

Massive Data Analytics' Project Proposal

Sandro Cavallari, Marco Giglio, Paolo Morettin

1 Introduction

Social networks experienced an exponential growth in the last five to ten years. In a few years they became one of the most used communication media: several people, nowadays, tend to spend many hours per day writing on their *walls*, *twitting* etc and basically every big company, important personality, or club, manages accounts on several social networks, using them as its most important communication media.

Given the increasing role of social networks in everyday's lives, researches became interested to them, questioning how they affect our privacy and behavior [1][2] or examining the role they fulfilled during some important recent events, such as the Arab Spring [3][4].

2 Project description

Our interest is to monitor trends on a social network, understand whether people are feeling positive or negative toward a certain topic and correlate this feeling with recent news coming from newspapers. In detail, our project aims in developing a methodology in order to:

1. understand when the common feeling about a certain topic shifts from positive to negative, as done in [5];
2. correlate this shift to news coming from newspapers and news agencies.

The social network we will focus on is Twitter, an online social network that allows users to upload short text messages (*tweets*) of up to 140 characters.

3 Work plan

We can divide our work in 3 different phases: the first one to collect data; the second one to find sentiment change in tweets; the last one to correlate the results coming from phase 2 to news coming from other sources.

3.1 Phase 1

All the analysis we want to perform are only possible on big datasets; otherwise, on small datasets, it becomes hard to correctly identify a sentiment shift.

For this reason, we need to collect a big amount of data coming both from Twitter and from newspapers and news agencies.

We are planning to collect these information for one month, exploiting a web server up 24 hours per day. In order to collect Twitter data, we will use a Java library called *twitter4j* whereas the news will be extracted from several RSS feeds. All the resulting data will be stored on a database on the server.

3.2 Phase 2

All the data collected in phase 1 will be download on faster machines to be analyzed. In this new phase we aim in finding some instant in which the common feeling toward a certain topic shifted from positive to negative.

Researchers already tried and succeeded in similar tasks [5][6][7], hence several techniques will be investigated and analyzed in order to find the one which better suits what we are trying to accomplish.

3.3 Phase 3

References

- [1] Debatin B. et al., “*Facebook and Online Privacy: Attitudes, Behaviors, and Unintended Consequences*”, Journal of Computer-Mediated Communication, 15, pg. 83-108 (2009)
- [2] Acar A., “*Antecedents and Consequences of Online Social Networking Behavior: The Case of Facebook*”, Journal of Website Promotion Vol. 3, N. 1-2, pg. 62-83 (2008)
- [3] Howard P. et al., “*Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring?*”, ICT4D Bibliography (2011)
- [4] Lotan G. et al., “*The Revolutions Were Tweeted: Information Flows During the 2011 Tunisian and Egyptian Revolutions*”, International Journal of Communication 5 (2011)
- [5] Albert B. et al., “*Detecting Sentiment Change in Twitter Streaming Data*”, JMLR: Workshop and Conference Proceedings 17 (2011) 5-11
- [6] Go A., Bhayani R., Huang, L., “*Twitter sentiment classification using distant supervision*”, CS224N Project Report, Stanford, 1-12 (2009)
- [7] Tsytsarau M., Palpanas T., “*Survey on mining subjective data on the web*”, Data Mining and Knowledge Discovery, Vol. 24, N. 3, pg. 478-514, Springer (2012)