Constructing Twitter Datasets using Signals for Event Detection Evaluation

Constructing Twitter Datasets using Signals for Event Detection Evaluation

Hugo Hromic and Conor Hayes

Insight Centre for Data Analytics
National University of Ireland Galway (NUIG)
{hugo.hromic,conor.hayes}@insight-centre.org
http://www.insight-centre.org

Abstract. Twitter is a very attractive real-time platform for research on event detection. However, despite the great amount of interest, datasets suitable for evaluating such methods are not easily available. The two most important reasons for this are Twitter's strict Terms and Conditions for data distribution and the vast amount of Tweets data generated at every minute. In this paper we show a first exploration of a signal processing method suitable for generating datasets for event detection evaluation. Our proposal is based on the notion of ADSR (attack-decay-sustain-release) envelopes commonly used in acoustics signals modelling and applied to Twitter dynamics such as hashtags usage. We show preliminary results over real-world data that support this idea and the potential of our method for the event detection task itself.

Keywords: Annotated Twitter data, Event detection evaluation, Signal processing, ADSR modelling

1 Introduction

Social Media services are widely integrated into many aspects of our modern digital lives. They developed from simple blogs into complex multi-functional platforms like Facebook, Twitter, or Youtube. One particular class of those services are *microblogging* sites, which focuses on broadcasting short messages from users among their friends and/or followers. Microblogging is specially interesting because of its fast and real-time nature for information delivery. Today, Twitter is possibly the most widely known and used platform for microblogging in the world, with more than 230 million monthly active users as of September 2013, generating a massive average of 500 million *Tweets* (short messages) per day. Moreover an impressive peak record of 143,199 Tweets per second was achieved on August 3, 2013.[1]

Twitter is also an attractive use case for different research topics, specially for event detection. Generally an event is understood as something happening

[1] http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/

at a certain place and time, however there is no clear agreement in the literature about a specific definition [1]. For example a set of topics can be first discovered using LDA (Latent Dirichlet Allocation) —a topic modelling approach— and then those topics measured during different points in time for detecting changes (considered as events) according to the words distributions computed by the model [2]. In another example, earthquakes (seen as one particular kind of events) can be monitored using a content-based method and estimate the location of those natural disasters very promptly using spatio temporal associations [3]. Research on event detection for Twitter is broad and most works do not use the exact same definition for what they consider an event [4–10].

On the other hand, datasets for testing and evaluating event detection are not easily available or not well suited to compare different techniques. This is in part because of Twitter's strict Terms and Conditions for redistribution that prevent researchers from sharing their datasets openly.[2] The vast volume of data being generated constantly also contributes to the problem because building datasets turns into a difficult and expensive task. All of the above produce a strong focus on *intrinsic* evaluation —how well the method performs in terms of own measures, for example an specific interpretation of Precision and Recall— instead of *extrinsic* —comparing the approach to other alternatives—. Without standard datasets and a clear definition of *event*, it is hard to compare between available systems. A dataset created to evaluate a content-based approach might not have the required properties for a structural-based approach or even what two systems are targeting for as an *event* may be totally different.

The above problems have no clear solution in literature. However, few recent works have proposed the idea of collecting raw data and then using this data for generating or *synthesizing* datasets for event detection evaluation [4, 1]. One of these provide a solid base for building suitable datasets, including a definition of *event* that agrees with as many other works as possible [1]. The authors built a large 120 million Tweets dataset comprising approximately 150,000 annotated Tweets that relate to about 500 annotated events curated by humans using a crowd-sourced methodology. The work makes an important contribution towards building datasets for event detection evaluation, however there is still the lack of flexibility such as easily and quickly fine-tune the built dataset, i.e. different signal-to-noise ratios to evaluate detection sensitivity and/or robustness.

We propose an initial dataset generation method based on modelling Tweets dynamics as *signals* from different features. Preliminarily we focused on Twitter *hashtags* usage.[3] Our method borrows the concept of ADSR (attack-decay-sustain-release) envelopes widely used in the music domain for modelling musical instrument acoustics [11–14]. The ADSR model uses four parameters: initial time of attack, decay time, sustain level and a release time (see Figure 1).[4] The same concept can be applied to Twitter hashtags usage signals: rising in certain time (attack), falling for another time (decay), maybe keeping at some level (sustain)

[2] `https://dev.twitter.com/terms/api-terms`
[3] In Twitter, a hashtag is a user provided tag of what the Tweet is about.
[4] Figure from `http://www.dragonflyalley.com/constructionJHtrapezoidVCA.htm`

and eventually diminishing for a period until no longer used (release). Any of the those parameters could be also set to zero, indicating the absence of any of the phases (i.e. instant attack, no decay or instant release). ADSR envelopes are very flexible and diverse variations exist such as the addition of an extra *hold time* parameter between attack and decay (referred as AHDSR).

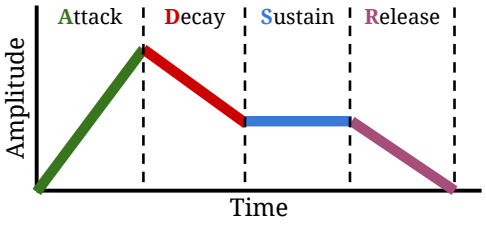


Fig. 1. Classic ADSR envelope shape. Four parameters are used to model almost any instrument tonal sound: (A)ttack time, (D)ecay time, (S)ustain Level and (R)elease time. More parameters can be found on more advanced variations. In our proposal, this can model Twitter hashtag dynamics during events.

The fundamental idea for our method is to separate interesting hashtags from noisy ones by curve-fitting individual hashtag usage signals into ADSR envelope parameters. Our **hypothesis** is that we can distinguish fitted ADSR parameters for event driven hashtags from those used for noisy topics, i.e. hashtags used for spamming or mundane everyday events. For the definition of an event, we decided on the following: *"a significant thing that happens at some specific time and place"* [1]. In this definition, *significant* means that the event may be discussed in traditional media such as news articles or sports reports. Fitting ADSR parameters from signals have been studied before, where a state of the art method for fitting from sound signals is proposed based on dynamic time warping [15].

We believe that an abstract model for hashtags usage dynamics based on ADSR envelopes parameters fitting is an important contribution. It is possible to construct sets of event hashtags differentiated from noise to be used as selectors for event Tweets. In this manner, fine-tuning of desired proportions of events and noise signals can be done to construct specific synthetic datasets.

In this paper we first present current existing approaches for Twitter datasets construction aimed for event detection evaluation, then we describe our proposed alternative for synthesizing Twitter datasets using signal analysis and raw Twitter data, following are the preliminary results we obtained with our method and finally we present conclusions and future directions for our work.

2 Related Work

Constructing Twitter datasets for evaluating event detection approaches has been done in almost every work related to this task, however most of the datasets are not available or can not be easily used into other approaches. Despite this,

more general datasets building methodologies have been proposed that could be used across different event detection approaches.

One work proposed a content-based event detection system that uses an LDA model to discover topics at regular intervals from a stream of Tweets [4]. Jensen-Shannon divergence is used over the found topics at each interval to detect changes (events). For evaluation, a synthetic dataset generation approach was also proposed. It uses real-world Tweets together with manually selected sets of hashtags designed to match both real events and noisy topics occurring during the capture time span. To minimise the effects of this manually (and possibly biased) selection, all hashtags are removed from the Tweets. Moreover for Tweets that are of events the text is replaced entirely with short text from news articles with similar topics as the hashtags used.[5] This procedure makes the LDA-based event detection unaware of the original hashtags used for creating the dataset. Despite of this methodology being adequate for text-based event detection, it is not suitable for approaches of different nature such as structural-based or burst-based methods that may not consider the text content of the Tweets at all.

The growing need for a comparable dataset inspired a construction methodology specifically designed for evaluation of event detection systems [1]. This approach uses two state of the art event detection methods (one based on Locality Sensitive Hashing (LSH) and the other on Cluster Summarization) over raw data complemented with input from Wikipedia's Current Events Portal to generate a pool of candidate events and their associated Tweets. Afterwards, this pool is merged and curated by humans using Amazon's Mechanical Turk (AMT) crowd-sourcing platform. Furthermore, this work recognises the need for a common definition of what is understood as an *event* based on current literature (see Section 1). Despite the advantages proposed, the generated datasets are static in nature and any further changes in their properties would require again input from the crowd-sourcing component, which unfortunately is not free or straightforward to apply.

3 Finding Hashtags using ADSR Envelopes

The task of generating synthetic datasets suitable for event detection evaluation can be defined in terms of the following core subtasks: (i) capture a collection of real-world Tweets to be used as input, (ii) build a pool of candidate events including their associated Tweets, (iii) build a pool of background Tweets to be used as noise and (iv) select Tweets from the above pools to be used in the final synthesized dataset.

For the first subtask usually the Twitter Streaming API is used.[6] This API offers real-time Tweets being published in Twitter. The second and third subtasks are the core of this section, and form the basis of our method. The fourth subtask can be done as shown in Section 3.3.

[5] The Topic Detection and Tracking (TDT) news articles corpus is used.

[6] `https://dev.twitter.com/docs/api/streaming`

3.1 Modelling Twitter Dynamics using Signals

The first step for our events finding model is to describe Twitter dynamics in terms of signals that can be further processed. Signals can be constructed by accounting for some features that change in time. In the case of a Tweets stream, there are plenty of features available that can be used: number of Tweets from a user, number of Tweets using a hashtag, number of mentions received by a user, number of references to a URL, number of references to named entities (prior to entity recognition), number of Retweets (references to Tweets from users) for all the above items, and more.

We adopted the work in [16] as a starting point for features. An approach for event detection is proposed by solely analysing Tweets and Retweets volumes for specific hashtags during certain world-wide events. Despite the simplicity of this concept, the authors demonstrated how effective it can be for prompt recognition of live popular events. The results motivated us to decide on using the following stream features for initial exploration: *Number of Tweets* (**NT**), *Number of Retweets* (**NRT**) and *Number of Replies* (**NR**) for each seen hashtag. Another work has used also words frequencies to build wavelet signals [7].

For all the features, we aggregated the counts into minute-sized bins to keep a reasonable amount of data points without overwhelming the ADSR parameter estimation. Example resulting signals can be seen in Section 4.

3.2 Estimating ADSR envelopes from Signals

We now need a method for separating event hashtags signals from background noise ones. For this we propose the usage of ADSR envelopes to profile and separate them accordingly. As mentioned, ADSR envelopes can model signals using four parameters: attack time, decay time, sustain level and release time. In the case of Twitter signals, hashtags that describe events must start from a certain state and then rise until an upper bound is met (attack phase). After this, either they start decaying at some rate or stay active constantly for some amount of time (decay and/or sustain phases). Finally, the hashtags dynamics must decay until they are not used any more (release phase). The ADSR parameters do not need to be the same for all events, for this reason it should be possible to model any hashtag usage signal by certain combinations of ADSR values.

The challenge then resides on estimating ADSR parameters that best fit hashtags signals. The approach in [15] provides a good ground. In their work, the authors investigated techniques for analysing tin whistle sound signals for estimating ADSR parameters from them. They examine two commonly used methods and propose a further novel method for more accurately estimating parameters using dynamic time warping (DTW). DTW is an alignment method for corresponding points in two time series (in this case signals and candidate ADSR envelopes). Their approach can be computationally expensive due to its combinatorial nature, however the authors also proposed a quantization method to reduce the complexity of the algorithm while maintaining reasonable accuracy.

A problem left to consider is the case when hashtags have a long running signal. In this situation, hashtags dynamics can rise again during the decay, sustain or release phases. When this happens, the ADSR envelope should be re-triggered to start from the attack phase again. This condition is not handled by the parameter estimation approach above, so it must be treated separately. This paper does not provide a solution for this issue, but ideas are proposed in Section 5 to further investigate possible approaches for re-triggering detection.

3.3 Selecting Tweets for the Synthetic Dataset

As we mentioned earlier, the last subtask for constructing a dataset is the selection of Tweets from the events and the background noise pools. The selected Tweets will be the annotated documents to be placed in the final dataset. The hypothesis for our proposal serves as foundation for this subtask.

We assume that hashtags dynamics modelled as signals can be used to estimate parameters for ADSR envelopes. After the parameters are estimated we could then differentiate event hashtags from background noise ones because both kind of hashtags may have clearly defined ADSR envelope shapes. To find evidence for the above idea, we performed a preliminary exploration over real-world Tweets datasets as described in Section 4.

As an example consider the case of a football match. Goal events can produce hashtags with a sharp attack time, then decay mildly to be sustained for a small period of time until the goal hashtags are quickly no longer used until the next goal. On the other hand, massive public events such as the Pope Election in the Vatican can have different dynamics. When the smoke used to indicate the result of the election is released this provokes again a sharp attack time, but the sustain and release phases can be much longer compared to the football match because of longer discussions held afterwards, specially if the smoke was white.

4 Preliminary Results

For our exploratory analysis we chose to use real-world Tweets from the Twitter Streaming API. This API offers three endpoints for capturing data: Firehose (access to all Tweets), Filter (access using provided matching criteria) and Sample (access to a 1% random sample of all Tweets).

Capturing from Firehose is restricted to only certain Twitter business partners, hence we captured data using the other two available endpoints. Because the statistical significance of them is questioned in the literature [17], we decided to capture from the Filter and Sample endpoints at the same time and compare the results. We obtained data for the Pope Conclave 2013 event held in the Vatican City between March 12 and March 14, 2013. Details of the datasets can be seen in Table 1. For the Filter endpoint, we used the hashtag *#conclave*, the keywords *conclave 2013* and a list of recommended users to follow by specialised media as matching criteria.[7]

[7] `http://expandedramblings.com/index.php/how-to-follow-the-conclave-to-select-the-next-pope-on-a-mobile-device/`

	FILTER	SAMPLE
Tweets	460,334	9,904,068
Users	277,121	6,571,390
Hashtags	21,072	414,776

Table 1. Datasets captured (and named) from the FILTER and SAMPLE endpoints.

First, we extracted the number of Tweets (**NT**), Retweets (**NRT**) and Replies (**NR**) for every appearing hashtag in both datasets in minute-sized bins. Then we built individual time series data per hashtag using normalised counts (we are interested in signal shapes instead of global amplitudes). We analysed the distribution of observations over time for each hashtag and found that they closely follow a power-law curve. Because we want to study signals with enough data points, we decided to filter out all hashtags that had less than 60 observations (approximately one hour of data). Please note that this does not guarantee consecutive observations. After filtering, 251 hashtags ($\approx 1.2\%$) remained from the FILTER dataset and 2896 ($\approx 0.7\%$) from the SAMPLE dataset. This result highlights the well-known usage sparsity of hashtags in Twitter.

As an overall view for the Conclave event, we plot the signals for the *#conclave* hashtag in Figure 2. Five major sub-events can be identified visually: the *conclave start*, the *first* and *second black smokes* (no pope elected), the *white smoke* (pope elected) and the *pope revelation*.

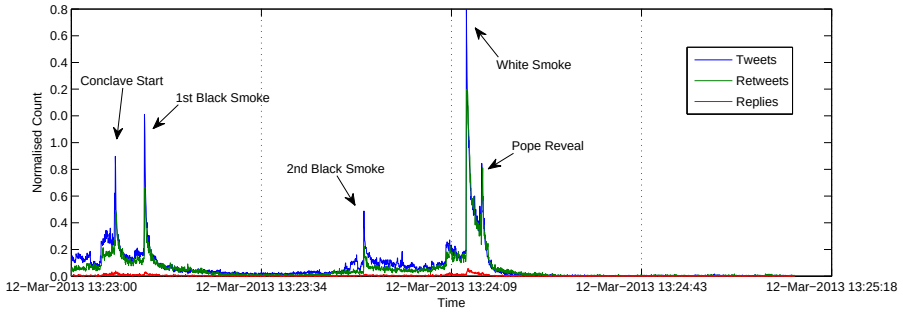
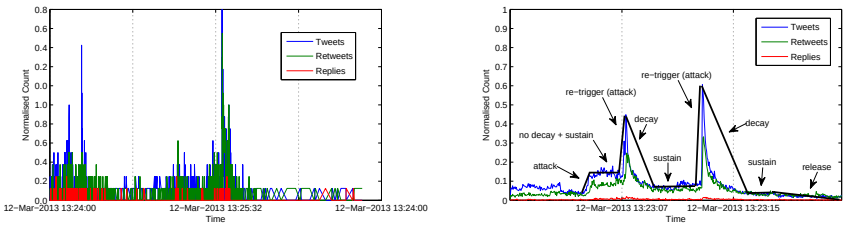


Fig. 2. NT, NRT and NR signals obtained for the hashtag *#conclave* in the FILTER dataset using normalised counts. Five sub-events can be visually identified.

This plot also shows that the three feature signals follow more or less the same trend, however in this (and the majority of the hashtags studied) the amplitude for NR does not provide enough dynamics. Moreover, the NT and NRT signals follow each other very closely (also seen in other hashtags). We also compared the signals for *#conclave* against the SAMPLE dataset with results shown in Figure 3(a). The 1% random sampling has clear effects on the result with similar signal shapes but at a much lower resolution. This suggests that the usage of the SAMPLE endpoint may not be well suited for signal analysis.

We are now interested in the feasibility of finding ADSR envelope shapes that would help in separating event hashtags from background noise. For this we explored the sub-events of the Pope Conclave more closely. In Figure 3(b) a

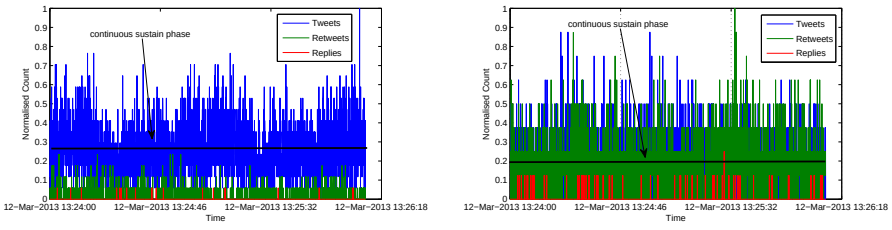


(a) In the SAMPLE dataset (b) Close-up with ADSR envelopes

Fig. 3. Different views for the signals extracted from the *#conclave* hashtag.

close-up of the *conclave start* and the *first black smoke* sub-events with ADSR envelopes superimposed is seen. For these particular kind of events the ADSR parameters result as follows: a fast attack phase, a mild decay phase, a long standing sustain phase and no visible release phase in the time window. It can be noted that after the second re-trigger the decay phase does not fit very well to the signal. This could be improved using ADSR envelopes that use polynomial curves instead of linear. The above result does not mean there will be no release phase. It is safe to assume that sooner or later the *#conclave* hashtag will be not used —or used very sparingly— in following hours as the commotion for the sub-events dissipates. For the *second black smoke*, the *white smoke* and the *pope revelation* sub-events the ADSR curves seem similar, suggesting that re-triggering is in effect multiple times after each interesting stage of the Conclave.

The FILTER dataset contains little noise compared to the SAMPLE one because of the filtering terms used for capturing. For this reason we decided to explore noisy hashtags from the SAMPLE dataset. In Figure 4 example signals for the hashtags *#android* (related to the mobile operating system) and *#followback* (a common practice done by Twitter users to gather attention and gain followers) can be respectively seen. These hashtags have a very distinctive constant/steady dynamics, dominated respectively by Tweets and Retweets.



(a) *#android* hashtag (b) *#followback* hashtag

Fig. 4. Signals for noise hashtags in SAMPLE. A constant trend is clear and can be modelled using ADSR using carefully chosen parameters.

Many other hashtags were found in SAMPLE with the same signal shape. Despite of the constant amplitude of the signals, plain dynamics can also be modelled using ADSR envelopes. For example consider the settings of nearly

instant attack and decay phases ($time \approx 0$), and a never ending sustain phase with a level approximate to the average observed NT, NRT or NR. Additionally it can be seen that the NT and NRT signals gave complementary non-overlapping curves, suggesting that they could be used independently for different kinds of signals. Some events can be triggered by a strong Tweets ADSR envelope, and others instead by strong Retweets dynamics.

All observations found during the exploration of the captured datasets suggest the viability of understanding Twitter hashtags dynamics —and possibly other Twitter objects such as users or URLs— using ADSR parameters estimated from signals created from stream features. As mentioned, if hashtags can be profiled through their fitted ADSR parameters it would be possible to classify them to separate event from noise hashtags.

5 Conclusions and Future Directions

In this paper we discussed the difficulty of evaluating different approaches for event detection in Twitter. The main challenges are Twitter's T&C agreements and also the huge amount of real-world data that makes almost impossible to manually build an annotated dataset. We proposed the usage of hashtags dynamics modelled as signals and curve fitting of these signals into ADSR parameters to address the task of separating event from noise Tweets. Our preliminary results suggest that our hypothesis is plausible and that curve fitting of ADSR envelopes may be a promising method to separate hashtags dynamics according to their ADSR parameters.

5.1 Future Directions

Future work mostly lies on expanding the features that can be extracted from Tweets to build different kinds of signals. Also we believe there may be other sources beneath hashtags, for example users being mentioned or retweeted. For the identified re-triggering detection problem we think it could be handled by using finite state machines with rules for the ADSR phases. We also believe our method can go beyond synthetic datasets generation and into the event detection task itself. Because there are always the same number of ADSR parameters, machine learning or clustering algorithms could be easily and efficiently used to classify event hashtags. Finally, curve fitting in a streaming fashion can be also explored instead of static historic datasets. This comes with interesting challenges such as scalability and signal processing without look-ahead.

6 Acknowledgements

This work was supported by Science Foundation Ireland (SFI) partly under Grant No. 08/SRC/I1407 (Clique: Graph and Network Analysis Cluster) and partly under Grant No. 12/RC/2289 (Insight Centre for Data Analytics).

References

1. Andrew J McMinn, Yashar Moshfeghi, and Joemon M Jose. Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 409–418. ACM, 2013.
2. Qiming Diao, Jing Jiang, Feida Zhu, and Ee-Peng Lim. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 536–544. Association for Computational Linguistics, 2012.
3. 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*, pages 851–860. ACM, 2010.
4. Jey Han Lau, Nigel Collier, and Timothy Baldwin. On-line trend analysis with topic models:\# twitter trends detection topic model online. In *COLING*, volume 12, pages 1519–1534, 2012.
5. Saša Petrović, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189. Association for Computational Linguistics, 2010.
6. Charu C Aggarwal and Karthik Subbian. Event detection in social streams. In *SDM*, volume 12, pages 624–635. SIAM, 2012.
7. Jianshu Weng and Bu-Sung Lee. Event detection in twitter. *ICWSM*, 11:401–408, 2011.
8. Saša Petrović, Miles Osborne, and Victor 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*, pages 338–346. Association for Computational Linguistics, 2012.
9. Hila Becker, Dan Iter, Mor Naaman, and Luis Gravano. Identifying content for planned events across social media sites. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 533–542. ACM, 2012.
10. Qiankun Zhao, Prasenjit Mitra, and Bi Chen. Temporal and information flow based event detection from social text streams. In *AAAI*, volume 7, pages 1501–1506, 2007.
11. Kristoffer Jensen. *Timbre models of musical sounds*. PhD thesis, Department of Computer Science, University of Copenhagen, 1999.
12. Marko Helen and Tuomas Virtanen. Perceptually motivated parametric representation for harmonic sounds for data compression purposes. In *Proc. DAFx*, 2003.
13. John M Grey and John W Gordon. Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America*, 63(5):1493–1500, 1978.
14. Stephen McAdams, James W Beauchamp, and Suzanna Meneguzzi. Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. *The Journal of the Acoustical Society of America*, 105(2):882–897, 1999.
15. J Timoney, T Lysaght, L Mac Manus, and A Schwarzbacher. Dynamic time warping for tin whistle partial envelope. In *submitted to Irish Signals and Systems conference, Belfast, Northern Ireland*, 2004.

16. Flavio Chierichetti, Jon Kleinberg, Ravi Kumar, Mohammad Mahdian, and Sandeep Pandey. Event detection via communication pattern analysis. In *Proceedings of the Eight International AAAI Conference on Weblogs and Social Media*, 2014.
17. Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *ICWSM*, 2013.