FTC 2016 - Future Technologies Conference 2016
6-7 December 2016 | San Francisco, United States

A Framework for Twitter Events Detection, Differentiation and its Application for Retail Brands

Olga Kolchyna
CS Department, University College London
Email: v.kolchyna@cs.ucl.ac.uk

Thársis T. P. Souza
CS Department, University College London
Email: tharsis.souza.14@ucl.ac.uk

Philip C. Treleaven
CS Department, University College London
Systemic Risk Centre,
London School of Economics
Email: p.treleaven@ucl.ac.uk

Tomaso Aste
CS Department, University College London
Systemic Risk Centre,
London School of Economics
Email: t.aste@ucl.ac.uk

***Abstract*—We propose a framework for Twitter events detection, differentiation and quantification of their significance for predicting spikes in sales. In previous approaches, the differentiation between Twitter events has mainly been done based on spatial, temporal or topic information. We suggest a novel approach that performs clustering of Twitter events based on their shapes (taking into account growth and relaxation signatures). Our study provides empirical evidence that through events differentiation based on their shape one can clearly identify clusters of Twitter events that contain more information about future sales than the non-clustered Twitter signal. We also propose a method for automatic identification of the optimum event window, solving a task of window selection, which is a common problem in the event study field. The framework described in this paper was tested on a large-scale dataset of 150 million Tweets and sales data of 75 brands, and can be applied to the analysis of time series from other domains.**

***Keywords*—Anomaly detection; clustering; event detection; event study; spikes; social media; Twitter**

I. Introduction

In the last decade social media have extended their application beyond their original domain changing the way people communicate, share ideas and opinions. Due to the large amount of information posted online it has become possible to track how people react to different world events in real time. In this context, detection of events in social-media has become an attractive problem in data mining. Event detection approaches have been proposed in the last few years [1]–[5]. For example, [2] suggested a multiscale event detection method, which takes into account temporal and spatial scales of events in social-media and [3] proposed to extract events from social data by representing textual data in form of sequences of numeric values.

As was highlighted by [6], the problem of event detection is closely related to the problem of event clustering. For example, identification of different types of news events can shed light on the problem of detecting the most predictive events [7]–[9]. The distinction between the different types of events in these studies was performed based on the temporal information and the topic that was discussed: initial public offerings, or earnings announcements, or stock splits. In the context of hugely popular social media networks, it is important to distinguish different types of events not only based on their content, spatial and temporal information, but to take into account the dynamic of how information spreads through social networks. For example, a discussion on Twitter would evolve differently through time depending on whether it was initiated by the brand through a marketing campaign or came as a result of a word-of-mouth information sharing. A post-event effect on sales might also be different for the two scenarios. Studying these internal dynamics is a challenging task and requires understanding the rules of human collective behaviour. Progress in this direction was achieved by [10]–[13] who were able to get insights into the nature of events in blogosphere activities [10], views of Youtube videos [11], Amazon books sales [12], [13] by studying the growth and relaxation signatures of those events.

In this paper we incorporate the knowledge about the internal dynamics of social-media events by introducing a framework that allows to automatically detect events and cluster them based on their growth and relaxation signatures. We used this framework for the analysis of a large-scale dataset that contained daily sales figures for 75 brands from the retail sector (supplied by Certona Inc.) and daily Twitter sentiment time series for the same brands. The objective of the analysis is to detect events in sales and in Twitter sentiment, and measure whether Twitter sentiment events could be used to predict events in sales. To the best of our knowledge, there are no extensive studies that measure the impact of a specific social-media event on sales of a company. The only study that mentions evaluation of Twitter events in relation to sales is the study by Dijkman et al. [14], however their analysis is limited to just one company and 12,780 tweets.

We demonstrate that clustering of events based on their shapes serves as a filter of Twitter signal by applying our clustering algorithm to Twitter sentiment events and revealing classes of Twitter events that contain more information about future sales than the non-clustered Twitter signal.

978-1-5090-4171-8/16/$31.00 ©2016 IEEE

In order to measure the significance of a Twitter event it is necessary to identify the event date and set the event window within which the analysis is performed ([15]), however, the necessary step of defining an event window is challenging [16]–[18]. To solve this problem, in our framework we proposed not to specify event window explicitly, but to analyse the predictive power of Twitter along a wide range of windows using cumulative probabilities. Our statistical approach allowed us to automatically identify the windows during which Twitter events had significant power to predict spikes in sales.

The main contributions of this paper are as follows:

- Defined a framework for automatic events detection, events differentiation and evaluation of their importance;
- Proposed a novel method for events clustering based on their growth and relaxation signatures;
- Extended the event study field by proposing a method for automatic identification of the optimal event window;
- Performed a large-scale application of proposed framework to retail brands;

II. Proposed Framework

In this study, we propose a framework for events detection, differentiation and evaluation of whether events in one time series can be used to predict events in the other time series. The framework consists of three steps: (A) events detection; (B) events clustering; (C) quantification of events significance. The suggested framework can be used for analysis of any kind of time series with the following necessary conditions: 1) time series should correspond to the same time period; 2) time series should have the same aggregation window.

In this paper we demonstrate the application of the framework to retail brands using the following datasets:

1) Daily Sales time series provided by Certona company, related to 75 brands over the period of one year, from November 1, 2013, to October 31, 2014. Data was normalised using a z-score [19] to make it comparable across different brands.
2) Daily Twitter Sentiment time series that cover the same time period. The dataset includes sentiments of more than 150 million tweets that mention the names of selected brands. Sentiment analysis of Twitter messages was performed using our own tool [20], which has shown improved performance compared to known benchmarks as well as has been successfully applied in other works [21]. The outcome of the algorithm is a label assigned to each Twitter message: positive, negative or neutral. Daily Twitter Sentiment time series were calculated as a ratio of number of positive messages to a number of negative messages in a day.

A. Events Detection

Definition: An event is a quantitatively significant change of behaviour of a dynamic phenomenon over time. In this paper, a Twitter event is an anomalous uplift of sentiment on Twitter and sales event is an extreme increase in volume of sales.

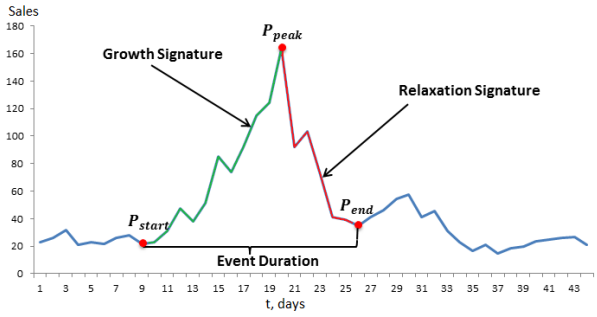

Fig. 1. Signatures of a sales event. The green color represents the growth signature of the event, the red color represents the relaxation signature and the red circle is the peak of the event.

Each event can be characterised by its duration (from few hours to few days), the growth signature, peak and the relaxation signature. In Figure 1, P_{start} denotes the start of the event, P_{peak} denotes the peak of the event, P_{end} represents the end date of the event; the subset of data points between P_{start} and P_{peak} is a growth signature, the subset between P_{peak} and P_{end} defines the relaxation signature. In our framework the process of identifying the event and its corresponding signatures consists of two steps: 1) peak detection; 2) extraction of growth and relaxation signatures.

1) Peak Detection: As the first step of the events detection process we identify peaks in data that indicate anomalous behaviour. For this purpose we compare the performance of three outlier detection methods:

1) **Extreme Studentized Deviate Test (ESD identifier)** [22]. For a sample $x_N = \{x_i\}_{i=1}^{N}$ it classifies any point more than t standard deviations away from the mean to be an outlier, where the threshold value t is most commonly taken to be 3. In other words, x is identified as an outlier if:

$$|x - X| \geq t\sigma \quad (1)$$

where X is the mean and σ is estimated standard deviation of the data sequence.

2) **Hampel identifier** [23], [24]. For this outlier detection method, the mean is replaced with median of the residuals and the standard deviation with the median absolute deviation estimate (MAD). MAD is a robust measure of the variability of univariate data. To compute MAD, one calculates the median of the absolute deviations of each historical value from the data's median.

$$MAD(x_N) = \overline{X}(|x_1 - \overline{X_N}|, ..., |x_N - \overline{X_N}|) \quad (2)$$

where $\overline{X}$ is the median. x is identified as an outlier if:

$$|x - \overline{X_N}| \geq g(N, \alpha_N)MAD(x_N) \quad (3)$$

Where g is a function related to the number of data points and a specified type I error (see [25], [26]).

3) **Median and InterQuartile Range (IQR)** [27]. For this outlier detection method, one calculates the 25th percentile and the 75th percentile of the data. The difference between the 25th and 75th percentile is the interquartile deviation IQR. The historical value x is classified as an outlier if it is outside of the closed range:

$$[Q_1 - K * IQR; Q_3 + K * IQR] \quad (4)$$

where $IQR = Q_3 - Q_1$, Q_1 and Q_3 are the 25th and the 75th percentiles respectively, and K is often selected equal to 1.5.

When performing peaks detection it is important to consider that Sales and Twitter time series are non-stationary. We observe weekly patterns in data, for example, tweets volume on Fridays and weekends is much higher than during the other days of the week. If the volume of Friday's tweets was compared to the volume of Thursday's tweets, the peak detection method would show a spike on Friday, however we want to detect only special events and not regular bursts. To account for non-stationarity, the three outlier detection measures described above (ESD/Hampel/IQR) are computed from the observations within a moving window that is comprised of data points from the same day of the week. For example, if the data point of primary interest is Friday, a moving window will include the data point of primary interest and the K prior Friday values. In this way, Friday values will only be compared to the K previous Fridays, Saturdays will be compared to K previous Saturdays and so on.

2) Extraction of Growth and Relaxation Signatures: For each peak P_{peak_i} that was detected during the first step, the goal is to identify the data point at which the event starts P_{start_i} and the data point at which the event finishes P_{end_i}.

Let us define the points at which the time series change its direction as change-points, and the time intervals between consecutive change-points as time segments. In Figure 2 points $C = C_1, C_2, ..., C_9$ represent change-points. To extract the growth signature, the immediate left neighbouring point of P_{peak_i} is analysed to determine whether the point is a change-point. The procedure is repeated with consecutive left neighbours until the stopping criterion is met. The first change-point on the left side from P_{peak_i} that meets the stopping criterion is considered to be the start point of the event P_{start_i}. To extract the relaxation signature, an identical procedure is performed with right neighbours of P_{peak_i}. The first change-point on the right side from P_{peak_i} that meets the stopping criterion is considered to be the end point of the event P_{end_i}.

The stopping criterion is considered to be met if any one of the following three conditions is fulfilled:

1) The first condition is fulfilled if the distance between the current change-point C_k and the peak P_{peak_i} exceeds the maximum distance D_{max}, predefined by the user. This condition alows to limit the duration of the event. Formally,

$$(C_k - P_{peak_i}) > D_{max} \quad (5)$$

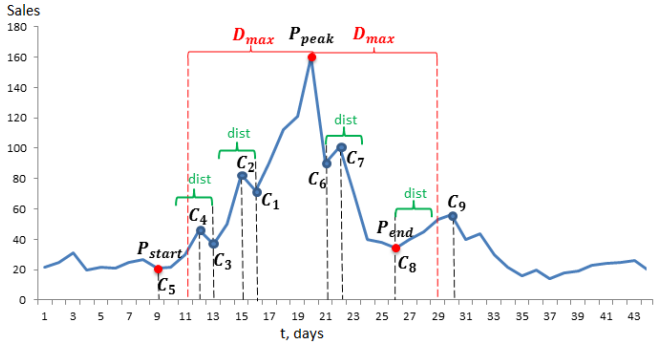

Fig. 2. An example of a sales event. $C = C_1, C_2, ..., C_9$ denote change-points. D_{max} and $dist$ are two measures used in the first and second stopping conditions.

2) The second condition is fulfilled if the distance between the current change-point C_k and the next change-point C_{k+1} exceeds the distance $dist$, predefined by the user. This rule allows to include noisy points, that do not effect the overall trend, as part of the signature. For example, points C_2, C_4 and C_7 in Figure 2. Formally,

$$(C_k - C_{k+1}) > dist \quad (6)$$

3) The third condition is fulfilled if the y value of the current change-point C_k (a sales figure or a sentiment value that corresponds to C_k) became lower than the local median (the median calculated over a moving window). This rule is important in order to avoid including non-peaking points as part of the event.

To illustrate, the change-point C_5 in Figure 2 became a starting event point P_{start}, since it fulfilled the first condition $(C_5 - P_{peak}) > D_{max}$; the change-point C_8 became the end point P_{end}, since it satisfied the second condition $(C_8 - C_9) > dist$.

All time segments between P_{start_i} and P_{peak_i} denote a growth signature, while all time segments between P_{peak_i} and P_{end_i} represent a relaxation signature.

The algorithm for extracting the growth signature:

```
currentPoint = X_peak − 1
stoppingCriterion = FALSE
while (stoppingCriterion != TRUE){
  isChangePoint = CheckIfChangePoint(currentPoint);
  if (isChangePoint){
    nextChangePoint = FindNextChangePoint(currentChangePoint)
    stoppingCriterion = AnalyseStoppingCriterion()
    if (stoppingCriterion != TRUE)
      currentPoint = nextChangePoint − 1
  }
  else
    currentPoint = currentPoint − 1
}
X_start = currentPoint
```

To extract the relaxation signature of the event, one should replace "-1" with "+1" in the above code, and X_{start} with X_{end}.

B. Events Clustering

One of the objectives of this study is to identify different types of Twitter Sentiment and Sales events based on their

shape. Solving events clustering problem is equivalent to solving a similarity matching problem for the collection of time series representing events. In this study, we used KMeans clustering [28] and compare three approaches for calculating the distance between data points, two known methods and one novel method that we propose. K-means was chosen because it is simple, widely used and computationally efficient.

1) **Euclidean Distance (ED)** is the most used distance function that calculates the similarity between two sequences of the same length by summing the ordered point-to-point distance between them.

$$d(T,S) = \sqrt{\sum_{i=1}^{N} (T_i - S_i)^2} \quad (7)$$

Where T and S are time series of length n.

2) **Dynamic Time Warping Distance (DTW)**. While Euclidean distance is a linear map between points, DWT [29] allows non-linear mapping. Given two time series $T = \{T_1, T_2, ..., T_n\}$ and $S = \{S_1, S_2, ..., S_m\}$ of length n and m respectively, a distance matrix $n * m$ is constructed where each element represents a pairwise Euclidean distance between points in the two sequences:

$$distMatrix = \begin{bmatrix} d(T_1,S_1) & ... & d(T_1,S_m) \\ ... & ... & ... \\ d(T_n,S_n) & ... & d(T_n,S_m) \end{bmatrix} \quad (8)$$

The objective of DTW is to find the warping path $W = \{w_1, w_2, ..., w_K\}$ of continuous elements on distMatrix that minimizes the following function:

$$DTW(T,S) = \min \left(\sqrt{\sum_{k=1}^{K} w_k} \right) \quad (9)$$

3) **First Derivative Based Distance (FDD)**. We propose a new way of calculating the distance measure based on the first derivatives of the time series. Our algorithm works as follows:

- Each time series $X_i = \{(t_i^1, y_i^1), ..., (t_i^n, y_i^n)\}$ is divided into L number of sequential stripes of equal length along the time-axis [30], where t_i is time and y_i is a corresponding value;
- The first derivative d_i^l for each stripe is calculated as

$$d_i^l = \frac{(y_i^{l+1} - y_i^l)}{\Delta t} \quad (10)$$

where (t_i^l, y_i^l) and (t_i^{l+1}, y_i^{l+1}) are the start end end coordinates for the l_{th} stripe of the i_{th} time sequence;
- The Euclidean distance is computed between the corresponding first derivatives of both time series.
- Parameter L is chosen to be equal to one third of the average length of the time series.

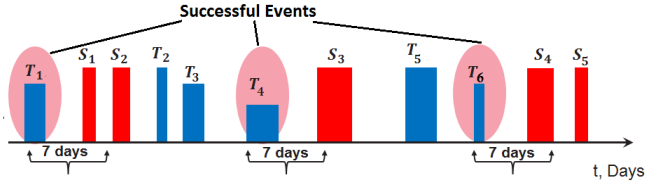

Fig. 3. Schematic representation of successful Twitter events for the time window of 7 days. Red bars represent sales events, blue bars represent Twitter events, pink circles highlight successful Twitter events.

C. Quantification of Events Significance

The objective of this study is to measure the ability of events in one time series to predict events in the other time series. Using Twitter sentiment time series and sales time series as a case-study, the null hypothesis, H0, can be defined as follows: sales events follow Twitter events in a random manner. To test the hypothesis we propose two significance tests.

1) Statistical Test One: This test evaluates how significant the number of successful Twitter events is. A Twitter event is considered to be successful if within a specified event window (7, 14, 21 days, etc.) there is a sales event following it (see Figure 3). By defining the event window within which the analysis is performed, we follow a traditional approach of the event study (as proposed by [15]).

To measure the significance of the observed number of successful events we randomise the positions of sales events 1000 times, calculate the number of successful events for each randomised scenario, and then compare the empirical results to the results after randomisation. The null hypothesis is rejected for a p-value of less than 0.05. If the number of observed successful Twitter events is outside of 95% confidence interval for the randomised case, we can conclude that occurrence of Twitter events before sales events is not random.

2) Statistical Test Two: Selection of the event window, as performed in the first test, is a challenging task. In the second test, we propose an algorithm that allows to simultaneously measure the importance of events for all possible event windows and then automatically determine which event window is the best choice.

In this test, we consider that a Twitter event has a power to predict sales if at least one sales event appears after the beginning of that Twitter event. A sales event S_i might happen after a Twitter event T_i at different distances. For each Twitter event we store the distance d_i at which the first sales event happened. In the situations when multiple Twitter events are followed by one sales event we consider that all these Twitter events contributed to the appearance of a single sale event. We assign a weight w_i to each Twitter event, which is inversely proportional to the distance between the Twitter event and the following sales event: the longer the distance between a sale and a Twitter event, the smaller the weight, and vice versa. The weights of Twitter events that have one sale following them should sum up to one. This is the most conservative approach, which prevents us from over-counting the number of predictive events, although it may result in in under-counting them.

For example, in Figure 4, we observe that a sale event S_3 has three Twitter events, T_2, T_3 and T_4, preceding it.

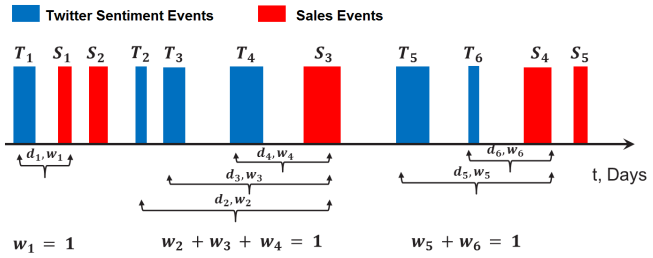

Fig. 4. Schematic representation of sales events together with Twitter events and their weights. Red bars represent sales events, blue bars represent Twitter events.

We consider that all three Twitter events contributed to the appearance of the sales event S_3. Each of the Twitter events is being assigned a weight w_2, w_3, w_4, respectively, with the sum of the weights being equal to one: $w_2 + w_3 + w_4 = 1$. The weights w_2, w_3, w_4 are inversely proportional to the distances d_2, d_3 and d_4 at which Twitter events occurred. For example, the longest distance is between the Twitter event T_2 and the sale event S_3, thus, T_2 event should have the smallest weight assigned to it. Conversely, the shortest distance is between event T_4 and the sales event, thus, it should be assigned the highest weight, assuming that Twitter event T_4 has the highest probability of contributing to the occurrence of the sale event S_3.

In this test, we are interested in analysing the probability of observing at least one sales event after a Twitter event for each event window. For this purpose we calculate a cumulative probability for each distance using the following steps:

- Calculate the time interval d_i between each Twitter event T and the first following it sales event S.
- Calculate the corresponding weight for each Twitter event w_i.
- Sort Twitter events in the incremental order of distances.
- For each Twitter event calculate the probability to have at least one sales event following it, by dividing the total number of events for each distance by the sum of their weights.
- Compute the cumulative probability for every distance by summing up the probabilities of the previous distances.

To identify the event windows at which Twitter events have significant power to predict sales events we perform a randomisation test. We randomise positions of all sales events, preserving their number and duration. The randomisation is made in a way that events do not overlap. The randomisation process is repeated 1000 times, and for each run the cumulative probability is calculated. To quantify the significance we compute the difference between the observed and randomised cumulative probabilities for each of 1000 runs. We then calculate the average difference, standard errors and confidence intervals. We reject the null hypothesis if the 2.5 percentile of the differences is higher than zero, which means that 97.5% of all differences are higher than zero. This allows us to conclude that that the observed system has a statistically significant advantage over the randomised system, and, therefore, sales are likely to follow Twitter events in a non-random manner.

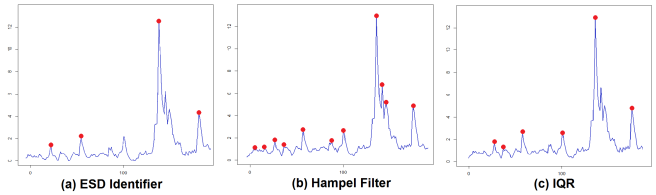

Fig. 5. Results of peak detection for sales data using three different methods: (a) ESD identifier; (b) Hampel filter; (c) IQR. Red dots denote peaks in time series.

III. Application of the Framework to Retail Brands

Today, consumers leave feedback about their customer experiences and express views about products on social media websites. This information can be relevant for retail brands to predict sales, achieve more insight on inventory management, plan marketing campaigns and adjust the offer in real time. The ultimate goal of any brand would be the ability to predict spikes or abnormal events in sales using the information from social-media.

In this study, we use the proposed framework for events detection and differentiation in order to meet two business objectives:

1) Quantifying the significance of all Twitter events in predicting sales events. This objective allows us to understand whether non-filtered Twitter signal contains useful information about sales.
2) Quantifying the significance of specific types of Twitter events in predicting sales events. The goal of this objective is to understand whether different dynamics of Twitter behaviour (events with different types of growth and relaxation signatures) have different effect on future sales, and to identify the dynamics that have significant predictive power.

A. Events Detection Results

The first step of the events detection process is identification of individual peaks in time series. In our analysis we used a moving window equal to 7 days for calculating the mean and the median.

Comparing the results of three peaks detection methods, we observed that the ESD identifier missed some relevant peaks. The problem with this method is that both the mean and the standard deviation are often extremely sensitive to the presence of outliers. In fact, if the level of outliers is higher than 10%, ESD detects no outliers at all. On the contrary, Hampel filter identified even small increases in sales as anomalous peaks. Hampel filter is much more resistant to the influence of outliers, however, it can be too aggressive in classifying values that are not really extreme. It can be shown that if more than 5% of the data points have the same value, MAD is computed to be 0, so any value different from the residual median is classified as an outlier. Compared to ESD and Hampel filter,

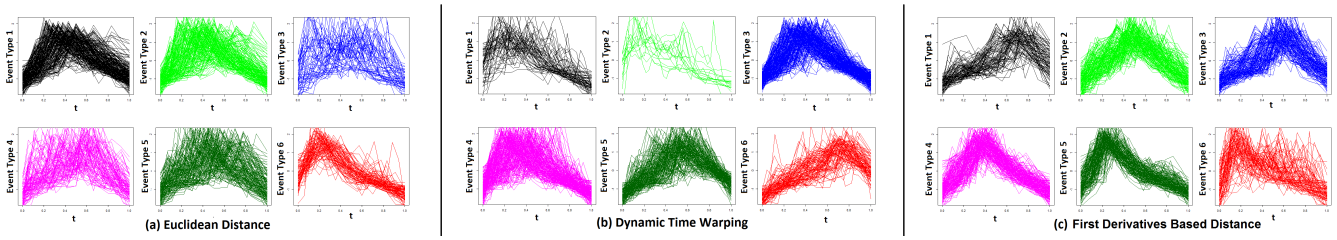

Fig. 6. Different types of Twitter sentiment events obtained using K-means with (a) ED, (b) DTW and (c) FDD.

the IQR method showed superior performance. It captured the big spikes in data and did not suffer from identifying small increases in sales as outliers. We therefore used IQR as the method in our final analysis, since it provided a good balance between the amount of false-positives and false-negatives.

In the next step, we performed extraction of events signatures as described in section II-A2. As a result, we identified 810 events in Twitter sentiment and 760 sales events across 75 brands.

B. Events Clustering Results

K-Means clustering of Twitter sentiment events revealed interesting results: three distinct clusters of shapes persisted across all methods (Fig. 6): a cluster with symmetric growth and relaxation dynamics; a cluster with a long growth signature and a short relaxation signature; a cluster with a short growth signature and a long relaxation signature. Using the elbow method [31] we identified that the optimum number of clusters for Twitter sentiment is six. This number of clusters allowed to capture the three main dynamics of events as well as variations in the slopes of growth/relaxation for each method.

Clustering based on all three types of distance measures (ED, DTW, FFD) allowed to capture the three main dynamics of Twitter sentiment events, however, the measure of spread between cluster objects was different for the three measures:

- **Euclidean distance results.** Extracted Twitter sentiment events have a duration in the range between 7 and 42 days. Events of different lengths might have similar shapes and should be clustered into one group. Clustering based on ED failed to produce this outcome. Since ED performs linear mapping, in most cases the ED between the time series of similar lengths is smaller than the distance between the time series of varying length, independently of the growth and relaxation shapes. As a result, ED grouped the time series primarily based on their length (Fig. 6(a)). Apart from that, Euclidean distance doesn't handle outliers, and it is very sensitive to signal transformations: shifting, amplitude and time scaling. These drawbacks make Euclidean inappropriate for our application.
- **DTW results.** DTW was designed to handle time sequences of varying length, solving the problem of ED. However, matching the shapes that do not line up in X-axis introduced a different kind of problem for our study: DTW often grouped together time series that have a non-matching location of the peak. This resulted in noisy clusters (Fig. 6(b)). Additionally, a non-linear mapping is computationally very expensive.
- **First Derivative based distance results.** Our new FDD approach, based on the first derivatives of the strips, produced the most clean results. Figure 6(c) shows that the distance between time series within each class is smaller than for the other two clustering methods. Our approach resolved the problems faced with the other two methods:
  1) FDD is able to cluster time series of different lengths by automatically normalising them to a number of data points equal to the number of stripes. The clustering considers the growth/relaxation signatures of the events, since the higher level feature of the first derivative allows to extract information about the shape. This solves the problem experienced with the Euclidean distance.
  2) FDD approach uses linear mapping between points which allows to capture the location of the peak in time and solves the problem of DTW.
  3) By taking into consideration only high-level features of the time series we reduce dimensionality and thus, reduce noise. This also significantly reduces the computation time.

Since results based on the FDD approach were better than the results for the other two methods, further analysis was performed using only the outcomes of FDD clustering.

C. Quantifying the significance of all Twitter events in predicting sales events

In this scenario, we performed analysis of the predictive power of Twitter events before clustering them into different groups.

1) Statistical Test One: In the first significance test, we calculated the number of successful Twitter events for both observed and randomised scenarios, as described in section II-C1. The analysis was performed for two event windows: 7 days and 21 days.

Table I summarises the results. For both event windows the number of empirically observed successful events appeared to be significantly larger than the number of successful events after randomisation of sales. For example, we empirically observed that 161 Twitter sentiment events had at least one sale event following them within 7 days. Conversely, in cases when the positions of sales events were randomised, the average number of Twitter events that had a sales event following them within 7 days was 136.93 with the 95% CI [134.9; 138.96]. Therefore, the empirical result of 161 was outside of the 95%

TABLE I. NUMBER OF SUCCESSFUL TWITTER EVENTS FOR TWO EVENT WINDOWS, 7 DAYS AND 21 DAYS.

Sentiment events	Predictive within 7 days	Predictive within 21 days
Empirical Results	161	434
Randomised Results 95% CI	136.93 [134.9; 138.96]	417.52 [415.09; 419.96]

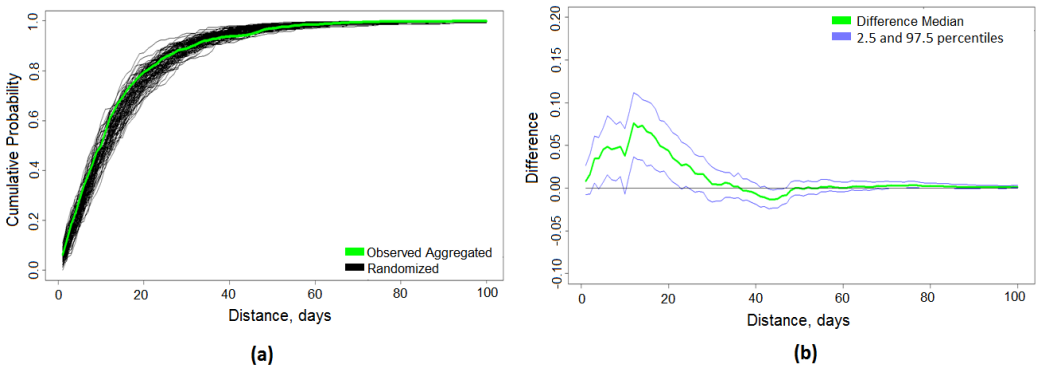

Fig. 7. (a) Cumulative probability of having a sale event within a specified time interval after a Twitter Sentiment event, (b) Difference between cumulative probabilities of observed and randomised data for Twitter Sentiment. The green line represents empirical results, the black lines represent randomised results and the blue lines represent 2.5 and 97.5 percentiles.

CI. Similar dynamic was observed for the event window of 21 days, suggesting that the cases when Twitter events precede sales events do not occur by chance.

2) Statistical Test Two: In this test, we calculated a cumulative probability of having at least one sales event after a Twitter event for both, observed and randomised scenarios, as described in section II-C2.

In Figure 7(a), the green line is the empirical cumulative probability and the black lines are cumulative probabilities for the randomised scenarios. We also calculated the differences between the observed cumulative probability and each of the randomised probabilities (Figure 7(b)). If there was no underlying relationship between Twitter and sales events, the median of the differences between the observed cumulative probability and probabilities obtained after randomisation of sales events, would be close to zero. However, Figure 7(b) shows that the median is above zero along with 2.5 percentile of differences for the first 21 days. According to these results, we can reject the null hypothesis with the confidence of 97.5%, and conclude that sales events follow Twitter events in a non-random manner. The period during which the 2.5 percentile line is above zero defines the optimum event window which corresponds to an interval of 21 days between the sales event and the Twitter event. In this scenario, there is a significant probability that the observed sale event did not happen after the Twitter event by chance.

D. Quantifying the significance of specific types of Twitter events in predicting sales events

In this section we clustered Twitter events into six classes using the FDD method as described in section II-B (Fig. 6(c)). We then analysed the predictive power of each Twitter event class independently. The null hypothesis, H0, was reformulated for the case of multiple Twitter events types as follows: Twitter sentiment events of different types appear before sales events in a random manner.

1) Statistical Test One: To test the null hypothesis we calculated the occurrence of each event type as a percentage of total number of successful events.

The results for a 7-days horizon are shown in table II, where the first line presents observed proportions of successful events of each type, and the second line presents proportions of each event type after randomisation. From 161 observed Twitter sentiment events that were classified as successful within 7-days the majority of events (26.09%) were classified as events of class 4, whereas the event class that had the smallest representation appeared to be class 6 (9.32%). Comparing the empirical proportions (the first row in table II) with the proportions after randomisation (the second row in table II), we observe that the results for event types 2, 3, 4 and 5 are significantly different from random results. Specifically, event types 3 and 5 are significantly under-represented while event types 2 and 4 are significantly over-represented. Very similar results can be observed for the 434 events that were successful within 21 day (table III). We observe that the relative frequencies of types 3, 4, 5 and 6 are significantly different from random.

Comparing the results for different distances, 7 and 21 days, we observe that relative frequencies of events types change depending on the time interval. For example, the event of type 4 is over-represented for the 7 days distance and under-represented for the 21 days distance.

Since we observed significant deviations from random in the proportions of different types of events, we can reject the null hypothesis and conclude that the occurrence of different types of Twitter events before sales events is not random. These means that different Twitter clusters have significantly different power to predict events in sales.

2) Statistical Test Two: As in III-C1, we calculated the cumulative probabilities of having at least one sales event after any Twitter event, however, in this test the probabilities were computed individually for every Twitter event type. To measure

TABLE II. RELATIVE FREQUENCIES OF SUCCESSFUL TWITTER EVENTS FOR SUCCESS TIME HORIZON OF 7 DAYS.

Event Types	Event Type 1	Event Type 2	Event Type 3	Event Type 4	Event Type 5	Event Type 6
Observed, %	10.55	23.6	13.66	26.09	16.77	9.32
Random 95% CI	10.88 [10.43; 11.32]	20.22 [19.56; 20.86]	16.91 [16.28; 17.54]	22.21 [21.6; 23.12]	20.18 [19.31; 20.71]	9.62 [9.18; 10.05]

TABLE III. RELATIVE FREQUENCIES OF SUCCESSFUL TWITTER EVENTS FOR SUCCESS TIME HORIZON OF 21 DAYS.

Event Types	Event Type 1	Event Type 2	Event Type 3	Event Type 4	Event Type 5	Event Type 6
Observed, %	11.06	20.51	17.74	22.35	19.35	8.99
Random 95% CI	10.91 [10.61; 11.08]	20.30 [20.07; 20.53]	16.43 [16.23; 17.14]	22.63 [22.38; 23.13]	19.90 [19.38; 20.13]	9.83 [9.61; 10.00]

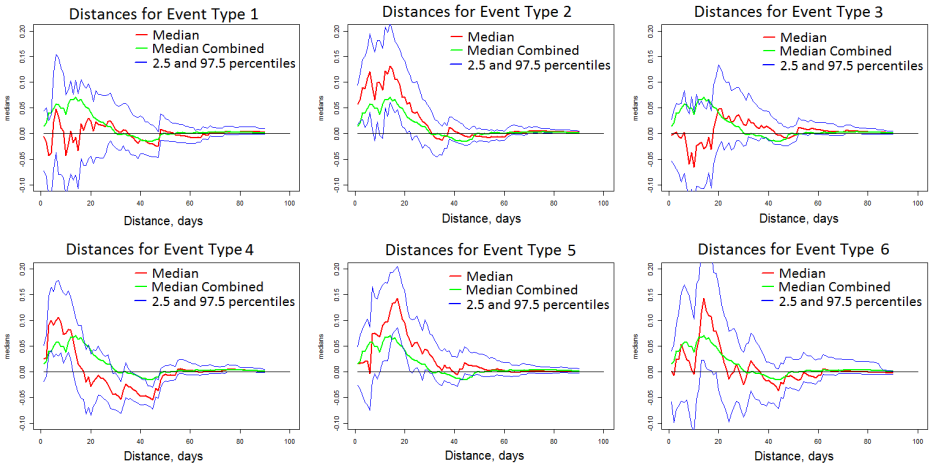

Fig. 8. Difference between cumulative probabilities of observed and randomised data for Twitter Sentiment events of different types. The red lines correspond to the difference for the specific Twitter type, the green lines correspond to the difference for the non-clustered Twitter signal, the blue lines represent the 2.5 and 97.5 percentiles.

the significance of predictive power of each individual Twitter cluster we calculated the differences between the empirical and randomised results (Fig. 8).

For events types 2, 4, 5 and 6 we observe that: 1) in the first few days/weeks, the 2.5 percentile of the differences between observed and randomised results is above zero, which indicates that sales events follow Twitter sentiment events of type 2, 4, 5 and 6 in a non-random manner; 2) the difference between observed cumulative probability for a specific Twitter type and the randomised sequences (red graph) is greater than the difference between the observed cumulative probability for the non-clustered Twitter signal and the randomised data (green graph), which means that signals of the event types 2, 4, 5 and 6 have better predictive power than the non-clustered Twitter signal. This is a very important finding which means that by using the signals solely of event types 2, 4, 5 and 6 we filter Twitter signal from noise, and can achieve higher accuracy of predicting sales events. It is interesting to notice that event type 2 has a consistent significant predictive power during the first 3 weeks after the event (the 2.5 percentile is continuously above zero), while event type 4 shows significant predictive power only during the first two weeks; event type 5 is predictive between 12 and 22 days; event type 6 is only predictive for a short period of time during the third week after the event.

This information can be incorporated into a forecasting model that considers the predictive power of different Twitter events at different distances. For example, the Twitter Sentiment events of type 2 can be used to predict sales events within the first 3 weeks after the twitter event, whereas Twitter events of type 4 can be used to predict sales events only within the first 2 weeks after a Twitter event.

IV. CONCLUSIONS

In the era of social-media networks, brands closely follow online discussions about their products and services in order to understand their consumers. In this context, detection of spikes in social-media could be of key importance for many retail brands.

To address these needs, we proposed a framework that allows automatic events detection, clustering and quantification of events' significance. The framework was tested on a large-scale dataset of 150 million Tweets and sales data of 75 brands.

Our research presents a contribution to the field of the event study by proposing a novel approach for filtering Twitter signal from noise by clustering it into different event types based on the growth and relaxation signatures. The predictive power of Twitter events in this study was evaluated using two scenarios: in the first scenario we performed the analysis for the non-clustered Twitter signal; in the second scenario we clustered Twitter sentiment events based on their growth and relaxation signatures and calculated the statistics of successful predictions separately for every event type. The main contribution of our research is identification of specific event types that have the power to predict events in sales.

The results can be summarised as follows:

- Twitter sentiment events can significantly improve prediction of events in sales.

- Events can be clustered into categories based on their shapes (position of the peak, growth and relaxation signatures).
- Different event shapes are differently associated with sales.
- Some sentiment event types have significantly higher predictive power than the non-clustered Twitter signal.

As the future direction of our research we aim to understand what different events' shapes represent in terms of Twitter dynamics and content (persistence of news, importance) and plan to incorporate the extracted knowledge into a forecasting model for consumer sales.

REFERENCES

[1] F. Atefeh and W. Khreich, *A Survey of Techniques for Event Detection in Twitter*, 2015, ch. 53, pp. 132–164.

[2] X. Dong, D. Mavroeidis, F. Calabrese, and P. Frossard, "Multiscale event detection in social media," *Data Mining and Knowledge Discovery*, vol. 29, no. 5, pp. 1374–1405, 2015. [Online]. Available: http://dx.doi.org/10.1007/s10618-015-0421-2

[3] D. T. Nguyen, D. Hwang, and J. Jung, *Event Detection from Social Data Stream Based on Time-Frequency Analysis*. Springer International Publishing, 2014, pp. 135–144.

[4] H. Sayyadi, M. Hurst, and A. Maykov, *Event detection and tracking in social streams*, 2009.

[5] H. Becker, M. Naaman, and L. Gravano, *Event identification in social media*, 2009.

[6] C. A. Charu and S. Karthik, *Event Detection in Social Streams*, 2012, ch. 53, pp. 624–635. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/1.9781611972825.54

[7] T. O. Sprenger, P. G. Sandner, A. Tumasjan, and I. M. Welpe, "News or noise? using twitter to identify and understand companyspecific news flow," *Journal of Business Finance and Accounting*, vol. 41, pp. 791–830, 2014.

[8] W. Antweiler and M. Z. Frank, "Is all that talk just noise? The information content of internet stock message boards," *Journal of Finance*, vol. 59, no. 3, pp. 1259–1294, 2004.

[9] J. E. Thompson, "More methods that make little difference in event studies," *Journal of Business Finance and Accounting*, vol. 15, pp. 77–86, 1988.

[10] Y. Sano, K. Yamada, H. Watanabe, H. Takayasu, and M. Takayasu, "Empirical analysis of collective human behavior for extraordinary events in the blogosphere," *Phys. Rev. E*, vol. 87, p. 012805, 2013.

[11] R. Crane and D. Sornette, "Robust dynamic classes revealed by measuring the response function of a social system," *Proceedings of the National Academy of Sciences*, vol. 105, no. 41, pp. 15 649–15 653, 2008.

[12] D. Sornette, F. Deschatres, T. Gilbert, and Y. Ageon, "Endogenous versus exogenous shocks in complex networks: An empirical test using book sale rankings," *Physical Review Letters*, vol. 93, 2004.

[13] D. Sornette, *Endogenous versus Exogenous Origins of Crises*, ser. The Frontiers Collection. Springer Berlin Heidelberg, 2006, pp. 95–119.

[14] R. Dijkman, P. Ipeirotis, F. Aertsen, and R. Helden, "Using twitter to predict sales: A case study," 2015. [Online]. Available: http://arxiv.org/abs/1507.00955

[15] C. M. A., "Event studies in economics and finance," *Journal of Economic Literature*, vol. 35, no. 1, pp. 13–39, 1997. [Online]. Available: http://EconPapers.repec.org/RePEc:aea:jeclit:v:35:y:1997:i:1:p:13-39

[16] W. Bessembinder, K. Kahle, W. Maxwell, and D. Xu, "New methodology for event studies in bonds," *Review of Financial Studies*, vol. 22, p. 42194258, 2009.

[17] Y. Konchitchki and D. E. OLeary, "Event study methodologies in information systems research," *International Journal of Accounting Information Systems*, vol. 12, pp. 99–115, 2011.

[18] S. J. Brown and J. B. Warner, "Using daily stock returns: the case of event studies," *Journal of Financial Economics*, vol. 14, pp. 3–31, 1985.

[19] G. Norman and D. Streiner, *Biostatistics: The Bare Essentials*, ser. Pmph USA Ltd Series. B.C. Decker, 2008. [Online]. Available: https://books.google.co.uk/books?id=y4tWQl_8Ni8C

[20] O. Kolchyna, T. T. P. Souza, P. Treleaven, and T. Aste, "Twitter sentiment analysis: Lexicon method, machine learning method and their combination," in *Handbook of Sentiment Analysis in Finance*, G. Mitra and X. Yu, Eds. Editors, 2015, ch. 5. [Online]. Available: http://arxiv.org/abs/1507.00955

[21] T. T. P. Souza, O. Kolchyna, P. C. Treleaven, and T. Aste, "Twitter sentiment analysis applied to finance: A case study in the retail industry," in *Handbook of Sentiment Analysis in Finance*, G. Mitra and X. Yu, Eds. Editors, 2015, ch. 23. [Online]. Available: http://arxiv.org/abs/1507.00784

[22] B. Rosner, "Percentage points for a generalized esd many-outlier procedure," *Technometrics*, vol. 25(2), pp. 165–172, 1983.

[23] F. R. Hampel, "A general qualitative definition of robustness," *Ann. Math. Statist.*, vol. 42, no. 6, pp. 1887–1896, 12 1971. [Online]. Available: http://dx.doi.org/10.1214/aoms/1177693054

[24] ——, "The influence curve and its role in robust estimation," *j-J-AM-STAT-ASSOC*, vol. 69, no. 346, pp. 383–393, Jun. 1974.

[25] H. Liu, S. Shah, and W. Jiang, "On-line outlier detection and data cleaning." *Computers and Chemical Engineering*, vol. 28, no. 9, pp. 1635–1647, 2004.

[26] L. Davies and U. Gather, "The identification of multiple outliers," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 782–792, 1993. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476339

[27] J. Tukey, "Exploratory data analysis," *Addison-Wesley*, 1977.

[28] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. M. L. Cam and J. Neyman, Eds., vol. 1. University of California Press, 1967, pp. 281–297.

[29] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowl. Inf. Syst.*, vol. 7, no. 3, pp. 358–386, Mar. 2005. [Online]. Available: http://dx.doi.org/10.1007/s10115-004-0154-9

[30] D. Toshniwal and R. Joshi, "Using cumulative weighted slopes for clustering timeseries data," *GESTS Intl Trans. Computer Science and Engr.*, vol. 20, 2005.

[31] R. L. Thorndike, "Who belong in the family?" *Psychometrika*, vol. 18(4), pp. 267–276, 1953.