Towards a deep analysis of Twitter: tagging tweets and discovering user relations

M.F. Aparicio, J.M. Zurita, and J.L. Castro

Dept. Computer Science and Artificial Intelligence
University of Granada

Acknowledgement. This work was supported by the research project FFI2016-79748-R (AEI/FEDER, UE).

1 Introduction

Data flow nowadays is huge, specially on the Internet. Social media has become really popular in the last decade, providing new and almost unlimited ways of interpreting reality, as population can be segmented by analysing the content that is being generated.

Several social network based tools have been created to take advantage from this situation, from surveillance systems to marketing oriented recommendations systems [1, 2]. However, to the best of our knowledge, no tool that can perform deep analysis to identify underlying relations between users and to reveal clusters with same interests and/or tendencies has been published to date.

If we want to extract knowledge from social networks, first we must *teach* the machine. One way to achieve this is to use labelled data set from where generalisation can happen. The goal of this project is the first step towards our investigation: collecting and labelling data.

2 Domain

Twitter is currently one of the most used social networks in the world. It allows to publish short messages from up to 140 characters (*tweets*), share them (or *retweet* them) and mark someone else's message as relevant (add to favourites list) messages from other users. Twitter has more than 300 millions active users ([3]) who generate around 500 millions of tweets per day. Having all of this in mind, it is not surprising that it constitutes a way to quickly react to breaking news and events.

Our challenge is to collect and analyse these tweets in real time, as they are being written in an average rate of 6000 per second [4]. Nevertheless, the system has to be able to manage peaks even larger, as activity varies over time, specially when talking about trending topics.

The system should be parallel and/or distributed, as well as easily scalable, and it must perform data pre-processing, relevance analysis (using ontologies) and semantic labelling in order to build a data set which can be used for applying machine learning techniques.

3 State of the art

Social networks are perfect targets for making experiments and running data analysis tools, mainly oriented to marketing [5, 6, 7] and advertising. Moreover, Twitter allows, because of its inherent characteristics, to analyse society real time response to events and hot news.

One of the bests examples is [8], a tool developed in 2012 to analyse the sentiment towards the candidates to the presidency of the United States in real time. 452 out of 535 members that conform the Congress are active on Twitter [9], and approximately 21% of the population has an account on this social network [10], so it is fair to say that the sample is quite significant. The tool can analyse tendencies of sentiments during public events by using a number of modules for analysing, pre-processing, classifying and visualizing.

The rest of the researched tools ([11, 12]) follow a similar structure, but with a sequential pipeline.

4 Goal

We could not find any tools that suited our needs and therefore we outlined the development of a multi-platform, efficient and scalable tool to monitor tweets in real time and to perform semantic labelling (using ontologies that can be expanded for multidisciplinary profiles). Ultimately we expect to supply the tool with the capacity of computing potentially relevant parameters for deep analysis (e.g. general sentiment) and integrate all of this into a data set building tool.

5 System Design

We opted for a distributed, highly-scalable architecture based on microservices. The processing pipeline has been divided into three main services and a web server that is used to manage the whole system, using the `ZMQ` library to communicate them through two channels: one for data transmission and another one for commands.

5.1 Producer

The producer node connects to the Twitter Streaming API and it is responsible for getting tweets in real time and transforming them into individual tasks that can be equally sent to the parser nodes. This service can be configured remotely and it keeps concurrently listening to commands that are sent by the web server.

5.2 Parser

Once the parser receive a task the analysis begins. It starts tokenizing the tweet's content, taking into consideration the particular structures of the social network (like usernames, hashtags, emojis...). Then we perform a semantic search based on the ontologies that where previously set by the user and update some statistics, like count of most used words and n-grams. Finally, we geocode localizations in geographic coordinates for those tweets that did not have them previously.

General Sentiment Analysis We could not use pre-trained models for sentiment analysis in Spanish [13], so we trained our own one using `scikit-learn` and *TASS* [14] corpus.

Using this data set, we performed text pre-processing removing stopwords, as well as a tokenizing and stemming the words. To convert documents into features, we used a dictionary based technique (bag of words) so each document can be covered with a binary vector that represents the presence or absence of the word in each document.

Lastly, we compared different models (see 1) and we chose the best of them (logistic regression) from its score in cross validation ROC AUC. Performing a grid search to optimise hyper parameters brought up a precision of 0.83175 for this model.

Model	*Accuracy*	*Recall*	*Precision*	F_1 *score*	*AUC*
LR	0.75084	0.80327	0.76850	0.78538	0.82929
kNN	0.61377	0.78444	0.62859	0.69751	0.66624
RF	0.68071	0.70454	0.72682	0.70844	0.68719
NB	0.69339	0.56851	0.84003	0.67743	0.77857
SVM	0.56803	0.99965	0.56797	0.72437	0.79060

Table 1: Classification model comparison for Spanish sentiment analysis. We present results for models based on logistic regression, k-Nearest Neighbours, Random Forest, Naive Bayes and SVM, all of them with *Bag of Words* as a feature extraction method.

5.3 Collector

Now that analysis is finished, the last node handles the connection to the parsers and requests statistics. This nodes also connects to the web server, so the actions that the user performs over the web interface are transmitted to the rest of the system through commands using this node.

5.4 Relations Visualisation Module

We manage to establish relations (TIN2013-48319-R) between the authors of the tweets, and some of them can be directly represented graphically by using a graph.

Relation environment Given a user u, their relation environment is composed of all the users that interact with them through responses or retweets.

Dialog environment Given a user u, their dialog environment is composed of all the users that interact with them through replies, as long as u also interacts with them.

Interest environment For each user u, their interest environment is defined as all the users that share interests with u. We establish that they share an interest if they have retweeted or replied to the same tweets.

Ideas environment For each user u, their ideas environment is defined as all the users that share ideas and interest with u. We establish that they share ideas if they have retweeted the same tweets.

Based on this relations, we transform each user into a node and each relation between them into edges of the graphs. By looking at this representation it is very easy to see which users are relevant and which clusters are related.

6 Conclusions and future works

Twitter, and all social media in general, constitute a wide and an important source of data whose segmentation might by useful in a lot of fields. However, the task is not easy and normally we will need to collect data and structure it. Our project allows to build semantic tagged data sets, and its main features are:

- Multiplatform, distributed higly scalable system based on microservices.
- Editable ontologies built by users (for a multidisciplinary profile).
- General sentiment analysis of tweet.
- Pleasant and easy-to-use web interface.

Although this is a good start, it is necessary to keep developing the tool:

– Multi-source support.
– Heterogeneous pipelines.
– Groups identification through clustering techniques.
– Precise rules to identify underlying relations between users and clusters.

Suggesting and identifying new relations and ways to analyse clusters of users will be very challenging. Which properties of tweets and authors could be included in the analysis? Which relations will lead us to extract more underlying information?

References

[1] Kathy Lee, Ankit Agrawal, and Alok Choudhary. "Real-time Disease Surveillance Using Twitter Data: Demonstration on Flu and Cancer". In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '13. New York, NY, USA: ACM, 2013, pp. 1474–1477. ISBN: 978-1-4503-2174-7. DOI: 10.1145/2487575.2487709. URL: http://doi.acm.org/10.1145/2487575.2487709.

[2] Zahra Ashktorab et al. "Tweedr: Mining twitter to inform disaster response." In: *ISCRAM*. 2014.

[3] Statista. *Twitter: number of active users 2010-2017*. URL: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ (visited on 12/27/2017).

[4] *Twitter Usage Statistics - Internet Live Stats*. URL: http://www.internetlivestats.com/twitter-statistics/ (visited on 12/27/2017).

[5] *Alltop, all the top stories*. URL: http://alltop.com/ (visited on 01/05/2018).

[6] *Social Media Software for Community Management & Social Support*. URL: https://www.lithium.com/ (visited on 01/05/2018).

[7] *Sysomos*. URL: https://sysomos.com/ (visited on 01/05/2018).

[8] Hao Wang et al. "A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle". In: *Proceedings of the ACL 2012 System Demonstrations*. ACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 115–120. URL: http://dl.acm.org/citation.cfm?id=2390470.2390490.

[9] *Tweet Congress*. URL: http://www.tweetcongress.org (visited on 12/27/2017).

[10] *Twitter MAU in the United States 2017 | Statistic*. URL: https://www.statista.com/statistics/274564/monthly-active-twitter-users-in-the-united-states/ (visited on 12/27/2017).

[11] Michael Mathioudakis and Nick Koudas. "TwitterMonitor: Trend Detection over the Twitter Stream". In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*. SIGMOD '10. New York, NY, USA: ACM, 2010, pp. 1155–1158. ISBN: 978-1-4503-0032-2. DOI: 10.1145/1807167.1807306. URL: http://doi.acm.org/10.1145/1807167.1807306.

[12] ExportTweet. *Twitter Analytics and Hashtag Tracking by ExportTweet*. URL: https://www.exporttweet.com (visited on 01/05/2018).

[13] Carlos Henríquez Miranda et al. "A Review of Sentiment Analysis in Spanish". In: *Tecciencia* 12.22 (June 2017), pp. 35–48. ISSN: 1909-3667. DOI: 10.18180/tecciencia.2017.22.5. URL: http://www.scielo.org.co/scielo.php?script=sci_abstract&pid=S1909-36672017000100035&lng=en&nrm=iso&tlng=en (visited on 12/08/2017).

[14] Julio Villena-Román et al. "TASS : Workshop on Sentiment Analysis at SEPLN". In: *Procesamiento del Lenguaje Natural* 50.0 (Apr. 2013), pp. 37–44. ISSN: 1989-7553. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4657 (visited on 12/05/2017).