Emotional Trends in Social Media
A State Space Approach

Sören Volgmann[1] and **Francisco Rangel**[2, 3] and **Oliver Niggemann**[1] and **Paolo Rosso**[3]

Abstract. In this paper, a new modeling and learning approach is presented which is based on two assumptions from the field of psychology: 1. The number of Tweets mainly depends on previous dynamics of the discussion, i.e. a state-space modeling approach is used for the first time. 2. Humans mainly react to emotional stimuli, i.e. Tweets are automatically characterized by their emotional content. Therefore, the emotions of conversations are extracted and used for system identification and parameter estimation of a state space model, which deals with events and its transitions.

The proposed approach is further evaluated with an example discussion about the Spanish corruption affair held on Twitter during summer 2013. The experimental results show a method to model and learn the evolution of social media discussion based on emotions.

1 INTRODUCTION

The use of social media, especially Twitter, is rapidly increasing. It allows users to share information in real-time and gains uprising attention in many different domains, such as political campaigns, disaster communication etc. The principal factor of data diffusion in Twitter is the possibility to broadcast information from the network of followees[4] to the network of followers[5]. Research has been paying attention to trending topic detection (i.e. [6]), but the estimation of such trends, which are covered by noise and cyclic effects, is still difficult.

This paper proposes a method to model the evolution of a certain Twitter discussion using a state space approach, by analyzing the emotionality of the discussion. The model is used to deduce the dynamic evolution of social media discussions and enabling to predict future trends. Contrary to this approach, many approaches in research use an absolute time base to analyze the dynamic behavior of Twitter discussions, i.e. the evolution changes because it is Monday morning (see [2], [1]). Additionally, there is an interest of obtaining emotions expressed in text [7], usually based on the six basic emotions of Ekman [4].

2 DATA AND METHODS

Data from Twitter, and also social media in general, is covered by variability and noise, which often makes trends imperceptible. Furthermore, trends are hidden through cyclic events, such as day and night effects. This is due to the fact that people publish more messages during the day in comparison to the night. Besides variability and noise, trends in social media discussions are very short. In [1] statistical models are used to predict future events, where the authors of [2] use normalization over at least two weeks of data in order to detect abnormalities and events. In comparison to [2], we apply a state-based approach considering the emotions of discussions, which works with less data to extract trends.

In this work, we focus on political discussions carried out on Twitter. In particular, Twitter data of the Spanish corruption scandal during summer 2013 is used. The data comprises 4,397,023 Tweets and was retrieved from 09-07-2013 to 02-10-2013.

For whole conversation, we obtain several features. The six basic emotions (*anger*, *joy*, *sadness*, *surprise*, *disgust* and *fear*) according to [4] are extracted using a lexicographical approach from [8]. Moreover, these classical features, available through Twitter API[6], are observed: *followers*, *followees*, *Retweets*, *unique Tweets*, *number of Tweets* and *unique users*.

Next, a state-based model expresses certain features depending on the previous state, i.e. the number of Tweets for the current hour depends on the number of Tweets from the previous hour. We use a two step approach to estimate an overall state space model S, which is defined in Definition 1 according to [3]. In the first step a predefined *local linear trend model with seasonal components* is used to decompose uni-variate data and estimate its trend. In the second step, the estimated trends are combined to a multivariate feature vector and an overall state-based model is estimated.

Definition 1 *A state space system S is:*

$$S : \begin{cases} \boldsymbol{x}_{t+1} & = \boldsymbol{F}\boldsymbol{x}_t + \boldsymbol{W}_t, \quad \boldsymbol{W}_t \sim MND(\boldsymbol{0}, \boldsymbol{Q}_t) \\ \boldsymbol{y}_t & = \boldsymbol{G}\boldsymbol{x}_t + \boldsymbol{V}_t, \quad \boldsymbol{V}_t \sim MND(\boldsymbol{0}, \boldsymbol{R}_t) \end{cases} \quad t{=}1{,}2{,}...$$

where $\boldsymbol{x}_t \in \mathbb{R}^n$, $\boldsymbol{y}_t \in \mathbb{R}^q$, $\boldsymbol{W}_t \in \mathbb{R}^n$ *and* $\boldsymbol{V}_t \in \mathbb{R}^q$.

A state-based system S consists of two equations: an *observation equation* $\boldsymbol{y}_t$ and a *state equation* $\boldsymbol{x}_{t+1}$. The observations describe the observable data from the system, in this case the extracted features from a discussion, and the state equation denote the internal system behavior. Additionally, the states and the observations are covered by a multi normal distributed noise term (*MND*).

The states are often hidden and not fully observable. Therefore, Kalman filtering, introduced by [5], is used to calculate the linear estimates, with the minimum mean square error, of the state vector $\mathbf{x}_t$ in terms of the observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$. In addition, the parameters $\boldsymbol{F}$, $\boldsymbol{Q}_t$ and $\boldsymbol{R}_t$ of the system are estimated through a maximum likelihood estimation.

[1] Fraunhofer Institute IOSB-INA, Lemgo, Germany, email: {soeren.volgmann, oliver.niggemann}@iosb-ina.fraunhofer.de

[2] CTO Autoritas Consulting S.A., Madrid, Spain, email: francisco.rangel@autoritas.es

[3] NLE Lab – PRHLT Research Center, Universitat Politècnica de València, València, Spain, email: prosso@dsic.upv.es

[4] People who are being tracked on Twitter platform.

[5] People who are tracking particular persons, groups, organizations, etc. on Twitter platform.

[6] https://dev.Twitter.com/docs/api/1.1

3 EXPERIMENTS AND DISCUSSION

Figure 1 depicts the raw data, in this case the number of Tweets, for a time period of 22 days. In the data a periodicity is strongly apparent – a regularly up and down evolution of the discussion is visible each day. This effect was also discovered by [2] and is due to usual day schedule. At around 2 am the activity on Twitter decreases, whereas in the morning around 7 am the activity increases again. The highest number of Tweets is usually observed during noon time.

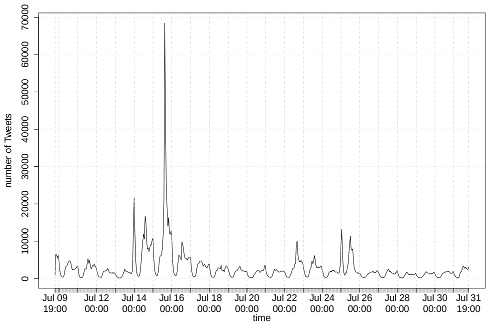

Figure 1. Raw data from Twitter platform (number of Tweets per hour).

Obviously, there are huge peaks in the data, which indicate certain events. For instance, the peak around 15th of July where some information about a private conversation between current Prime Minister of Spain *Rajoy* and his party treasurer *Bárcenas* was published, after *Rajoy* denied to have talked to *Bárcenas*.

The peaks cause an exponential growth of the conversation, which is suppressed by taking the logarithm of the data, shown in (a) of Fig. 2. Applying the state space model with Kalman filtering decomposes the data, where the first state represents the local linear level and the trend component, depicted in (b) of Fig. 2. In (c) of Fig. 2, the slow varying trend component of (a) is estimated, which is the slope of (a). The cyclic events of the data are removed and modeled into the remaining states of the state space model.

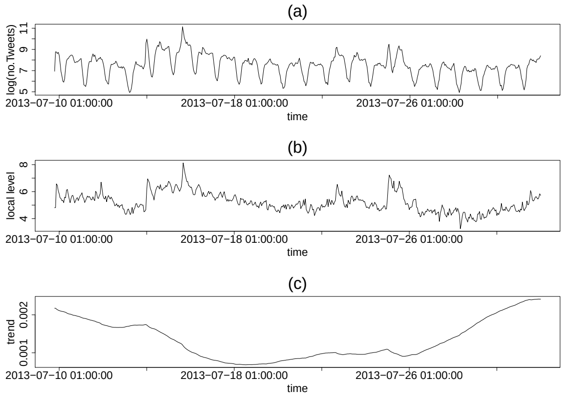

Figure 2. Decomposition of no. of Tweets per hour: (a) logarithmic data , (b) local linear trend component, (c) trend component.

Each feature is decomposed according to Fig. 2 and normalized to its maximum to visualize differences among the data. The result is a slow varying trend for each feature where the cyclic components and the noise are separated. After applying Kalman filtering, trends of the features are more apparent and interactions between features are visualized. In comparison to [2], abnormalities in the evolution of the discussion can be visualized even within a short period of time. Despite the observation time has an influence on the machine learning approach, there is no restriction on the amount of data.

Figure 3 illustrates a showcase of one state estimate from the learned model. In this example, the solid line denotes the trend for the emotion *anger* and the dotted line denote the estimated trend using Kalman filtering. Kalman filtering estimates the observed data with a slight difference and the trend estimation closely follows the original trend, despite the values are very small.

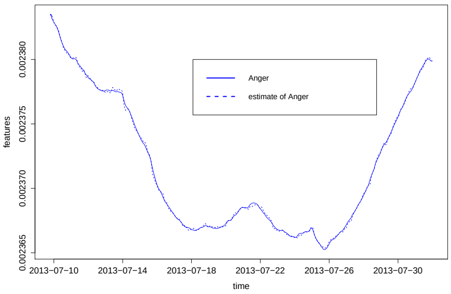

Figure 3. Extracted trends of *anger* and its estimates.

4 CONCLUSION AND FUTURE WORK

We have shown a method to learn a state-based model out of data from social media, which deals efficiently with cyclic events and noise. The approach is not limited to a certain amount of data compared to [2]. Furthermore, the derived system is useful to deduce the evolution of discussion upon emotional features and make hidden changes more apparent.

In future work, instead of predicting future events like [1] our approach will be used to predict slow varying future trends. Additionally, external events, happening outside of social media platforms (i.e. news, stock markets, etc), could be included into the used model.

REFERENCES

[1] G. Amodeo, R. Blanco, and U. Brefeld, ‘Hybrid models for future event prediction’, in *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1981–1984. ACM, (2011).

[2] K. Balog, G. Mishne, and M. de Rijke, ‘Why are they excited?: identifying and explaining spikes in blog mood levels’, in *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pp. 207–210. Association for Computational Linguistics, (2006).

[3] P. J. Brockwell and R. A. Davis, *Introduction to time series and forecasting*, Taylor & Francis US, 2002.

[4] P. Ekman, ‘Universals and cultural differences in facial expressions of emotion.’, in *Nebraska symposium on motivation*. University of Nebraska Press, (1971).

[5] R. E. Kalman et al., ‘A new approach to linear filtering and prediction problems’, *Journal of basic Engineering*, **82**(1), 35–45, (1960).

[6] N. Pervin, F. Fang, A. Datta, K. Dutta, and D. Vandermeer, ‘Fast, scalable, and context-sensitive detection of trending topics in microblog post streams’, *ACM Transactions on Management Information Systems (TMIS)*, **3**(4), 19, (2013).

[7] F. Rangel and P. Rosso, ‘On the identification of emotions in facebook comments’, in *In Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI (ESSEM 2013)*. A workshop of the XIII International Conference of Italian Association for Artificial Intelligence (AI*IA 2013), (2013).

[8] G. Sidorov, S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Treviño, and J. Gordon, ‘Empirical study of machine learning based approach for opinion mining in tweets’, in *Advances in Artificial Intelligence*, 1–14, Springer, (2013).