What was Hillary Clinton doing in Katy, Texas?

Toni Gruetze   Ralf Krestel   Konstantina Lazaridou   Felix Naumann

Hasso Plattner Institute
Prof.-Dr.-Helmert-Straße 2-3
14482 Potsdam, Germany
{firstname.lastname}@hpi.de

ABSTRACT

During the last presidential election in the United States, Twitter drew a lot of attention. This is because many leading persons and organizations, such as U.S. president Donald J. Trump, showed a strong affection to this medium. In this work we neglect the political contents and opinions shared on Twitter and focus on the question: Can we determine and track the physical location of presidential candidates based on posts in the Twittersphere?

Keywords

Politician; Location; Tracking; Twitter; US Election

1. MINING TWITTER

Whilst reading this work, millions of people publish content of various types in different social networks, such as Facebook and Twitter. These networks enable users to share their current thoughts and experiences as they happen, facilitate an agile spreading among their peer group and allow faster reactions of other users. In particular, due to the public visibility of most messages, Twitter is said to "break down the communication barriers" [5]. This yields an opportunity to mine and analyze the content of the shared thoughts and experiences for specific use cases.

Mining Twitter for the location of prominent people like politicians is challenging. One has to distinguish between tweets containing relevant insights, i.e., mentions of politicians or locations, and the large amount of tweets with irrelevant and misleading information. Moreover, it is difficult to analyze tweets in isolation, because the messages lack adequate contextual information due to their length restriction. The recipient is usually familiar with the context, since it is consumed only in a small time window.

In contrast to conventional mass media, Twitter does not enforce any editing rules or guidelines for the content. Both important news as well as minor events are discussed and the information is spread as they happen, like the tweet of user @joexhunt 'I drove by #TMCC and was like ”Why is @KOLO8 here?” Oh, yeah. Our next prez #Hillary in #Reno today.' Thus, Twitter can be seen as a fast, decentralized, anarchic news media [4].

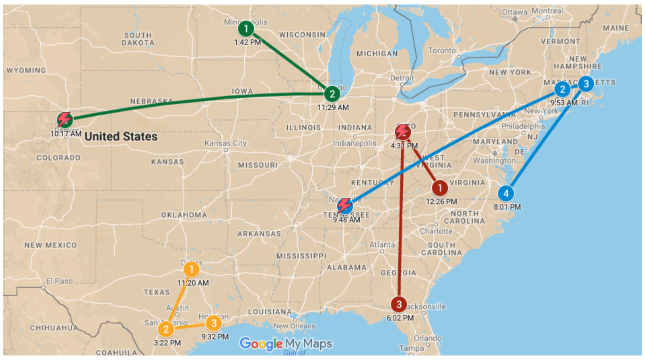

Figure 1: Trails of four presidential candidates extracted from Twitter on February 29th 2016

Previous research dealt with the topical comparison of newspapers and tweets [8], the extension of articles with additional facts and opinions from microblog posts [1], and the clustering of news stories according to the geographic locations of the sharing users [5]. Furthermore, tweets were used to estimate the user's location [2] and harvested for collaboratively collected geospatial information [7].

In this paper, we present an approach to track the location of U.S. politicians prior and up to the presidential election by crowd sourcing the Twittersphere. In particular, we discuss the following three steps in more detail: the retrieval of tweets about the U.S. presidential election 2016, the detection of politicians and locations mentioned within the tweets, and the collaborative reasoning based on the vast amount of statements shared in the Twittersphere. The resulting dataset containing the detected locations of politicians during the election campaign, all project resources, and visualizations can be found on our project home page.[†]

2. U.S. POLITICIANS ON TWITTER

For this work, we collected a set of over 770M tweets by more than 25M users mentioning candidates and other persons relevant for the U.S. presidential election during the 13-month period starting on November 2015. Due to the rate limits of the Public API of Twitter, we had to ensure that we retrieve as many relevant tweets as possible, without

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3041021.3054231

[†] https://hpi.de/naumann/projects/us-election-on-twitter

exceeding the 1% of the Twitter traffic. The retrieved tweet set is based on a carefully selected list of 241 queries containing politician names, their Twitter user aliases, and hashtags related to the U.S. election campaign. For instance, query terms like Ben for the resigned candidate Ben Carson are not appropriate, because they yield too many irrelevant tweets. You can find the dataset and a detailed format description on our home page.†

3. ENTITY MENTIONS

In order to identify tweets containing information about a specific candidate, we analyze the message text. Similarly to the well-known entity linking problem [3], the task is to identify the textual mentions and disambiguate among different potential entities (e.g., Hillary and Bill Clinton). Due to the short tweet length and the consequential lack of context, we do not apply sophisticated coherence reasoning strategies. Instead, we used a list of reliable entity aliases retrieved from Wikipedia link anchor texts to minimize the chances of false positively detected entity mentions. This was achieved, by scoring the aliases according to the commonness (also surface prominence) in Wikipedia [3] and retaining all entries with a score above 0.5. We further extend aliases by 50 manually collected Twitter profiles of the corresponding relevant politicians (e.g., @realDonaldTrump).

Moreover, the geolocalization of tweets is a difficult problem. Previous research showed that 99% of the tweets do not include geotags [6]. To geolocalize the tweets containing candidate mentions, we focus on spatial indicators in the tweet text. We limit the granularity of the detection to cities found as Wikidata entries, which resulted in over 25k cities. Similarly to the politicians, the location aliases are based on Wikipedia link anchor texts. To extend the bare city aliases by sub-locations (e.g., quarters or points of interest located within a city), we group further geolocations according to the cities they are located in, based on a polygonal overlay approach [6]. For instance, the Empire State Building is an alias for the City of New York. We limit these sub-locations to entities that can be assigned to exactly one city. By implication, we ignore geolocations like federal states or geographical plains that span over several cities.

4. SO, WHAT WAS HILLARY DOING IN KATY, TEXAS?

Nothing, actually! Due to our basic approach for entity detection, we mistakenly identify tweets mentioning "Katy", as evidence that Hillary is in the city of Katy, Texas. However, the tweets were actually referring to Katy Perry, a prominent supporter of the Clinton campaign. The resulting set of candidate-location pairs from tweets also includes further misleading information. For instance, the reference to Benghazi indicates Clinton's involvement in the attacks of 2012, but not her actual location.

To tackle these issues, the semantics of a tweet have to be classified to match a '*is current location of*'-pattern. We consider the contextual information given in the tweet (i.e., the text without the entity mentions) as a bag of words. To identify the context of '*is current location of*'-pattern matches, we initially use the set of Hillary Clinton's campaign event locations for the last week in January 2016 (i.e., prior to the Iowa caucuses). We retrieved them from the official campaign web site. Based on the tweet contexts mentioning these locations, we perform the well-known Apriori algorithm to retrieve frequent item sets of context tokens. These frequent term sets are then filtered based on the relative frequency among all locations of Hillary Clinton found in tweets of this period. The remaining frequent term sets are subsequently used to remove irrelevant tweets. Therefore, the final tweets can be interpreted as crowdsourced indicators for the politicians' locations. The actual time of an event is estimated based on the tweet publication times. For each day, we consider all tweets of a candidate-location pair and calculate the median of tweet times.

Figure 1 depicts the automatically retrieved trail of Donald J. Trump (red), Hillary Rodham Clinton (blue), Bernie Sanders (green), and Ted Cruz (yellow) two days prior to the Super Tuesday on Google Maps. The order is based on the estimated event time. Marked with a red flash are incorrectly recognized locations. The two errors of the democratic trails (Nashville and Fort Collins) stem from campaign events from the previous day (February 28th), whereas Donald Trump actually visited Columbus one day later.

5. CONCLUSION

In this work, we presented an approach to detect the current location of prominent persons like candidates of the U.S. presidential election by crowdsourcing the Twittersphere. While targeting a challenging problem, the introduced approach shows promising results. In future work, we will investigate on how to improve the detection quality based on different constraints (e.g., persons cannot travel with the speed of light). Furthermore, we will examine whether the methodology is also applicable to other domains or not (tracking of popstars, sports teams, religious figures, etc.).

Acknowledgments. This research was partly funded by the German Research Foundation, DFG grant no. FOR 1306.

6. REFERENCES

[1] X. Cao, K. Chen, R. Long, G. Zheng, and Y. Yu. News comments generation via mining microblogs. In WWW, pages 471–472, 2012.

[2] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: A content-based approach to geo-locating twitter users. In CIKM, pages 759–768, 2010.

[3] T. Gruetze, G. Kasneci, Z. Zuo, and F. Naumann. CohEEL: Coherent and efficient named entity linking through random walks. JWS, 37(C):75–89, 3 2016.

[4] M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. Ma. Breaking news on Twitter. In SIGCHI, pages 2751–2754, 2012.

[5] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: News in tweets. In SIGSPATIAL, pages 42–51, 2009.

[6] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. Mühlhäuser. A multi-indicator approach for geolocalization of tweets. In ICWSM, 2013.

[7] A. Stefanidis, A. Crooks, and J. Radzikowski. Harvesting ambient geospatial information from social media feeds. GeoJournal, 78(2):319–338, 2013.

[8] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing Twitter and traditional media using topic models. In ECIR, pages 338–349, 2011.