Early prediction and characterization of high-impact world events using social media

Janani Kalyanam\textsuperscript{1}, Mauricio Quezada\textsuperscript{2}, Gert Lanckriet\textsuperscript{1} and Barbara Poblete\textsuperscript{2}

\textsuperscript{1}Department of Electrical and Computer Engineering, University of California, San Diego, USA
\textsuperscript{2}Department of Computer Science, University of Chile, Santiago, Chile

Abstract

On-line social networks publish information about an enormous volume of real-world events almost instantly, becoming a primary source for breaking news. Many of the events reported in social media can be of high-impact to society, such as important political decisions, natural disasters and terrorist actions, but might go unnoticed in their early stages due to the overload of other information. We ask, is it possible to clearly and quickly identify which of these news events are going to have substantial impact before they actually become a trend in the network?

We investigate real-world news discussed on Twitter for approximately 1 year, consisting of 5,234 news events that are composed of 43 million messages. We show that using just the first 5\% of the events’ lifetime evolution, we are able to predict with high precision the top 8\% that have the most impact. We observe that events that have high impact present unique characteristics in terms of how they are adopted by the network and that these qualities are independent of the event’s size and scope. As a consequence, high impact news events are naturally filtered by the social network, engaging users early on, much before they are brought to the mainstream audience.

Social media has become a primary source of breaking news information for millions of users all over the world (8). On-line social networks along with mobile internet devices have crowdsourced the task of disseminating real-time information. As a result, both news media and news consumers have become inundated with much more information than they can process. To handle this data overload, it is important to find ways to quickly filter information that has the potential of creating an impact on society. Fast identification of high-impact news events, at both

\textsuperscript{*}Corresponding author: jkalyana@ucsd.edu. JK and GL acknowledge the support of Yahoo Faculty Research Engagement Program.
global and local scales, can help improve information delivery, as well as journalistic coverage and emergency responses from authorities during crisis situations. We address this challenge by analyzing the properties of real-world news events in online social networks and present our main findings of how high-impact events can be clearly identified in the early stages of their outbreak.

The study of information propagation on the Web has sparked tremendous interest in recent years. The literature primarily considers studying the process through which a meme, usually a piece of media (like a video, an image, or a specific Web article), gains popularity (1, 3, 9, 11–13, 15, 17). Although that research is relevant, it faces two major shortcomings when it comes to the identification of impactful newsworthy events. The first limitation is that a meme represents a simple information unit and its propagation behavior does not correspond necessarily to that of more complex information such as news events. News events are usually diffused in the network in many different formats, e.g., a particular news story such as earthquake in Japan can be communicated through images, URLs, tweets, videos, etc. The second limitation of prior work comes from how the impact of information is measured. In scientific research, impact has been traditionally related to the total amount of attention received by a subject (2, 5–7, 10). Content in online social networks which received votes/comments/shares above a certain threshold is deemed viral or popular, and therefore having impact. One problem with this definition is that it does not consider events of different sizes and scopes, and is biased toward large events. For instance, a global event with world-wide coverage, such as the death of political leader Nelson Mandela, is big news because of the volume of social media posts produced by online social networks over a short period of time. Therefore, it is reasonable to consider this news as high-impact. Nevertheless, there are events that create high-impact in smaller online communities, such as a particular country (e.g., the withdrawal of the main right wing presidential candidate in Chile due to psychiatric problems, just before elections[1]). Clearly, local events do not produce as many social media posts as global events, but do create a very strong and immediate reaction from users in local networks, and should also be considered high-impact. Some other types of events, such as The Oscars, span a long period of time and are discussed by social network users for weeks or even months. This type of event, when considered in full, does generate a great volume of social media posts, but these posts are distributed over time without producing an immediate reaction from the network’s users. All in all, it is evident that a variety of aspects need to be considered when determining whether or not an event can be deemed high-impact. Indeed, when consulting journalists on how news media sources measure the impact of news, we learn that they too face the issue of not having a clear way to approach this problem.

We focus on high-impact news in social media, contributing by (i) defining a new concise way for measuring information impact, (ii) determining the existence of unique characteristics that differentiate how high-impact exogenous events are propagated in the network, and (iii) show, through the creation of an early prediction model for high-impact events, that these types of news events are naturally identified and filtered by the network at very early stages of their

Figure 1: **Illustrative examples of two events summarized by our method.** The event [nelson, mandela] was collected on 2013-12-05. Since there is a high concentration in the first histogram bin, we conclude that the social media posts for this event occur in cascades of quick successions, almost instantaneously. The second event, [may, oscar], was collected on 2014-03-23 discussing The Oscars event that was held a few weeks before. The arrival times of these posts are much more spread out. The y-axis is in square root scale.

We define an event as a conglomerate of information that encompasses all of the social media content related to a real-world news occurrence. Using this definition, which considers an event as a complex unit of information, we define an event’s impact in terms of the strength or immediacy of the networks’ reaction to the information about a particular news topic. We define an event’s impact on the community based on the *arrival time intervals* of consecutive social media posts. High-impact events are those events for which the *distribution* of the arrival time intervals between posts is most heavily skewed towards the smallest possible arrival time interval (zero). In other words, those events which have posts arriving almost instantaneously are defined as high-impact. Figure 1 shows an example of the distribution of the arrival time intervals of the events (a) *The death of Nelson Mandela* and (b) *The Oscars*.

We study a dataset of 5,234 news events gathered from news headlines from a *manually*
curated list of news media accounts (e.g., @CNN, @BreakingNews, @BBCNews, etc.) in the microblogging platform Twitter\(^2\) (a full list of all the news media accounts is provided in the supplementary material). Headlines were collected periodically, every hour, over the course of approximately one year. In parallel, all the Twitter posts (called tweets) were extracted about each news event using the public API\(^3\), based on a set of descriptive keywords extracted for each event. The dataset is composed of 43,256,261 tweets about events. The descriptive keywords were automatically extracted from the headlines using a variation of frequent itemset extraction (14). Figure 2 illustrates an example of an event showing the set of keywords and the tweets associated with the event. These keywords form a semantically meaningful event; they refer to the incident where soccer player Luis Suarez was charged for biting another player during the FIFA World Cup in 2014. Note that the collected set of social media posts associated to an event in this case is a more exhaustive collection, encompassing several memes, viral tweets etc. about the event when compared to existing studies (1, 3, 9, 11–13, 15, 17) which typically focus on simpler pieces of information (e.g., one particular meme, one viral tweet etc.).

The proposed model is based on the vector quantization of the lengths of time intervals for the entire collection of events. This approach is taken from the field of multimedia content analysis where it is known as codebook-based representation, and used, for example, in audio processing and computer vision (4, 16). After learning a codebook of the most representative time intervals, each event is represented as a histogram where each bin corresponds to one representative time interval. The entries of the resulting vector are obtained as the percentage of consecutive message pairs of the event that are assigned to the corresponding bin, based on their arrival times. This representation aptly and adequately captures the immediacy and intensity of the reaction to the overall event. We call this representation an impact-based event model. Using this model, we identify events as having a similar impact if their arrival time intervals are similarly distributed, as that would imply a similar intensity and immediacy of the overall reaction to the event. In particular, we perform clustering on all the event vectors to identify groups with similar distributions of their arrival time intervals. We sort the resulting groups of events from highest to lowest impact, according to the concentration of social media posts in the bins that correspond to short time intervals. We consider the events that fall in the top few clusters to be high-impact as their associated social media posts have the shortest arrival time intervals. These correspond to roughly 8% of the events. We consider the next few clusters in the sorted ranking to form medium-high impact events, and so on. Thus we end with four groups of event histograms: high, medium-high, medium-low and low. Note that this way separating the events is independent of the message volume and the duration of the event, allowing us to identify local and global high-impact occurrences.

Figures 3a and 3b show the average histograms of the events in the high and low impact clusters, respectively. Visually, high-impact event vector representations seem to follow an exponential distribution. We test this hypothesis by fitting an exponential function of the form

\[ f(t) = \lambda e^{-\lambda t} \]

\(\lambda\) denotes the parameter of the exponential distribution.
Luis Suarez’s bite is by no means the first controversial World Cup moment. Here are six of the most shocking:

- **Luis Suarez** faces disciplinary proceedings for biting Italy defender Giorgio Chiellini in World Cup - via @Telegraph
- Maradona, Zidane, now Suarez. Memorable WorldCup scandals, we forget any? (@geraldimrayAP)
- Norwegian wins bet on Suarez World Cup bite via @sharethis
- World Cup 2014: Why are so many goals being scored? - via @Telegraph
- Ghana’s government sends a plane carrying over $3m in cash to Brazil to pay appearance fees owed to World Cup players
- (Hey, soccer fans!!) River City hosts free World Cup viewing party
- Suarez risks World Cup ban as Fifa charges him with biting Italian defender - The Times of India
- Disgrace to the Beautiful Game: How social media chewed up Luis Suarez
- Luis suarez’s bite. The trolls have started rolling. This one is my favorite
- BREAKING: FIFA charges Uruguay’s Luis Suarez with biting. He faces a maximum two-year ban.

Figure 2: A representative event, collected on 2014-06-25 with keywords (left) and sample user posts (right) collected from the Twitter Search API. Collected user posts contain at least one pair of keywords.
Figure 3: Figure 3a and 3b show the average histogram of the high-impact and lowest impact groups respectively in our dataset. The y-axis is in square root scale.

$f(x) = ae^{bx}$ to the event histograms. Figure 3 shows two scatter plots with the resulting exponential parameters $a$ and $b$. We observe that the majority (97.4%) of high-impact events have an exponent $b \leq -50$, separating them unequivocally from other events.

Further analysis of the high-impact events shows significant differences to other events, in the following aspects: (i) how the information about these events is propagated, (ii) the characteristics of the conversations that they generate, and (iii) how focused users are on the news topic. In detail, high-impact events have a higher fraction of retweets (or shares) relative to their overall message volume. On average, a tweet from a high-impact event is retweeted 2.36 times more than a tweet from a low impact event. The most retweeted message in high-impact events is retweeted 7 times more than the most retweeted message in a medium or low impact event. We find that a small set of initial social media posts are propagated quickly and extensively through the network without any rephrasing by the user (plain forwarding). Intuitively, this seems justified given general topic urgency of high-impact events. Events that are not high-impact did not exhibit these characteristics.

Our investigation also revealed that high-impact events tend to spark more conversation
between users, 33.4% more than other events. This is reflected in the number of replies to social media posts. The number of different users that engage in high-impact events is 32.7% higher than in events that are not high-impact. Posts about high-impact events are much more topic focused than in other events. The vocabulary of unique words as well as hashtags used in high-impact events is much more narrow than for other events. Medium and low impact events have over 7 times more unique hashtags than high-impact events. This is intuitive, given that if a news item is sensational, people will seldom deviate from the main conversation topic.

In a real-world scenario, in order to predict if an early breaking news story will have a considerable impact in the social network, we will not have enough data to create its impact-based model, i.e., we will not know yet the distribution of the speed at which the social media posts will arrive for the event. For instance, an event can start slowly and later produce an explosive reaction, or start explosively and decay quickly to an overall slower message arrival rate. Still, reliable early prediction of very high-impact news is important in many aspects, from decisions of mass media information coverage, to natural disaster management, brand and political image monitoring, and so on.

We approach the early prediction task of high-impact events by using machine learning algorithms over the early features of events. We use event features that are independent of our impact-based model, such as the retweets, the sentiment of the posts about the event, etc. This prediction task is carried out independently of the number of social media posts about the event, and the type of event (local or global). In our experiments, using off-the-shelf logistic regression, we identify the high-impact events with 82% precision, using only the earliest 5% of the data of each event. Additionally, we were able to identify with high accuracy a considerable percentage of all high-impact events (≈ 46%) at an early stage, with very few false positives. There were a total of 5,234 events with 426 high impact events. We performed 5 fold cross-validation with randomly selected 60% training, 20% validation, and 20% test splits. We investigated the performance of several classifiers for this prediction task, and discuss about the results obtained from logistic regression.
The precision using only the early tweets is almost as much as using all tweets in the event (0.819 to 0.830). This suggests that the social network somehow acts as a natural filter in separating out the high-impact events fairly early on. The recall goes from 0.455 to 0.540. This indicates that there are some high-impact events which require more data in order to determine what kind of impact they will produce, or events for which impact occurs due to random conditions. A detailed description of the features and different classification settings are provided in the supplementary material.

We show that there are several properties that separate how high-impact news events evolve in Twitter in comparison to other events. We have created a model for events that allows us for an unambiguous classification of high-impact events based on their impact in the social network, in terms of the distribution of their inter-message arrival rates. This definition avoids problems with current notions of virality and popularity. Some characteristics of high-impact events are that they are forwarded more often by users, and generate a greater amount of conversation than other events. Social media posts from high-impact news events are much more focused on the news topic. Our experiments show that there are several properties that can suggest early on if an event will have high-impact on the on-line community. We can predict a high number of high-impact events before the network has shown any type of explosive reaction to them. This suggests that users are collectively quick at deciding whether an event is important or not. However, there does exist a fraction of events which will become high impact, despite not presenting patterns of other high impact events during their early stages. These events are likely to be affected by other factors, such as random conditions found in the social network at the moment and require further investigation.

References and Notes


