Market Research Using Social Media

Tran Van Canh, L3S Research Center, Hannover, Germany
Room 1519, Appelstraße 9A, 30167 Hannover
Contact email: ctran@l3s.de
Ayser Armiti, head of Data Science, Moovel GmbH, Stuttgart

Abstract
Correct strategic decision is a critical task to any business or investor. To minimize risk, it is required to understand the product and services offered by such a business and to study the performance of any competitor. These days, social media provides a very rich content about people's opinions, habits, and expectations. Utilizing this content to study the market saves a lot of time and effort. In this thesis project, we propose to apply machine learning techniques including topic models and ranking algorithms to recommend trending topics in the social media that are related to a business interest.

Motivation
Market research is defined as the process of gathering and analyzing the market to understand people's opinions related to a product or a service. For years, this was done through surveys and domain experts. However, the revolution of social media provides rich data sources that give a deep insight from the market. In the following, we summarize the main benefits for utilizing social media for market research.

- Save money and time: conducting telephone surveys and face-to-face interviews require a lot of money in addition to the weeks needed to organize a survey and then to analyse the outcome of the whole process.

- Reach a wider audience: traditional approaches on market research take a sample of the whole audience and most of the time does not follow a normal distribution. However, these days nearly everybody use social media, which means that one can reach the whole market.

- Get fast feedback in real-time: today, social media and microblogging services spread events occurring in a certain place even faster than via news agencies. This real-time feedback enables stakeholders to react in real time with any change in the market.

- Get genuine customer reviews: in traditional surveys, the user is guided by a set of questions that could lead to misleading interpretations. However, users using social media express their options precisely and in their languages.

Problem Statement
In this project we are interested in Twitter as our data source for market research. Twitter proved to be a good data source for many application domains such as semantic-based community detection [5] event detection [4,6,8], social opinion mining [12], and

recommender systems [9]. Given a set of keywords W, one can use Twitter public API [1] to retrieve a subset of tweets that contain any keyword w in W.
Extracting useful information from tweets is a very challenging task and an open problem that attracts researches from different domains [4,5,11]. In the context of this research topic, there are a number of challenges to tackle: (1) the huge number of tweets published on a certain topic, (2) short messages: each tweet has a maximum length of 140 characters. Hence, people usually use abbreviations and shortcuts to convey their ideas and thoughts. (3) Twitter content contains a lot of typos, grammatical errors and slang. As a result, this prevents us from using traditional NLP techniques to pre-process and analyze this unique textual content. (4) the dynamics of Twitter content. Most of the approaches to extract trending topics assume a corpus of documents with a fixed vocabulary set. Unfortunately, at every single moment, there are a large number of new words, which complicates processing the incoming stream in an online fashion. On the other hand, the temporal usage pattern of each word is time-dependant, showing different usage behaviors based on the posting time [3]. Our targets to provide a useful content from the Twitter stream can be divided into two main components as follows:

1. **Topic extraction**: initially a set of topics is extracted from the crawled tweets, where each topic is described by a set of keywords. A weight is associated with each keyword, describing how relevant this keyword is to the topic it belongs to. Topic extraction from twitter data is a challenging problem. Traditional topic models like LDA [7] assume a predefined set of vocabularies. However, such techniques might not adapt well to the streaming nature and dynamics of twitter messages [3].

2. **Tweets ranking**: extracting topics gives a general idea about the content retrieved from the Twitter stream. However, for market research it is important to reach the audience and read their feedback. For this, it is important to rank the tweets related to each topic. Each tweet is given a score reflecting how semantically-close this tweet is to the topic described by other keywords.

Objectives

The main goals of this thesis project can be summarized as follows, however, the student is expected to contribute more ideas and creativities.

- Use Twitter Public API to crawl tweets that are relevant to a certain topic under study.
- Adopt the state-of-the-art NLP-based technologies to preprocess the content of Twitter messages and to mitigate the noisy nature of their content.
- Online extracting the latent topics from a stream of twitter messages.
- Provide a ranked list of recommended tweets that semantically match a set of predefined keywords received from users as input.

References

1. Public Twitter API, http://dev.twitter.com/pages/streaming_api

2. PearAnalytics. Twitter study, August 2009. http://pearanalytics.com/wp-content/uploads/2009/08/Twitter-Study-August-2009.pdf

3. Hamed Abdelhaq, Michael Gertz, and Christian Sengstock. Spatio-temporal characteristics of bursty words in twitter streams. In Proceedings of the 21th International Conference on Advances in Geographic Information Systems, SIGSPATIAL ’13, pages 149–158, 2013.

4. Hamed Abdelhaq, Christian Sengstock, and Michael Gertz. Eventweet: Online localized event detection from twitter. Proceedings of the VLDB Endowment, 6(12):1326–1329, 2013.
5. Tran Van Canh and Michael Gertz. rLinkTopic: A Probabilistic Model for Discovering Regional LinkTopic Communities. 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ASONAM 2014. Beijing, China, August 17-20, 2014

6. Charu C. Aggarwal and Karthik Subbian. Event detection in social streams. In SDM, pages 624–635, 2012.

7. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022, March 2003.

8. Adrien Guille and Cécile Favre. Event detection, tracking, and visualization in twitter: a mention-anomaly-based approach. Social Network Analysis and Mining, 5(1), 2015.

9. Einat Minkov, Ben Charrow, Jonathan Ledlie, Seth Teller, and Tommi Jaakkola. Collaborative future event recommendation. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM ’10, pages 819–828, 2010.

10. Fred Morstatter, Ju¨ rgen Pfeffer, Huan Liu, and Kathleen M. Carle. Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose. In Proceedings of ICWSM, 2013.

11. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World Wide Web, WWW ’10, pages 851–860, 2010.

12. Pawel Sobkowicz, Michael Kaschesky, and Guillaume Bouchard. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. Government Information Quarterly, 29(4):470 – 479, 2012.