/view?usp=sharing)

Survey on mining subjective data on the web

Mikalai Tsytsarau University of Trento, 38100 Trento, TN, Italy,

Themis Palpanas, University of Paris Descartes, 75005 Paris, France

Accepted: 8 September 2011, Published online: 19 October 2011

[Link of the publication](https://drive.google.com/open?id=1n_Jqa0_uRE_ev-J9Nhw6b-SLiCEXkMwz)

#Implement the solution:

The main subject is sentiments Analysis on Twitter a microblogging platform where people can easily share their thoughts about anything and their habits too. We have a lot of publications about sentiments analysis but not enough researches 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 to precision of results.

Event it’s an enormously continuous stream of data, Twitter is a good extra sentiment though an online community.

We will have several aims about what we would do with those data.

We don’t want to focus just on the tweets and sentiments themselves and but also getting the knowledge on how they impact people themselves too and their locations. Then it is relevant to learn how news affect people, since we live in a century that makes us extremely confronted constantly.

We’re also interested in focusing on the Big Data aspect because it’s important nowadays.

Finally, 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, real-time analysis and batch analysis.

A very quick presentation of all steps in the project:

1. Fetch data with API (main language we will use is Python), manage collections of stream data.
2. Pre-processing on data with Spark.
3. Build multiple module.
4. Analysis those data with multiple methods of machine learning.
5. Web interface.

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 precision of results.

Event it’s an enormously continuous stream of data, Twitter is a good extra sentiment though an online community.

Finally, 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.

A very quick presentation of all steps in the project:

1. Fetch data with API (main language we will use is Python), manage collections of stream data
2. Pre-processing on data (False account, spams, etc…)
3. Build multiple module (POMS, news, sentiment effects, etc…)
4. Analysis those data with multiple methods of machine learning.
5. Conception of Databases on user’s news psychological profile, locations.
6. Web interface

The main subject is sentiment Analysis on Twitter Data, Twitter is a microblogging platform allows users to share 140-character posts. The fact that it is small messages sanded continuously is making our work easier. Even it's an enormously continuous stream of data, there are now open source framework like Spark. Twitter is a good extra sentiment through an online community, and it has a real-life application such as recommendation systems.

The limited size of the messages is a positive for databases structure because we don't want to store a lot of data and negative for because we can't decide based on small messages in a short time.

Data storages like HDFS and HBase are becoming popular and useful for our project and algorithm exploits the hashtags and emoticons inside a tweet, as sentiment labels, and proceeds to a classification procedure of diverse sentiment types in a parallel and distributed manner.

Project Pipeline

The project is divided into four steps:

1. Fetch Data API Twitter/ Collections of Stream Datas
2. Organize sentences and useful words/ Wordnet and POMS
3. Analysis Data
4. Database of Users

Ideas about "Fetch Data API Twitter/ Collections of Stream Datas":

1. Use the Twython API.
2. Store data using Spark and Stream data options.
3. Think about a conception of a table about to organize sentences.
4. Various time of synchronization with the other table of data (Wordnet/POMS), hours/days/months
5. Checking false accounts
6. Storage about hundreds of users

Ideas about "Organize sentences and useful words/ Wordnet and POMS":

1. Get some information about Wordnet
2. How to gather adjectives for POMS calculation
3. How to add Lexicons of Emoticons with sentences to make (maybe) a better classification

Ideas about "Analysis Data":

1. Use unsupervised Learning methods for classification of sentiments
2. Use supervised learning methods based on POMS for predicating the profile of users

Ideas about "Database of Users":

1. Conception of the final DB that contains users and results

Conception of different part of the project and technologies

First, we will talk about scripts using Pyspark with Scikit-learn and TextBlop.

Scikit-learn allows us a a lots simple and efficient tools for data mining and data analysis, it’s accessible to everybody and reusable in various contexts. It’s Open source and commercially usable.

TextBlob is a library fro processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as sentiment analysis, classification, translation, and more. TextBlop use Google translate for sentences that are not in English, so in free and anonymous usage are limited 1000 words/day which is why we could not have analysed every tweet on the database.

The Spark Python API (PySpark) exposes the Spark programming model to Python. We used it because it simpler and we have a gain of productivity against language such as Scala or Java. Python is dynamically type, so RDDs can hold objects of multiple types.

To run all those libraries, we used Amazon EMR, an Amazon EMR release is a set of open-source applications from the big-data ecosystem. Each release comprises different big-data applications, components, and features that we selected to have Amazon EMR installed and configured when we create a cluster (more [here](https://github.com/yannistannier/twitter-sentiment-analysis/tree/master/spark-aws-emr)).

For the real-stream part, we created a module for Data Stream with Kinesis MQTT ([exemple](https://docs.aws.amazon.com/iot/latest/developerguide/kinesis-rule.html)). We used a StreamListener that will permanently get data from Twitter and with TextBlop we run an analysis on each tweet and sum all of them.

MQTT is a publish-subscribe messaging protocol based on the TCP / IP protocol, it was originally developed by Andy Stanford-Clark (IBM) and Arlan Nipper (EurtoTech), then offered to the Open Source community (For information, MQTT v3.1.1 is now an OASIS standard).

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, click [here](https://www.brianmac.co.uk/pomscoring.htm).

We developped the website using React.js, we could design simple views for each state in our application, and React efficiently update and render just the right components when your data changes. ([source code](https://github.com/yannistannier/twitter-sentiment-analysis/tree/master/interface-reactjs)).

Technologies/Methodologies/Software/Keywords:

Apache Spark, Big data, Machine Learning methodologies, Natural Languages Processing techniques, Spark Streaming.

Réferences:

Survey on mining subjective data on the web

Opinion aggregation over product reviews can be very useful for product marketing and positioning, exposing the customers’ attitude towards a product and its features along different dimensions, such as time, geographical location, and experience. Tracking how opinions or discussions evolve over time can help us identify interesting trends and patterns and better understand the ways that information is propagated in the Internet.

Quote this publication at page 8-9.

An Apache Spark Implementation for Sentiment Analysis on Twitter Data

Quote this publication at page 3-4.

An Extensive study of Sentiment Analysis tools and Binary Classification of tweets using Rapid Miner

Quote this publication at page 5-6.

Sentiment analysis on Twitter A text mining approach to the Syrian refugee crisis

Quote this publication at page 2 and page 6-7.

Sentiment Analysis for Twitter: Going Beyond Tweet Text

Quote this publication at page 1-2-3.

Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis

Quote this publication at page 2-3.

A Deep Learning Approach to Sentiment Intensity Scoring of English Tweets

Quote this publication at page 4-5. (Do a Pearson correlation between sentiment/emotions).

Emotions from text machine learning for text-based emotion prediction

Quote this publication at page 3.

Determining Word–Emotion Associations from Tweets by Multi-Label Classification

Quote this publication at page 2.