Knowledge Discovery from News Events on Social Media*

(Ph.D. thesis summary)

Mauricio Quezada

Department of Computer Science, Universidad de Chile
Millenium Institute for Foundational Research on Data
Santiago, Chile
mquezada@dcc.uchile.cl

Abstract. Online activity involves the consumption and production of event-related content. There are about 500 million Twitter messages published every day, and according to surveys, 59% of its users use the platform as a way to get the news. Its high rate of production of multimodal content (text, images, and videos) necessitates having flexible models to understand the dynamics of the information disseminated on social media. This thesis proposes the creation of context models from user-generated messages on Twitter to discover knowledge as a way to perform high-level quantitative analysis of news events. These models are useful in three perspectives: the spatio-temporal context in which the events develop, the activity of users that react when a high-impact event happens, and the multimodal content that can be exploited to generate a comprehensive summary of the event. Our current work involves the creation of a geopolitical model that relates events and countries, allowing us to discover international relations; the study of what features make an event susceptible to provoke high activity from users, and a characterization that allows us to predict with high precision which events are going to produce high activity. This includes our ongoing work on generating automatic multimodal summaries of events based on the assumption that the users describe the non-textual content in their tweets when they express their facts and opinions around events.

Keywords: Social Media · News events · Document models.

The so-called *Web 2.0* represents a change of state in how users interact with the Web. It is mainly defined as a platform destined to encourage end-users to publish content. One of the main manifestations of this phenomena are the *online social networks*, also called *microblogging platforms*. Users online make

* This work was supported by the Millennium Institute for Foundational Research on Data (IMFD), and by CONICYT PCHA/Doctorado Nacional 2015/21151445.

connections with others based on different criteria, and start to produce content that might be interesting to other users. Microblogging services such as Facebook, Twitter, or Sina Weibo are nowadays among the most used platforms for users to connect with family, friends, acquaintances, co-workers, or strangers with similar interests. Users interact with each other and produce or share content, which could be about their lives, their thoughts, about what is happening around the world, etc. This collective of information published in these Internet-based applications, such as microblogging platforms, blogs, wikis, etc., is what is called *social media* [5].

Content in social media is multimodal. Twitter, for example, encourages users to publish short texts (initially limited to 140 characters, now 280), but it recently started to incite users to share more photographs[1]. And with the proliferation of smartphones and internet-connected devices, more of this information is also geo-tagged and "real-time". Whether via text, images, videos, sounds or hyperlinks, social media lowered the entry barriers to content producers and made it easy for consumers to access to a myriad of different pieces of information.

The influence of social media on society can not be denied. It has facilitated the communication between people and speeded-up the difussion of information online. For instance, it is believed that the revolutionary wave of protests and uprisings in the Arab states (known as the Arab Spring which began in 2010) was highly influenced by social media as a means to organize and facilitate communication [3]. Also, for instance, it has permitted many applications in emergency management and detection, such as earthquake alert systems using Twitter [9, 10, 6]. The usefulness of quantiative analysis of events through time is undeniable, and social media offers a window to see and capture information about those events, how they develop, and how the world interprets them.

One of the main usages of social media platforms is the consumption and generation of event-related content. According to a recent 2018 study[2], about two thirds of U.S. adults get their news on social media. The most used platforms to get the news are Facebook, Youtube, and Twitter, while over 70% of Twitter users surveyed use that platform to do so. Furthermore, nowadays almost every news outlet has a presence in social media, in order to attract readers and viewers. In this way, users comment on the news events, reacting to them according to a myriad of factors, and many of these characteristics are present in one way or another in social media. We see social media as a medium that reflects an important part of what society thinks about what is happening the world.

However, the popularity of social media is not without issues. *Information overload* refers to the problem of being unable to manage or to make decisions based on data, due to the high volume of information available and the limited capabilities of the person who is dealing with it. Humans have limited cognitive processing capacities, and when they are overloaded with information, their

[1] https://techcrunch.com/2019/03/13/twitter-news-camera/ (Accessed: Jun 27, 2019)
[2] http://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/ (Accessed: Jun 27, 2019)

quality of decision making suffers [2]. In the context of social media, the high availability of diverse information may prevent users to find relevant content.

Finding relevant content in social media is not easily solved by search engines. Publications on social media, or *posts*, can be of *variable quality*. Posts are composed of multimedia pieces of content, but often they are brief or short. For instance, a post can be a very short text, an hyperlink, a single image or video with little context. They can be also irrelevant to the user's interest, for example, spam posts, which contain relevant keywords but in a misleading way, in order to lure users into a irrelevant website. Posts can be also out-dated, delivering incorrect or obsolete information. Many posts can be duplicate ones, published by automated agents, or by users using "share buttons" in websites which publish posts with a template text; they can be also near-duplicate, with little text differences, or sharing the same resource from different URLs. Another important characteristic of posts is that they are written in natural language, so they can be incorrectly capitalized, misspelled, or with ambiguous meaning. Users make use of colloquial language and different forms of expression when publishing content, e.g., abbreviations, hashtags (tags to describe content), emojis (ideograms), etc. Finally, messages can be also misleading, sharing false information. All of these particularities of social media make it difficult to apply standard techniques in order for users to find relevant content.

We regard news events as a higher level abstraction than single posts. In related work, an event is deemed as *something that happens in a certain place and time* [11], while other definitions consider an event as a collection of documents related to a certain occurrence [1]. Throughout this dissertation, we will consider an event as a collection of social media posts describing or commenting on a real-world occurrence. In this sense, an event is a more complex piece of information compared to single posts as it leads to new tasks, such as event detection, tracking, or summarization. Also, an event is comprised of posts of heterogeneous quality, from different locations, and at different times. This yields to new problems and challenges when studying social media.

In this dissertation, we tackled the problem of *extracting useful knowledge* from events on social media. In order to be able to infer and extract useful information from events, we propose the development of event representations that leverage specific features according to the desired goal when analyzing social media data. For this, we propose different models or representations of events based on three perspectives:

1. **User activity.** When users react to an event, they may manifest this reaction on social media, producing or sharing content relevant to the event. The characteristics of these manifestations are dependent on the proper features of the occurrence, and not all are equal. We look at how the activity of users can give us insights on the proper features of an event, and incorporate this behavior in a compact representation.
2. **Spatio-temporal context.** Events develop in different locations. On the other hand, users from different locations may react differently to the same event. We study the development of events based on user activity conditioned

by the location users are from, proposing a representation of events and locations based on social media posts.

3. **Common features in content.** Users may publish similar pieces of content in social media in reaction to events. However, each post can contribute to a different aspect of the event, while having some features in common. We leverage these commonalities in content to produce a compact model that preserve topical information in events.

Our main objective is to define event representations through different data aggregations to perform quantitative analysis of news events on social media. Currently, it is very difficult to manage and analyze the high volume of information being published when a event happens in the world. In particular, we study events through three perspectives: user reaction and activity, spatio-temporal context of events, and content aggregation. Understanding user reaction involves discriminating which events are more important or produce more impact in a community. Analyzing spatio-temporal context refers to understanding how communities from different locations are affected by different events, as seen on social media, and identifying similar communities and events based on this context. Understanding content refers to the identification of the core aspects of an event, without having to go through all the –potentially various– posts. In particular, our goal is to propose different models for representing events. These models should be flexible enough to apply diverse methodologies to discover useful knowledge from information published on social media about real-world events, from the perspectives described above.

We chose Twitter as our data source for this work. Twitter provides a simple way to obtain data and via its API (Application Programming Interface), from which we can obtain tweets automatically and programatically. Furthermore, it is not as restrictive as other sources, such as Facebook, which incentive users to maintain a private profile, hence making the data collection much more difficult or impossible to perform. Also, services like Facebook encourage users to share potentially personal content, and not just event-related. Twitter is primarily dedicated to encourage users to share event-related posts, and it is mainly used as a news source by its users (about 70% of its users use Twitter to get news). For instance, its website asks "What's happening?" to users when publishing content, as opposed to Facebook's "What's in your mind?" Therefore we see Twitter as a suitable platform for this type of work.

The study of news events on social media has several applications in the proposed setting. How the community reacts to different events would allow us to identify the characteristics of these events by a measure of reaction or other features given by the social network or the content of the events. By these characteristics, it would be possible to identify or even predict which events are going to cause a significant reaction from the community, improving journalistic coverage or better response from authorities facing an emergency. Additionally, by studying not only the response, but the context of different communities and how they respond to certain events, may give us insights about the communities themselves, for example, by revealing unexpected relations between differ-

ent communities, or by measuring event similarity using the context, instead of content-based features. On the other hand, the study of the content is useful to understand the different points of view ahead of an event. Users accustomed to the same perspectives given by other users or sources may be oblivious of other angles of the same news event. Being capable of identify the different aspects of an event and then present these aspects in a concise summary can deal with this problem. All in all, the proposed framework can be of utility to understand social behavior, to study and decrease the effects of the information overload, as well as to perform comparative historical research[3].

Thesis statement. This dissertation defines flexible models for events on the social networking platform Twitter. Having three perspectives in mind, user reaction, spatio-temporal context, and content, the defined models should be able to allow us to discover new insights about news events reflected on Twitter. The thesis statement is as follows:

Modeling news events from user-contributed content on Twitter, based on their spatio-temporal context, the reaction the users had on them, or the multimedia content which the events contain, is novel and effective to perform high-level quantitative analysis of news events.

Quantitative analysis of news events is useful to understand how the news impact society. For example, how are they perceived by users, or how can we archive this kind of information for future use, as more content is produced in a digital-only format.

Challenges. We identified three main challenges.

- **Retrieval of relevant posts.** Social media offers a partial view of the world. Also, mainstream topics obfuscate distinct points of view, which can obstruct retrieval of diverse content. Because users are frequently posting messages about their own lives, daily situations, or general topics, trends can be only visible when looking at large volumes of data. This makes identification of events and relevant content a very difficult task. And due to the characteristics of Twitter (or any other social networking service), usually messages are very short and with grammar and spelling errors. Also, users spontaneously create new ways to refer to the same entities (e.g. via the use of hashtags, emojis, or abbreviations), which makes it difficult to identify more relevant content when detecting events. The challenge comes in how to identify such content in an efficient way, how to deal with duplicated or quasi-duplicated content, and how to evaluate the effectiveness of methodologies when presenting multimedia content.
- **Biases in sampled social media data.** As we stated above, social media offers a partial view of the world. Furthermore, the employed methodologies

[3] https://en.wikipedia.org/wiki/Comparative_historical_research (Accessed: Jun 27, 2019)

to retrieve or identify events from social media may be biased depending on several factors. For instance, our dataset is collected using news outlets as sources, being the majority of the outlets coming from the USA or the UK. Also, our sources use specific words and ways to express the information, which can also create a bias in the way we further retrieve more tweets. This is a huge challenge in order to provide generalizable results from the proposed methodologies. Also, it is challenging to ensure that our results are as diverse as the utilized data source.

- **Lack of ground-truths.** As data in social media is being published at all times, it is unfeasible to apply standard measures, such as recall, when evaluating a methodology, because we do not have available all the relevant content. On the other hand, there are no *gold standards* we can contrast our models with. We need to come up with methodologies to validate our results, in order to provide generalizable results.

Contributions. Our contributions are the following:

1. A novel event representation based on user activity triggered by news events on Twitter [4]. This representation allows us to rank events into different levels of activity. We also show that the activity can be determined by other event features, and that these features appear early on the development of events. We show that it is possible to early *predict* the level of activity of an event using aggregated post features.
2. A spatio-temporal representation based on the location where an event happens, and the locations the users commenting on the news are from [7]. With this type of representation, we can compare events and locations based on different factors, and track the evolution of an event based on the locations involved in it.
3. A lightweight representation of content based on shared URLs [8]. We aggregate event-related posts based on common relevant URLs, retweets and replies, generating a compact representation of an event. In our preliminary experiments, we observed that the representation is one order of magnitude smaller than the original data. At the same time, we observed that with our representation we can achieve comparable clustering results, with a fraction of running time and memory required.
4. An event collection methodology based on *seed news outlets* (as part of [4]). Given a set of news outlets, we extract every hour the most relevant keywords from their headlines and use them to retrieve relevant tweets from regular users. We also made available a dataset of 193 million tweets of 25 000 news events, from 2013 to 2015.

Even though the different points of view posed as themes for this project cover mostly independent approaches of event mining, they have in common the goal of exploring and studying how different data aggregations can be useful to extract useful knowledge from events. This dissertation can be viewed as an exploration on how different data aggregation strategies applied to events on

social media are useful to easily extract knowledge or to serve as building blocks for new models and methodologies.

References

1. Becker, H., Naaman, M., Gravano, L.: Learning similarity metrics for event identification in social media. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining. pp. 291–300. WSDM '10, ACM, New York, NY, USA (2010). https://doi.org/10.1145/1718487.1718524, http://doi.acm.org/10.1145/1718487.1718524
2. Gross, B.M.: The managing of organizations: The administrative struggle, vol. 2. [New York]: Free Press of Glencoe (1964)
3. Howard, P.N., Duffy, A., Freelon, D., Hussain, M.M., Mari, W., Maziad, M.: Opening closed regimes: what was the role of social media during the arab spring? Available at SSRN 2595096 (2011)
4. Kalyanam, J., Quezada, M., Poblete, B., Lanckriet, G.: Prediction and characterization of high-activity events in social media triggered by real-world news. PloS one **11**(12), e0166694 (2016)
5. Kaplan, A.M., Haenlein, M.: Users of the world, unite! the challenges and opportunities of social media. Business horizons **53**(1), 59–68 (2010)
6. Mendoza, M., Poblete, B., Valderrama, I.: Nowcasting earthquake damages with twitter. EPJ Data Science **8**(1), 3 (Jan 2019). https://doi.org/10.1140/epjds/s13688-019-0181-0, https://doi.org/10.1140/epjds/s13688-019-0181-0
7. Peña-Araya, V., Quezada, M., Poblete, B., Parra, D.: Gaining historical and international relations insights from social media: spatio-temporal real-world news analysis using twitter. EPJ Data Science **6**(1), 25 (2017)
8. Quezada, M., Poblete, B.: A Lightweight Representation of News Events on Social Media. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM (2019). https://doi.org/10.1145/3331184.3331300, (to appear in the proceedings.)
9. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: Real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web. pp. 851–860. WWW '10, ACM, New York, NY, USA (2010). https://doi.org/10.1145/1772690.1772777, http://doi.acm.org/10.1145/1772690.1772777
10. Sarmiento, H., Poblete, B., Campos, J.: Domain-independent detection of emergency situations based on social activity related to geolocations. In: Proceedings of the 10th ACM Conference on Web Science. pp. 245–254. WebSci '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3201064.3201077, http://doi.acm.org/10.1145/3201064.3201077
11. Yang, Y., Carbonell, J.G., Brown, R.D., Pierce, T., Archibald, B.T., Liu, X.: Learning approaches for detecting and tracking news events. IEEE Intelligent Systems and their Applications **14**(4), 32–43 (1999)