May 2018

The Nature of Twitter Trending Topics

Analysing Intrinsic Factors Associated with the Twitter Ecosystem

Marco A. PALOMINO [a,1], Quentin RIBAC [b,2] and Giovanni MASALA [a]
[a] *Big Data Group — School of Computing, Electronics and Mathematics*
University of Plymouth, United Kingdom
[b] *Département Informatique, IUT de Lannion*
Université de Rennes 1, France

Abstract. We are inundated with data—companies such as Twitter deal with petabytes of information on a daily basis. However, some users, especially the new ones, often find it difficult to cope with the overwhelming and disorganised deluge of information. Scientists have already worked out ways to identify Twitter trending topics, as a means to index information and make sense of it. However, we know little about the impact on trending topics of various intrinsic factors associated with the Twitter ecosystem. For example, anecdotal evidence suggests that trending topics are characterised by highly polarised tweets, or that large audiences typically host the emergence of trending topics. However, no study has yet addressed these issues formally. To remedy this situation, we have launched an investigation on the nature of trending topics. Our initial observations indicate that there is a correlation between strong sentiment polarity and the emergence of trending topics—we can also confirm that the strength of the polarity drops as the trending topics fade away. Conversely, our experiments highlight that there is no correlation between the size of a Twitter audience and the rise of trending topics.

Keywords. Twitter, trending topics, sentiment analysis, clustering, TF-IDF, news information retrieval

1. Introduction

Twitter [1] has become a real-time instrument to measure the impact of worldwide events. Indeed, Twitter has shown an extraordinary potential to monitor natural disasters—like wildfires [2] and earthquakes [3]—and also sports and entertainment events—like the *World Cup* [4] and the *Academy Awards*, which involved the publication of the message with the largest audience in Twitter's history [3] [5].

While monitoring Twitter posts has encouraged the development of multiple applications, analysing such posts is an increasingly challenging task. On average, around 6,000 posts—or *tweets*—are published on Twitter every second [6]. Hence, individual users are often sunk in the flood of information.

[1] Corresponding Author: Marco Palomino; E-mail: `marco.palomino@plymouth.ac.uk`.

[2] Work carried out while participating in the University of Plymouth's Undergraduate Internship Programme.

[3] One of the most famous celebrity pictures of recent times, "Ellen DeGeneres' Oscars selfie" is, officially, the most widely-spread message in Twitter's history. The informal snap published during the 86th Academy Awards in 2014 was redistributed over 2.6 million times in a span of 10 hours [5].

To index the massive amount of tweets constantly published and make sense of them, Twitter currently employs a proprietary algorithm to identify *trending topics*—i.e., themes that experience a surge in popularity. Such topics typically refer to terms and phrases that reflect current events—for instance, `World Cup`—and often include keywords extracted from popular conversations—for instance, `#Russia2018`.

While trending topics offer an alternative to cope with the information deluge, Twitter's algorithm is tailored for individual users—trending topics are based on the location of each specific user and the accounts that the user follows [7]—and do not contemplate the wider context.

Detecting trending topics is so valuable to journalists, e-marketing specialists, and social media researchers that we have decided to study the impact on trending topics of various intrinsic factors associated with the Twitter ecosystem. To be more precise, we are currently looking into the answers to the following questions:

What is the role of sentiment in trending topics? We wish to ascertain whether trending topics are characterised by a strong sentiment polarity. We expect that highly polarised tweets—i.e., very polemic, controversial or belligerent tweets—are likely to spark off wide discussion, which in turn encourages the support or opposition of large numbers of users and become part of a trend.

What is the relationship between trending topics and *likes*? "Likes" are used in Twitter to show appreciation for a tweet [7]. We expect some trending topics to be clusters of highly "liked" tweets. Of course, trending topics may also comprised highly criticised tweets. However, we believe that if a group of tweets is "liked" by a large number of users, then such a group will foster the emergence of a new trending topic.

What is the relationship between trending topics and the size of their audience? We conjecture that a large audience is indispensable for a trending topic. In other words, a small audience cannot host a trending topic.

What is the relationship between trending topics and their range of dissemination? We assume that trending topics are composed of greatly re-tweeted messages: if it is not retweeted, it cannot be a trending a topic.

What is the impact of multimedia content and associated URLs on trending topics? It might be the case that multimedia content makes some tweets more attractive and therefore more likely to reach larger audiences. It might also be the case that the occurrence of URLs in tweets fosters their dissemination.

To the best of our knowledge, no study has yet addressed the questions stated above. Casual observations and anecdotal evidence may support different answers to these questions. However, there is a need for a study that looks into the subject in depth. Such a study is the main motivation behind this paper.

It should be noted that we are not particularly interested in developing a new algorithm to detect trending topics—nor are we planning to refine an existing algorithm. Our emphasis is on understanding the impact of the factors highlighted above on the nature of trending topics. While our investigation could be carried out with the support of any existing system to detect trending topics, we have decided to produce our own system, simply to have full control of the implementation and modify it as required.

Our system is based on the ideas published by Benhardus and Kalita [8], and we have made it available to other researchers who wish to reproduce our experiments—our code is available at `https://github.com/ribacq/twitter-trending-topics`.

To test our topic detection system with realistic information, we have concentrated on the news domain. Although our work can be extended to other domains easily, we have chosen news for our experiments, because we have found it straightforward to validate the output of our trending topics system by comparing it with the headlines retrieved from a news API. If the trending topics that we identify are reported by well-regarded news sources, such as *BBC News* [9] or *Google News* [10], then we can confirm that our system has accurately discovered trending topics.

The remainder of this paper is organised as follows: We begin with an overview of related work. Then, we describe the dataset that we used for our experimentation. Afterwards, we describe our topic detection system and we use it to evaluate the impact on trending topics of the various factors listed above. Finally, we present our results and offer some conclusions.

2. Related Work

Twitter is a microblogging service that enables people to share their interests and opinions [11]. The word “tweet”—*short post*—has already entered our lexicon just as *Xerox* did for copying and *Google* for searching [12]. Twitter users can communicate with each other, with groups and with the public at large; thus, when Twitter conversations surface, they are often experienced by broader audiences than just the interlocutors. A comprehensive description of Twitter can be found in *The Twitter Book* [13].

As the body of research involving Twitter continues to grow, it has become clear that tweets contain plenty of useful information. However, the challenges for retrieving, storing and processing the colossal number of tweets continuously published keep mounting—it takes less than two days for one billion new tweets to be published [6].

An attempt to index Twitter content is characterised by *TweetMotif* [14], a service that summarises and groups publicly available tweets to discover trending topics, and provide an overview of what people are discussing. Although TweetMotif is no longer operational, the detection of Twitter trending topics has remained an active subject of research [15–19]. Benhardus and Jugal [8], for instance, used the *TF-IDF* information retrieval technique [20], combined with a number of heuristics, to assign weights to the different terms comprised in a tweet—the terms with the greatest weights were used to identify trending topics. We have also employed TF-IDF in our study. However, we are not interested in experimenting with different heuristics, because our current work is exclusively about measuring the impact of the factors stated above, rather than tuning the performance of a particular algorithm.

Shamma *et al.* [21] looked into ongoing temporal conversations to find *peaky* topics—topics that show highly localised, temporary interest. To mine the text across tweets, Shamma *et al.* employed a weighting approach similar to the TF-IDF. However, Shamma *et al.* did not consider any of the intrinsic factors of the Twitter ecosystem that we are pondering in this study.

We will now explain how the dataset that we used for our experimentation was retrieved, and we will describe its main features.

3. Dataset

As explained above, we used a news API to corroborate the accuracy of our trending topics discovery system. There are several news services APIs on offer—some of them correspond to specific media outlets, such as the *New York Times API* [22], or belong to specific software providers, such as the *Bing News Search API* [23]. However, we looked for an option that was free-of-charge and global in terms of its coverage. Hence, we chose *News API* [24], an API that supplies metadata for news headlines and articles published all over the Web. We do realise that other options might be suitable too, but we found News API particularly straightforward to integrate into our development.

All our experiments were carried out using a collection of tweets retrieved between May 8th 2018 at 12:07 BST and May 9th 2018 at 07:28 BST. We retrieved tweets published, exclusively, by the Twitter news sources listed in Table 1. We chose these sources, because they are English-speaking sources crawled by News API to retrieve breaking news headlines [24]. Thus, the tweets in the collection belong to the same domain as the headlines that we can obtain from News API[4].

Table 1. Twitter news sources monitored

@ABC	@abcnews	@AJEnglish	@arstechnica
@AssociatedPress	@AP	@FinancialReview	@axios
@AxiosWorld	@BBCNews	@BBCSport	@BleacherReport
@business	@BreibartNews	@businessinsider	@BIUK
@BuzzFeed	@CBCNews	@CBSNews	@CryptoCoinsNews
@MailOnline	@DailyMailUK	@DailyMailUS	@engadget
@EW	@espn	@financialpost	@FinancialTimes
@FT	@ftlive	@FortuneMagazine	@FourFourTwo
@FourFourTwoUSA	@FourFourTwoOz	@FoxNews	@foxnewsalert
@FOXSports	@googlenews	@Independent	@mashable
@mnt	@MetroUK	@DailyMirror	@MSNBC
@MTVNews	@mtvuknews	@NatGeo	@NRO
@NBCNews	@News24	@newscientist	@newscomauHQ
@Newsweek	@NYMag	@NFLNewsdesk	@politico
@Polygon	@Recode	@Reuters	@ReutersUK
@rte	@talkSPORT	@TechCrunch	@techradar
@amconmag	@TheEconomist	@globeandmail	@guardian
@GuardianAus	@thehill	@the_hindu	@HuffPost
@IrishTimes	@ladbible	@nytimes	@TheNextWeb
@sportbible	@Telegraph	@TelegraphNews	@timesofindia
@verge	@WSJ	@washingtonpost	@WashTimes
@TIME	@USATODAY	@vicenews	@WIRED
@WiredUK			

[4] The only source of news that is crawled by News API and we could not monitor is *Football Italia*—`https://www.football-italia.net/`. At the time of writing (May 2018), we could not find any verified Twitter user account for Football Italia.

The total number of tweets comprised in the collection that we used for our experiments is 510,000. We could have continued gathering tweets for longer, until retrieving a very large dataset. However, we wanted a collection that was manageable, so that we could easily reproduce our experiments and carry out our various comparisons without concerns for processing time, storage space and other practicalities.

To retrieve public tweets, we employed *python-twitter* [25], a Python-based wrapper that interacts with the *Twitter API* [26]. Given that we used the *Twitter Streaming API* [27], a stream listener retrieved the tweets published by the sources listed in Table 1 as soon as they were posted. Figure 1 displays the percentage of tweets published by the 30 most prolific sources in our dataset.

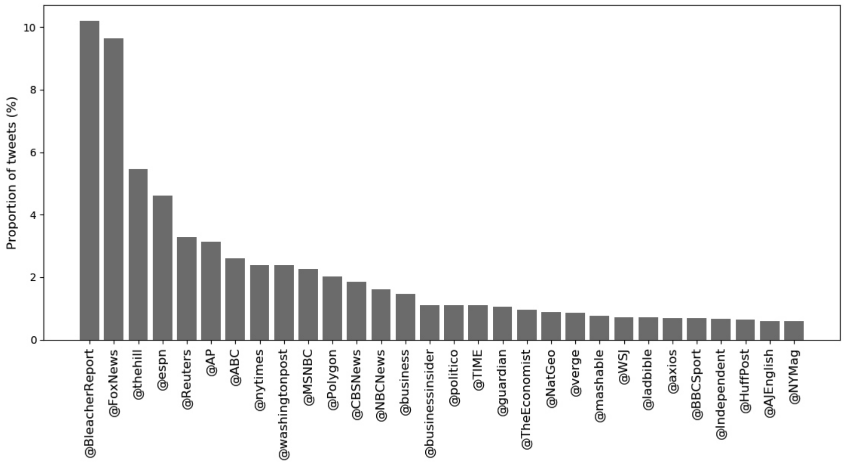

Figure 1. Percentage of tweets published by the 30 most prolific sources in our dataset

The streaming API provided the data in JSON format, and we wrote a Python parser to extract the tweets and other relevant metadata, such as the time when the tweets were published and the identifiers of the users who published those tweets. To store and manage the tweets that we collected, we uploaded them to a *MongoDB* collection [28] after retrieving them. This allowed us to keep, classify and manipulate separately specific subsets of tweets for our analysis.

4. Methods

The definition of trending topic that we used is a modified version of the one proposed by Benhardus and Kalita [8]: *a trending topic is a word, or combination of words, that experiences an increase in usage, both in relation to its long-term usage and in relation to the usage of other words*. Our system to identify trending topics is fully written in Python, which makes our implementation computationally economical, and easy to reproduce by other researchers. Figure 2 shows a diagram representing the sequence of steps that we follow from the start of the retrieval of tweets to the identification of trending topics and the determination of the sentiment expressed in the tweets. We will elaborate on each component of the diagram in the following subsections.

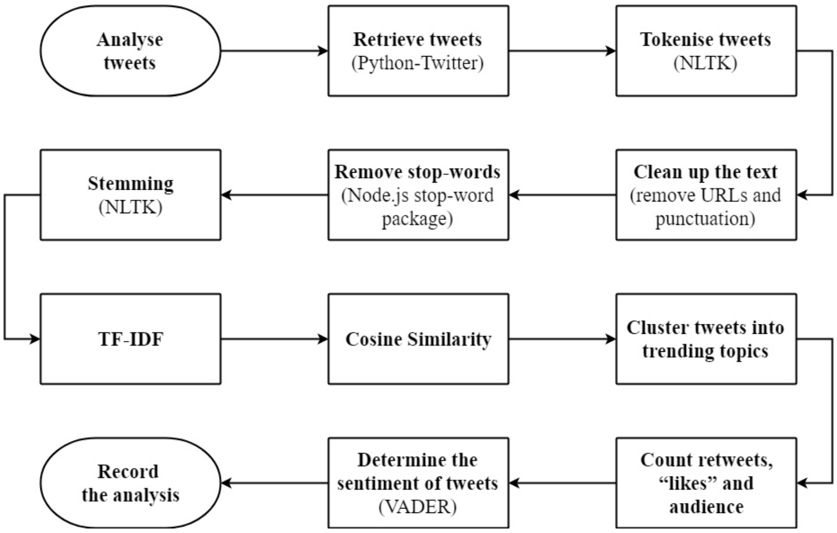
Analyse tweets
Retrieve tweets (Python-Twitter)
Tokenise tweets (NLTK)
Clean up the text (remove URLs and punctuation)
Remove stop-words (Node.js stop-word package)
Stemming (NLTK)
TF-IDF
Cosine Similarity
Cluster tweets into trending topics
Count retweets, "likes" and audience
Determine the sentiment of tweets (VADER)
Record the analysis

Figure 2. Tweet Analysis: Determination of trending topics and sentiment analysis

4.1. Tokenisation

Tokenisation is the process of splitting a piece of text into the different terms that compose it, while disposing of certain characters, such as punctuation [29]. Upon extracting the text of a tweet, we tokenised it, using the *Natural Language ToolKit* (NLTK) [30]. We chose the NLTK to perform this task, because it has become a standard suite for natural language processing (NLP) [31]. Hence, by using the NLTK, we are facilitating the reproducibility of our work.

Given that tweets, as any other form of online text, usually contain a great deal of "noise" and uninformative content [32]—such as shortened URLs, unicode characters and punctuation—we also took advantage of the tokenisation to clean up the text.

4.2. Stop-words

Stop-words are extremely common and semantically non-selective words [33], such as *the*, *is*, *at*, *which* and *on*. By removing these words, we guarantee that our analysis concentrates on "meaningful" terms within tweets.

To remove stop-words, we used the stop-word list provided by the *Node.js stop-words package* [34], which contains 661 terms. We have noticed that the Node.js stop-word list is a superset of the list built by Salton and Buckley for the experimental *SMART* information retrieval system [35].

4.3. Stemming

Upon removing stop-words, we *stemmed* the text of each tweet. The goal of stemming is to reduce inflectional forms and derivationally related forms of a word to a common base form [29]. The most popular algorithm for stemming English words, and one that has repeatedly shown to be effective, is *Porter's algorithm* [36]. This is precisely the stemming algorithm that we chose.

Evidently, stemming reduces the number of terms contained in our collection of tweets, by merging some words together. Our empirical evidence suggests that stemming reduces the amount of processing time involved in determining trending topics.

We stem tweets using the NLTK implementation of the *Porter's stemmer* [37]. The reason why we chose this particular implementation of the Porter's stemmer is, once again, that the NLTK has become a standard for the development of NLP code—thus, its use favours the reproducibility of our work.

4.4. TF-IDF

TF-IDF weights a document's relevance to a query based on a composite of the query's *term frequency* and *inverse document frequency* [20]. For the purpose of our work with tweets, *term frequency* can be defined as

$$tf_{ij} = n_{ij}$$

where $n_{i,j}$ is the number of times word i occurs in tweet j and

$$\sum_k n_{kj}$$

is the total number of words in tweet j. *Inverse document frequency* is defined as

$$idf_i = \log \frac{T}{d_i}$$

where d_i is the number of tweets that contain word i and T is the total number of tweets. Roughly speaking, the weight of a tweet will be higher if the number of times a word occurs in a tweet is higher, or if the number of tweets containing that word is lower; similarly, the weight of a tweet will be lower if the number of times a word occurs in a tweet is lower, or if the number of tweets containing that word is higher [38].

4.5. Cosine Similarity

To cluster the tweets into trending topics, we calculate the distance between every new tweet in the dataset and all the trending topics identified so far. The distance is estimated through the *cosine similarity* [29], a technique to measure cohesion within clusters. If the distance between a tweet and a trending topic is smaller than 0.3, the tweet is added to the topic; otherwise, a new topic with this tweet as its only element is formed. Clusters comprising less than 5 tweets are not considered trending topics.

Trending topics are identified by the terms in the cluster with the eight greatest TF-IDF weights. Examples of the tweets included in our dataset and their classification within the identified trending topics are listed in Table 2. Note that Table 2 also indicates the number of tweets per topic, the terms that identify the topics, the size of the audience for each particular topic, and the number of times that each tweet was retweeted.

Table 2. Examples of trending topics and their associated tweets

Topic 1:	288 tweets in total
Terms:	north china met Xi presid korean leader kim
Audience:	82,731,608 users
Tweets:	(237x)>>> RT @AP: BREAKING: China reports President Xi Jinping has met with North Korean leader Kim Jong Un in northern China. (22x)>>> RT @ABC: JUST IN: Chinese Pres. Xi Jinping has met with North Korean leader Kim Jong Un in northern China. https://t.co/t5iNbCCIhw (19x)>>> RT @AP: The Latest: China says President Xi Jinping met with North Korean leader Kim Jong Un in a northern Chinese port city, the second m...
Topic 2:	616 tweets in total
Terms:	trump presid iran deal nuclear withdraw french told
Audience:	131,073,530 users
Tweets:	(594x)>>> RT @nytimes: Breaking News: President Trump will withdraw from the Iran nuclear deal, he told the French president, fulfilling a campaign v... (13x)>>> RT @HuffPost: BREAKING: Trump reportedly tells French President Macron that he's withdrawing from the Iran nuclear deal... (3x)>>> RT @Telegraph: Donald Trump reportedly told French president Emmanuel Macron he will withdraw from the Iran nuclear deal and...

Figure 3 plots the number of headlines produced by News API every 30 minutes within the length of our data collection. Figure 3 also plots the number of trending topics that we identified every 30 minutes. To identify trending topics, we consider the intersection between *all* the terms in our clusters of tweets and the headlines retrieved from News API. If the number of terms in the intersection is equal to, or greater than, 40% of the terms in the headlines, we consider our cluster a trending topic.

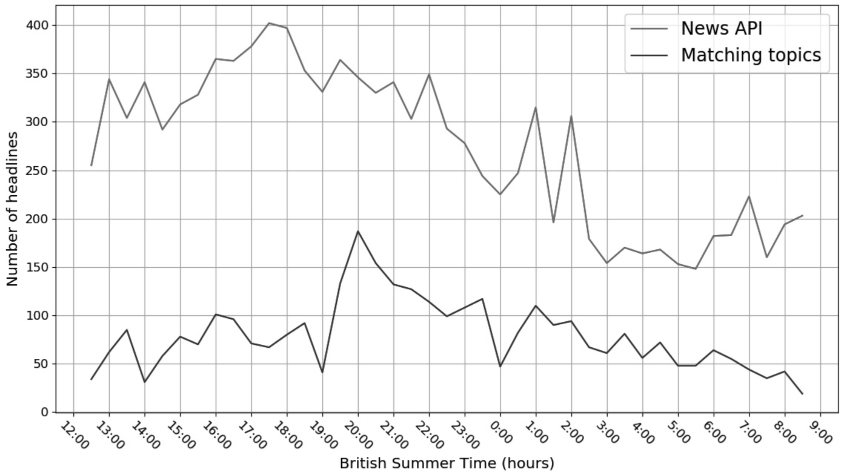

Figure 3. Number of trending topics matching actual news headlines

As it can be seen in Figure 3, our system missed a great deal of headlines, though this is not necessarily because our system requires some tuning. The truth is that several headlines never become trending topics in Twitter. For instance, during the aftermath of the Manchester Attack on May 22nd 2017, *Channel Four Television* gathered, and continuously updated, a large number of headlines—readers may want to see: `https://www.channel4.com/news/topic/manchester-attack`. While some of these headlines became trending topics in Twitter, others were overlooked by Twitter audiences—for example, the news referring to the MI5 review procedures for dealing with warnings did not spark a surge of activity in Twitter. In any case, producing a fully-optimised system to discover trending topics is not the goal of our work. Our focus is on the nature of the actual trending topics; thus, the remainder of this paper concentrates, exclusively, on the clusters of topics that match news headlines.

5. Results

We provide below a description of the results yielded by the analysis of our dataset. We begin by discussing our work on *sentiment analysis*.

5.1. Sentiment Analysis

Sentiment analysis is the process of computationally identifying and categorising opinions expressed in a piece of text [39]. The most basic task in sentiment analysis is classifying the polarity of a given opinion—i.e., determining whether an opinion expressed towards a particular subject is *positive*, *negative* or *neutral* [40].

To determine the polarity of the tweets in our dataset, we used *VADER* [41]—*Valence Aware Dictionary and sEntiment Reasoner*—a rule-based tool created to identify sentiments expressed in social media. Figure 4 shows the proportion of positive, negative and neutral tweets in our dataset according to VADER. The statistics were apprised and updated every 30 minutes, from the start of our retrieval of tweets to the end.

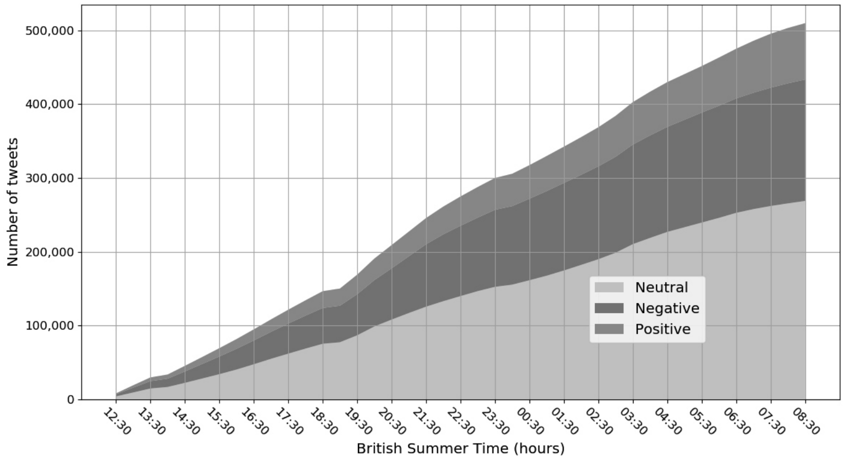

Figure 4. Number of positive, negative and neutral tweets in the collection estimated by VADER

Figure 5 compares the sentiment expressed by tweets that are in trending topics with the sentiment expressed by tweets that are not in trending topics. As it can be seen in Figure 5, tweets that are not in trending topics are largely neutral, whereas tweets that are included in the trending topics identified in our dataset are largely negative. This confirms our original conjecture that trending topics are characterised by a strong sentiment polarity—either positive or negative.

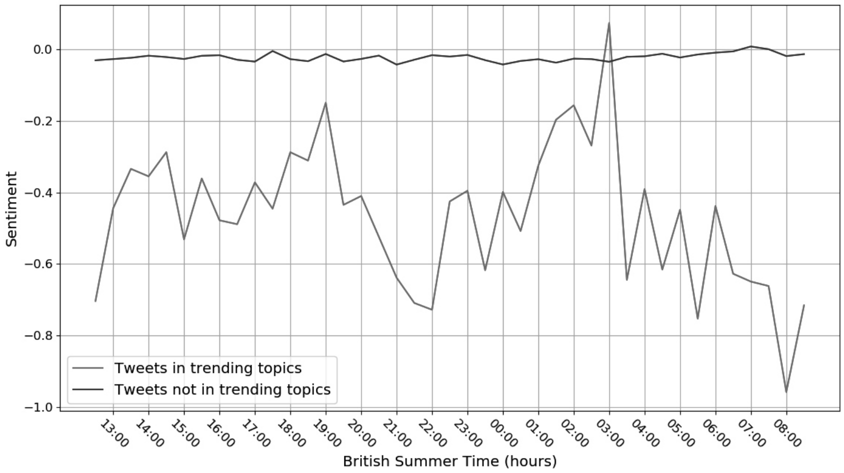

Figure 5. Sentiment polarity expressed in tweets comprised in trending topics

Figure 6 displays five trending topics that lasted for longer than a 30-minute period in our dataset. The plot in Figure 6 shows how the sentiment expressed in these five topics evolved over time. For example, the blue topic—characterised by the terms: `stupid`, `plan`, `trump`—remained negative during the entire period monitored; yet, it reached its highest polarity strength on May 8th 2018 at 19:00 and 21:00.

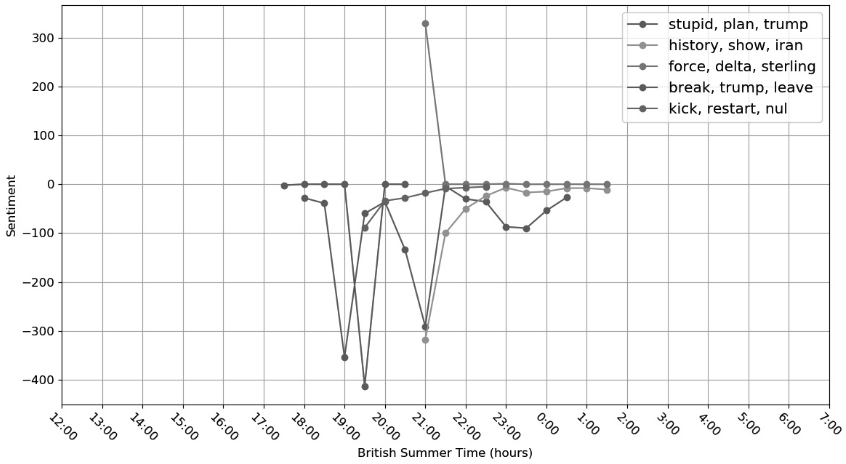

Figure 6. Evolution of polarity strength in trending topics over time

Note that the sentiment of the trending topics becomes neutral, or close to neutral, prior to their disappearance. For example, the red trending topic—characterised by the terms: `break`, `trump`, `leave`—became very negative on May 8th 2018 at 19:30, but had a very neutral polarity on the same day at 21:30, when we were able to track it for the last time. Indeed, it seems that the strength of the polarity of all our trending topics approaches neutrality as they fade away.

5.2. Audience

We define the *size of the audience* of a tweet as the total number of followers that the user that published the tweet had at the time the tweet was published. Similarly, the *size of the audience of a trending topic* is defined as the total number of followers that the users that published the tweets, or retweets, comprised in the trending topic have at the time the trending topic is formed. Figure 7 shows that the audience of the tweets in the trending topics that we identified in our dataset is smaller than the audience of tweets that were not part of a trending topic.

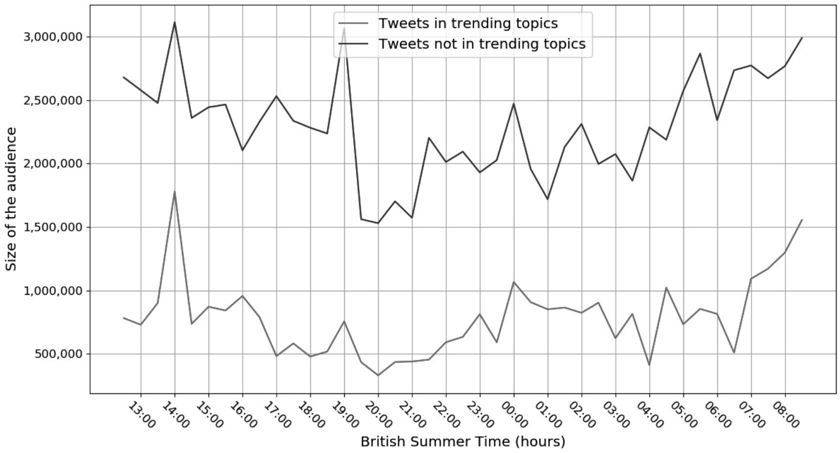

Figure 7. The size of the audience of tweets in trending topics

Certain values, such as the number of retweets for a particular tweet, or the number of "likes" that a tweet has received, cannot be determined immediately after retrieving the tweets—at the time of the retrieval, some tweets that we captured in our dataset had never been retweeted, because they had just been published. Obviously, the longer we wait after the publication of a tweet, the more retweets it may have. Hence, a week after the original retrieval, we retrieved, for the second time, every single tweet that we collected between May 8th 2018 at 12:07 BST and May 9th 2018 at 07:28 BST. The second time we retrieved these tweets we were able to know how many times they had been retweeted in the span of a week, and how many times they had been "liked". Had we waited for longer than a week, some of our tweets might have been retweeted or "liked" more times. However, trending topics are so temporary that variations that take place more than a week after their publication were not considered relevant.

5.3. URLs

Figure 8 compares the number of URLs encountered in tweets that are part of trending topics with the number of URLs in tweets that are not part of trending topics. As it can be seen, there are more URLs outside the trending topics than within them. Hence, the presence of URLs in tweets do not foster the emergence of trending topics.

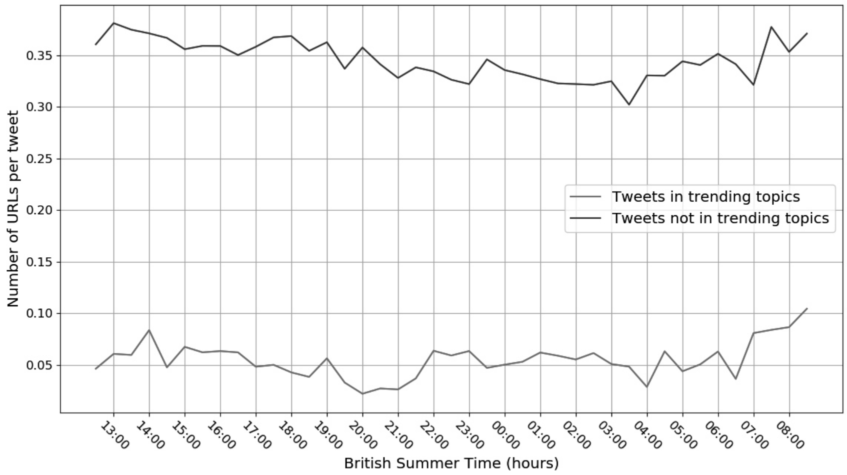

Figure 8. Number of URLs present in tweets comprised in trending topics

6. Conclusions

The need for a study on the nature of trending topics in Twitter has been evidenced. The available literature does not document in sufficient depth the features that characterise trending topics, their emergence and evolution over time. In this paper, we have produced preliminary results concerning the analysis of the impact on trending topics of some intrinsic factors associated with the Twitter ecosystem. Such results reveal that highly polarised tweets foster the development of trending topics. Moreover, we can confirm that existing trending topics tend to disappear from Twitter when the strength of their polarity declines.

Unexpectedly, our results show that large audiences are not indispensable for a trending topic—at least, this was not the case for our dataset. Similarly, the numbers of URLs present in tweets is not related to the incidence of trending topics. We had thought that URLs would attract audiences and, in turn, spark off the discussion that could lead to the formation of trends. However, this does not seem to be the case in our dataset. On the contrary, large numbers of URLs seem to drive tweets away from trending topics.

We expect to continue our investigation and include other domains of information in our research too—focusing on news headlines is sensible as a starting point. However, other domains, such as politics, might point out further insights on the subject.

References

[1] Twitter Inc. (2018, Mar.) Twitter. [Online]. Available: https://twitter.com/

[2] Z. Wang, X. Ye, and M.-H. Tsou, "Spatial, Temporal, and Content Analysis of Twitter for Wildfire Hazards," *Natural Hazards*, vol. 83, no. 1, pp. 523–540, 2016.

[3] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors," in *Proceedings of the 19th International Conference on World Wide Web*. ACM, 2010, pp. 851–860.

[4] Y. Yu and X. Wang, "World Cup 2014 in the Twitter World: A Big Data Analysis of Sentiments in US Sports Fans Tweets," *Computers in Human Behavior*, vol. 48, pp. 392–400, 2015.

[5] Lynch, Kevin. (2018) 10 Years of Twitter: Five Key Tweets that Made Record-Breaking History. [Online]. Available: http://www.guinnessworldrecords.com/news/2016/3/10-years-of-twitter-five-key-tweets-that-made-record-breaking-history-421461

[6] Internet Live Stats. (2018) Twitter Usage Statistics. [Online]. Available: http://www.internetlivestats.com/twitter-statistics/

[7] Twitter Inc. (2018, Mar.) Twitter Help Center — How Can We Help? [Online]. Available: https://help.twitter.com/en

[8] J. Benhardus and J. Kalita, "Streaming Trend Detection in Twitter," *International Journal of Web Based Communities*, vol. 9, no. 1, pp. 122–139, 2013.

[9] BBC. (2018) BBC News. [Online]. Available: http://www.bbc.co.uk/news

[10] Google. (2018) Google News. [Online]. Available: https://news.google.com/

[11] D. Boyd, S. Golder, and G. Lotan, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter," in *Proceedings of the 43rd Hawaii International Conference on System Sciences (HICSS)*. Honolulu, HI: IEEE, Mar. 2010, pp. 1–10.

[12] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter Power: Tweets as Electronic Word of Mouth," *Journal of the Association for Information Science and Technology*, vol. 60, no. 11, pp. 2169–2188, 2009.

[13] S. Milstein and T. O'Reilly, *The Twitter Book*. O'Reilly Media, May 2009.

[14] B. O'Connor, M. Krieger, and D. Ahn, "TweetMotif: Exploratory Search and Topic Summarization for Twitter," in *Proceedings of the 4th International AAAI Conference on Web and Social Media*, Washington, DC, 2010, pp. 384–385.

[15] H. Becker, M. Naaman, and L. Gravano, "Beyond Trending Topics: Real-World Event Identification on Twitter," *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, vol. 11, no. 2011, pp. 438–441, 2011.

[16] S. Petrović, M. Osborne, and V. Lavrenko, "Using Paraphrases for Improving First Story Detection in News and Twitter," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2012, pp. 338–346.

[17] S. Petrovic, M. Osborne, R. McCreadie, C. Macdonald, I. Ounis, and L. Shrimpton, "Can Twitter Replace Newswire for Breaking News?" in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Boston, MA, 2013.

[18] A. Guille and C. Favre, "Event detection, tracking, and visualization in twitter: A mention-anomaly-based approach," *Social Network Analysis and Mining*, vol. 5, no. 1, p. 18, 2015.

[19] S. Liang and M. de Rijke, "Burst-Aware Data Fusion For Microblog Search," *Information Processing & Management*, vol. 51, no. 2, pp. 89–113, 2015.

[20] G. Salton and D. Harman, *Information Retrieval*. John Wiley and Sons Ltd., 2003.

[21] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Peaks and Persistence: Modeling the Shape of Microblog Conversations," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. ACM, 2011, pp. 355–358.

[22] NYTimes.com. (2018) The New York Times Developer Network. [Online]. Available: https://developer.nytimes.com/

[23] Microsoft Azure. (2018) Bing News Search. [Online]. Available: https://azure.microsoft.com/en-gb/services/cognitive-services/bing-news-search-api/

[24] News API. (2018) News API - Access Worldwide News with Code. [Online]. Available: https://newsapi.org/

[25] Mike Taylor. (2018) Python-Twitter. [Online]. Available: https://github.com/bear/python-twitter

[26] Twitter Inc. (2018, Mar.) API Reference Index. [Online]. Available: https://developer.twitter.com/en/docs/api-reference-index

[27] ——. (2018, Mar.) Filter Realtime Tweets. [Online]. Available: https://developer.twitter.com/en/docs/tweets/filter-realtime/overview

[28] S. Agrawal, J. P. Verma, B. Mahidhariya, N. Patel, and A. Patel, "Survey on MongoDB: An Open-Source Document Database," *Database*, vol. 1, no. 2, p. 4, 2015.

[29] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to Information Retrieval*. Cambridge University Press, 2008, vol. 39.

[30] NLTK 3.3. (2018) NLTK.Tokenize Package. [Online]. Available: http://www.nltk.org/api/nltk.tokenize.html#module-nltk.tokenize.casual

[31] J. Perkins, *Python 3 Text Processing with NLTK 3 Cookbook*. Packt Publishing Ltd, 2014.

[32] E. Haddi, X. Liu, and Y. Shi, "The Role of Text Pre-Processing in Sentiment Analysis," *Procedia Computer Science*, vol. 17, pp. 26–32, 2013.

[33] W. J. Wilbur and K. Sirotkin, "The automatic identification of stop words," *Journal of information science*, vol. 18, no. 1, pp. 45–55, 1992.

[34] Huned Botee. (2018) Node.js Stopwords Package. [Online]. Available: https://github.com/huned/node-stopwords

[35] G. Salton, "A New Comparison between Conventional Indexing (MEDLARS) and Automatic Text Processing (SMART)," *Journal of the Association for Information Science and Technology*, vol. 23, no. 2, pp. 75–84, 1972.

[36] M. F. Porter, "An Algorithm for Suffix Stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[37] NLTK 3.3. (2018) NLTK.Stem.Porter Module. [Online]. Available: http://www.nltk.org/api/nltk.stem.html#module-nltk.stem.porter

[38] D. Hiemstra, "A Probabilistic Justification for Using TF× IDF Term Weighting in Information Retrieval," *International Journal on Digital Libraries*, vol. 3, no. 2, pp. 131–139, 2000.

[39] R. Feldman, "Techniques and Applications for Sentiment Analysis," *Communications of the ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[40] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment Classification Using Machine Learning Techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2002, pp. 79–86.

[41] C. J. Hutto and E. Gilbert, "Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," in *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM-14)*. Ann Arbor, MI: The AAAI Press, 2014.