Sentiment Analysis on Twitter : Effects of a Social Network

Mohamed Ben Hamdoune and Yannis Tannier
University of Paris-Descartes

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.

Categories and Subject Descriptors: H.2.8 [**Database Applications**]: Big Data and Real Stream—*Cloud Computing*; I.3.7 [**Learning**]: Apache Spark and Hadoop—*Social Media Analytics*

General Terms: Twitter, Psychological Profil

Additional Key Words and Phrases: Data Mining, Natural Languages Processing, Sentiment Analysis, Classification, Text Mining

ACM Reference Format:

Mohamed Ben Hamdoune, Yannis Tannier. 2018. Sentiment Analysis for Twitter : Effects of a Social Network.

1. INTRODUCTION

The main subject is [Aitor García-Pablos 2017] Sentiment Analysis on Twitter, [Öztürka and Ayvazb 2017] 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 [Felipe Bravo-Marquez and Eibe Frank and Saif M. Mohammad and Bernhard Pfahringer 2016] research about impacts and their effect on society. The maximum characters [Zhao Jianqiang and Gui Xiaolin 2017] 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[Hardik Meisheri and Rupsa Saha and Priyanka Sinha and Lipika Dey 2017] accuracy of results. Event [Zhao Jianqiang and Gui Xiaolin 2017] its an enormously continuous stream of data, Twitter is a good extra sentiment though an online community. Therefore, how to optimize all those streaming data [Alexandros Baltas and Andreas Kanavos and Athanasios K. Tsakalidis 2017] and build a web interface for users who want to get data. Our project will [Vishal Vyasa and V.Umab 2017] use many methodologies from Machine Learning like unsupervised methods to make a classification of sentiments, and supervised method [Themis Palpanas and Mikalai Tsytsarau 2011] 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 [Poddar et al. 2016] results for example.

> The best way to engage honestly with the marketplace via Twitter is to never use the words "engage," "honesty," or "marketplace."
>
> (Jeffrey Zeldman)

In this study, we introduce readers to the problems of Data Processing and Cloud Computation, which have been rapidly developing over the last decade. The rest of this document is organized as follows. In Sect. 2, we provide a specific view of Tweet format. Development, problems, definitions and main trends of this area are described in Sect 3. We analyze in Sect 6, and discuss in Sect 7 about several problems we have been through. Finally, we talk about others idea and features that were not implemented due to a lack of time in Sect 8.

2. RELATED WORK

To begin, ye can refer to this article [Themis Palpanas and Mikalai Tsytsarau 2011] 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 peoples feelings, opinions, thoughts and behaviors from users text data using Natural Language Processing (NLP) methods. For methodologies on preprocessing, and feature generation, we based our work on [Hardik Meisheri and Rupsa Saha and Priyanka Sinha and Lipika Dey 2017] process. Then we can also cite [Zhao Jianqiang and Gui Xiaolin 2017] 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. Based on 36 million tweets collected from Twitter, Wang et al. proposed a real-time sentiment analysis system for classification of political tweets during 2012 US presidential elections. Their model achieved 59% accuracy in predicting the sentiments of political tweets (Wang et al., 2012). We had over 800 million tweets in English and all

---

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.

DOI:http://dx.doi.org/

that data to analyze where possible just because of the MapReduce Model [Alexandros Baltas and Andreas Kanavos and Athanasios K. Tsakalidis 2017] in majority.

3. METHODOLOGY

blabla

3.1 Preprocessing (Data clean)

blabla

3.2 Feature Generation

The data for this task consists of tweets across various domains, classified into four emotions: joy, sadness, anger and fear. The training data additionally carries a real-valued score between 0 and 1 per tweet, indicating the degree of the emotion (that the tweet is classified as) the present in the tweet.

4. CLOUD COMPUTING

Cloud computing was used for the project, the trend today is machine learning, which is a form of artificial intelligence that uses algorithms to learn from data. These systems build models from incoming transactional data, then find patterns in that data to make predictions. These predictions can be a simple as providing a recommendation to a shopper on an e-commerce website or as complex as determining if a brand of automobile should be retired. As with their learning-system forebears, the overhead of machine-learning systems is typically huge. But today we have the option to place these systems in the cloud. Amazon Web Services, for example, supports machine learning using AWS's algorithms to read native AWS data (such as RDS, Redshift, and S3). Google has supported predictive analysts for some time with its Google Prediction API, and Microsoft provides an Azure machine-learning service. The ability to predict the future for both tactical and strategic purposes has eluded us because of prohibitive resource requirements. But today, thanks to the cloud for machine learning as a service, you can apply this technology far and wide on all that data enterprises have been collecting.

4.1 AWS (Amazon Web Service)

blabla

4.2 MapReduce Model and Spark Framework

Here

4.3 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 [Alexandros Baltas and Andreas Kanavos and Athanasios K. Tsakalidis 2017] for building Web interfaces to databases.

5. IMPLEMENTATION

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, Hadoops 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.

6. RESULTS

7. APPLICATION AREA: PSYCHOLOGICAL PREDICTION

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.

8. 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. The term data mining refers loosely to process of semiautomaticcally analyzing large databases to find useful patterns. Like knowledge

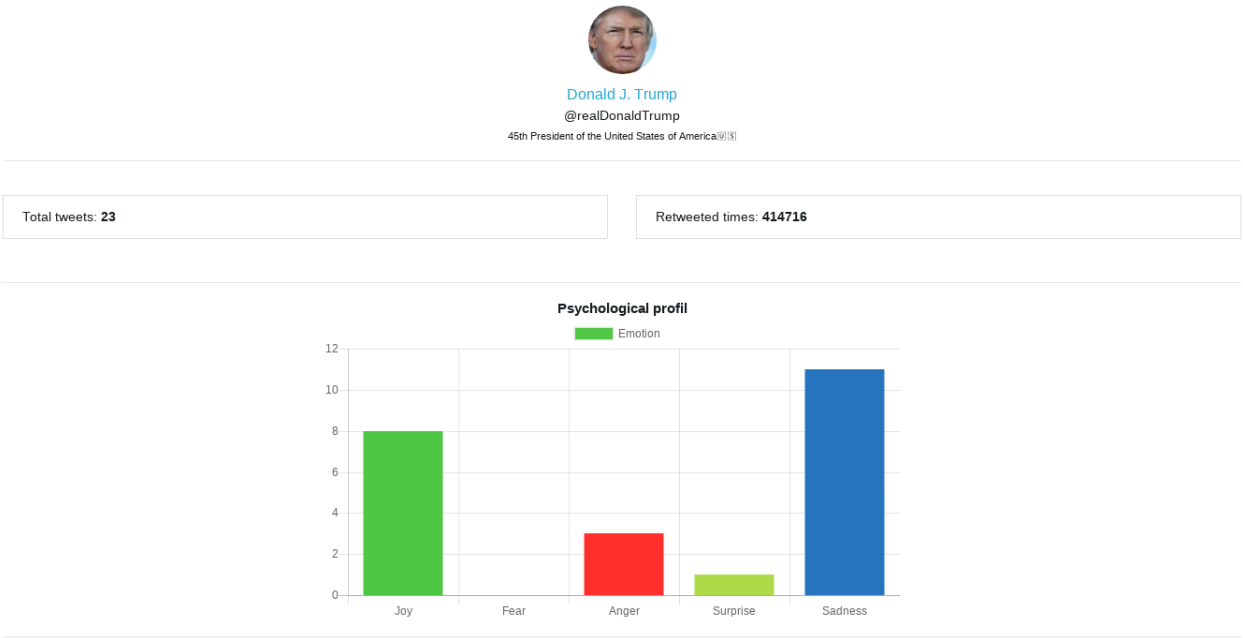

Fig. 1. Results predicted by our models with Donald John Trump (born June 14, 1946), the 45th and current President of the United States.

discovery in artificial intelligence (also called machine learning) or statistical analysis, we use data mining to discover rules and patterns from data. However, data mining differs from machine learning and statistics in that it deals with large volumes of data, stored primarily on disk. That is, data mining deals with knowledge discovery in databases. 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 postprocessing 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. Prediction is one of the most important types of data mining. 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.

9. 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 couldnt discuss the accuracies of the two enhancements. In the future, we plan to use the as feedback mechanism to classify new tweets.

10. REFERENCES

ACKNOWLEDGMENTS

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).

REFERENCES

German Rigau Aitor García-Pablos, Montse Cuadros. 2017. W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis. *Expert Systems With Applications* 91 (Sept. 2017), 127–137. http://dx.doi.org/10.1016/j.eswa.2017.08.049

Alexandros Baltas and Andreas Kanavos and Athanasios K. Tsakalidis. 2017. An Apache Spark Implementation for Sentiment Analysis on Twitter Data. (Aug. 2017), 15–25. DOI:http://dx.doi.org/10.1007/978-3-319-57045-7_2

Felipe Bravo-Marquez and Eibe Frank and Saif M. Mohammad and Bernhard Pfahringer. 2016. Determining WordEmotion Associations from Tweets by Multi-Label Classification. (oct 2016), 536–539.

Hardik Meisheri and Rupsa Saha and Priyanka Sinha and Lipika Dey. 2017. A Deep Learning Approach to Sentiment Intensity Scoring of English Tweets. *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* 8, 7 (Sept. 2017), 193–199. DOI:http://dx.doi.org/10.1038/s41598-018-20132-7

Nazan Öztürka and Serkan Ayvazb. 2017. Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. *Telematics and Informatics* 35 (Oct. 2017), 136–147. DOI:http://dx.doi.org/10.1016/j.procs.2017.12.044

Lahari Poddar, Kishaloy Halder, and Xianyan Jia. 2016. Sentiment Analysis for Twitter : Going Beyond Tweet Text. *CoRR* abs/1611.09441 (2016).

Themis Palpanas and Mikalai Tsytsarau. 2011. Survey on mining subjective data on the web. *Expert Systems With Applications* 24 (Oct. 2011), 478–514. DOI:http://dx.doi.org/10.1007/s10618-011-0238-6

Vishal Vyasa and V.Umab. 2017. An Extensive study of Sentiment Analysis tools and Binary. *Procedia Computer Science* 125 (Dec. 2017), 193–199. DOI:http://dx.doi.org/10.1016/j.procs.2017.12.044

Zhao Jianqiang and Gui Xiaolin. 2017. Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. *IEEE Access* 5 (Feb. 2017), 2870–2879. DOI:http://dx.doi.org/10.1109/ACCESS.2017.2672677