Sub-Event Detection from Twitter Streams as a Sequence Labeling Problem

Giannis Bekoulis **Johannes Deleu** **Thomas Demeester** **Chris Develder**
Ghent University - imec, IDLab
Department of Information Technology
firstname.lastname@ugent.be

Abstract

This paper introduces improved methods for sub-event detection in social media streams, by applying neural sequence models not only on the level of individual posts, but also directly on the stream level. Current approaches to identify sub-events within a given event, such as a goal during a soccer match, essentially do not exploit the sequential nature of social media streams. We address this shortcoming by framing the sub-event detection problem in social media streams as a sequence labeling task and adopt a neural sequence architecture that explicitly accounts for the chronological order of posts. Specifically, we (i) establish a neural baseline that outperforms a graph-based state-of-the-art method for binary sub-event detection (2.7% micro-F_1 improvement), as well as (ii) demonstrate superiority of a recurrent neural network model on the posts sequence level for labeled sub-events (2.4% bin-level F_1 improvement over non-sequential models).

1 Introduction

Social media allow users to communicate via real-time postings and interactions, with Twitter as a notable example. Twitter user posts, i.e., tweets, are often related to events. These can be social events (concerts, research conferences, sports events, etc.), emergency situations (e.g., terrorist attacks) (Castillo, 2016), etc. For a single event, multiple tweets are posted, by people with various personalities and social behavior. Hence, even more so than (typically more neutral) traditional media, this implies many different perspectives, offering an interesting aggregated description.

Given this continuous and large stream of (likely duplicated) information in Twitter streams, and their noisy nature, it is challenging to keep track of the main parts of an event, such as a soccer match. Automating such extraction of different sub-events within an evolving event is known as sub-event detection (Nichols et al., 2012). For tracking each of the sub-events, the timing aspect is an important concept (i.e., consecutive tweets in time). Thus, a sequential model could successfully exploit chronological relations between the tweets in a Twitter stream as an informative feature for sub-event detection.

Several methods have been proposed for sub-event detection: clustering methods (Pohl et al., 2012), graph-based approaches (Meladianos et al., 2015), topic models (Xing et al., 2016) and neural network architectures (Wang and Zhang, 2017). None of these studies exploits the chronological relation between consecutive tweets. In contrast, our work does take into account that chronological order and we predict the presence and the type of a sub-event exploiting information from previous tweets. Specifically, we (i) propose a new neural baseline model that outperforms the state-of-the-art performance on the binary classification problem of detecting the presence/absence of sub-events in a sports stream, (ii) establish a new reasonable baseline for predicting also the sub-event *types*, (iii) explicitly take into account chronological information, i.e., the relation among consecutive tweets, by framing sub-event detection as a sequence labeling problem on top of our baseline model, and (iv) perform an experimental study, indicating the benefit of sequence labeling for sub-event detection in sports Twitter streams.

2 Related Work

Twitter streams have been extensively studied in various contexts, such as sentiment analysis (Kouloumpis et al., 2011), stock market prediction (Nguyen and Shirai, 2015) and traffic detection (D'Andrea et al., 2015). Specifically, for sub-event detection in Twitter, several approaches

Proceedings of NAACL-HLT 2019, pages 745–750
Minneapolis, Minnesota, June 2 - June 7, 2019.

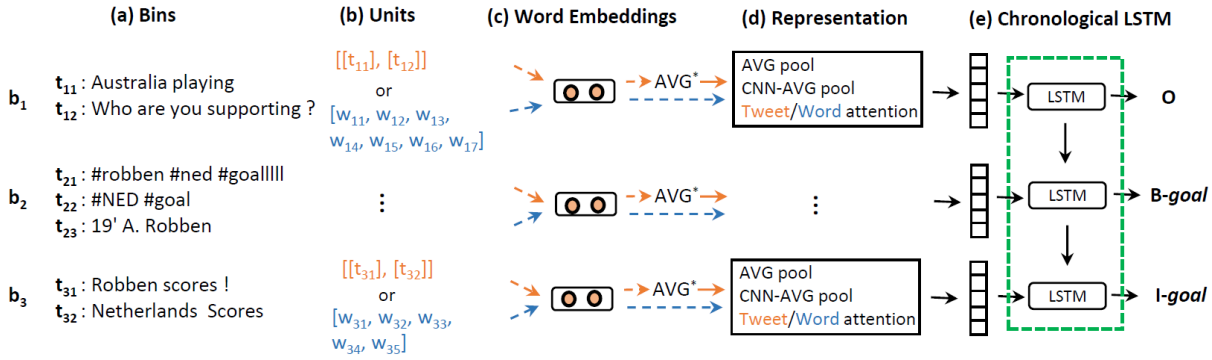

Figure 1: Our sub-event detection model comprises: (a) a bin layer, (b) a unit layer, (c) a word embeddings layer, (d) a representation layer and (e) a **chronological LSTM** layer to model the natural flow of the sub-events within the event. We represent each bin using either (i) a **tweet**- or (ii) a **word**-level representation. The AVG* represents an average pool operation, performed either directly on the embeddings or on the tweet's LSTM representation.

have been tried. *Unsupervised methods* such as clustering aim to group similar tweets to detect specific sub-events (Pohl et al., 2012; Abhik and Toshniwal, 2013) and use simple representations such as tf-idf weighting combined with a similarity measure. Other unsupervised algorithms use topic modeling approaches, based on assumptions about the tweets' generation process (Xing et al., 2016; Srijith et al., 2017). Several methods (Zhao et al., 2011; Zubiaga et al., 2012; Nichols et al., 2012) assume that a sub-event happens when there is a 'burst', i.e., a sudden increase in the rate of tweets on the considered event, with many people commenting on it. Recently, neural network methods have used more complicated representations (Wang and Zhang, 2017; Chen et al., 2018). Also *supervised methods* have been applied (Sakaki et al., 2010; Meladianos et al., 2018) for the sub-event detection task. These methods usually exploit graph-based structures or tf-idf weighting schemes. We believe to be the first to (i) exploit the chronological order of the Twitter stream and take into account its sequential nature, and (ii) frame the sub-event detection problem as a sequence labeling task.

3 Model

3.1 Task Definition

The goal is, given a main event (i.e., soccer match), to identify its core sub-events (e.g., goals, kick-off, yellow cards) from Twitter streams. Specifically, we consider a *supervised* setting, relying on annotated data (Meladianos et al., 2018).

3.2 Word- vs Tweet-Level Representations

Similar to previous works, we split a data stream into time periods (Meladianos et al., 2018): we form bins of tweets posted during consecutive time intervals. E.g., for a soccer game, one-minute intervals (bins) lead to more than 90 bins, depending on the content before and after the game, halftime, stoppage time, and possibly some pre-game and post-game buffer. Thus, for each bin, we predict either the presence/absence of a sub-event (Section 3.3) or the most probable sub-event type (Section 3.4), depending on the evaluation scenario.

We consider representing the content of each bin either on (i) a word-level or (ii) a tweet-level (see Fig. 1). Formally, we assume that we have a set of n bins $b_1, ..., b_n$, where each bin b_i consists of m_i tweets and k_i words (i.e., all words of tweets in bin b_i). Then, the *tweet-level representation* of bin b_i is symbolized as $t_{i1}, ..., t_{im_i}$, where t_{im_i} is the m_i^{th} tweet of bin b_i. In the *word-level representation*, we chronologically concatenate the words from the tweets in the bin: $w_{i1}, .., w_{ik_i}$, where w_{ik_i} is the k_i^{th} word of bin b_i.

3.3 Binary Classification Baseline

To compare with previous work (Meladianos et al., 2018), we establish a simple baseline for binary classification: presence/absence of a sub-event. For this case, we use as input the word-level representation of each bin. To do so, we use word embeddings (randomly initialized) with average (AVG) pooling (Iyyer et al., 2015) in combination with a multilayer perceptron (MLP) for binary classification, i.e., presence/absence of a

sub-event. Note that we experimented with pre-trained embeddings as well as max-pooling, but those early experiments led to performance decrease compared to the presented baseline model. We found that training based on average bin representations works substantially better than with max-pooling, and we hypothesize that this is related to the noisy nature of the Twitter stream.

3.4 Sequence Labeling Approach

Building on the baseline above, we establish a new architecture that is able to capture the sub-event types as well as their duration. We phrase sub-event detection in Twitter streams as a sequence labeling problem. This means we assume that the label of a bin is not independent of neighboring bin labels, given the chronological order of bins of the Twitter stream, as opposed to independent prediction for each bin in the binary classification baseline above. For instance, when a *goal* is predicted as a label for bin b_i, then it is probable that the label of the next bin b_{i+1} will also be *goal*. Although a sub-event may occur instantly, an identified sub-event in a Twitter stream can span consecutive bins, i.e., minutes: users may continue tweeting on a particular sub-event for relatively long time intervals. For this reason, we apply the well-known BIO tagging scheme (Ramshaw and Marcus, 1995) for the sub-event detection problem. For example, the beginning of a *goal* sub-event is defined as B-*goal*, while I-*goal* (inside) is assigned to every consecutive bin within the same sub-event, and the O tag (outside) to every bin that is not part of any sub-event. To propagate chronological information among bins, we adopt an LSTM on the sequence of bins as illustrated in Fig. 1, layer (e). Note that this tagging approach assumes that sub-events do not overlap in time, i.e., only at most one is ongoing in the Twitter stream at any point in time.

4 Experimental Setup

We evaluated our system[1] on the dataset from Meladianos et al. (2018), with tweets on 20 soccer matches from the 2010 and 2014 FIFA World Cups, totalling over 2M pre-processed tweets filtered from 6.1M collected ones, comprising 185 events. The dataset includes a set of sub-events, such as *goal*, *kick-off*, *half-time*, etc. To compare our binary classification *baseline system* to previous methods (Table 1), we use the same train/test splits as Meladianos et al. (2018), where 3 matches are used for training and 17 matches as test set. In this setting, we predict only the presence/absence of a sub-event. Similar to previous work, we count a sub-event as correct if at least one of its comprising bins has been classified as a sub-event. For the experimental study of our proposed *sequence labeling approach* for sub-event detection, where sub-event types are predicted, we have randomly split the test set into test (10 matches) and development (7 matches) sets. We use the development set to optimize the F_1 score for tuning of the model parameters, i.e., the word/tweet embedding representation size, LSTM hidden state size, dropout probability. We adopt 2 evaluation strategies. The first one, referred to as *relaxed* evaluation, is commonly used in entity classification tasks (Adel and Schütze, 2017; Bekoulis et al., 2018a,c) and similar to the binary classification baseline system evaluation: score a multi-bin sub-event as correct if at least one of its comprising bin types (e.g., *goal*) is correct, assuming that the boundaries are given. The second evaluation strategy, *bin-level*, is stricter: we count each bin individually, and check whether its sub-event type has been predicted correctly, similar to the token-based evaluation followed in Bekoulis et al. (2018b).

5 Results

5.1 Baseline Results

Table 1 shows the experimental results of our baseline model. The Burst baseline system is based on the tweeting rate in a specific time window (i.e., bin) and if a threshold is exceed, the system identifies that a sub-event has occurred. We report evaluation scores as presented in Meladianos et al. (2018). The second approach is the graph-based method of Meladianos et al. (2018). We observe that our baseline system (Section 3.3) has a 1.2% improvement in terms of macro-F_1 and 2.7% improvement in terms of micro-F_1, compared to the graph-based model from Meladianos et al. (2018), mainly due to increased precision, and despite the recall loss.

5.2 Sequence Labeling Results

Table 2 illustrates the predictive performance of our proposed model (i.e., using the chronological

[1] https://github.com/bekou/subevent_sequence_labeling

LSTM) compared to models making independent predictions per bin. The upper part of Table 2 contains models without the chronological LSTM. Our experiments study both *word-level* and *tweet-level* bin representations (see Fig. 1), as reflected in the 'Word' vs. 'Tweet' prefix, respectively, in the Model column of Table 2.

The simplest *word-level* representation uses the tf-idf weighting scheme (as in Pohl et al. (2012)) followed by an MLP classifier. For the other word-level models, we exploit several architectures: AVG pooling (Iyyer et al., 2015), a CNN followed by AVG pooling (Kim, 2014) and hierarchical word-level attention (Yang et al., 2016).

For *tweet-level* representations, we adopt similar architectures, where the AVG, CNNs and attention are performed on sentence level rather than on the word-level representation of the bin. In this scenario, we have also exploited the usage of sequential LSTMs to represent the tweets. When comparing models with and without tweet-level LSTMs, we report the strategy that yields the best results, indicated by ✓ and ✗ in the tweet-level LSTM (TL) columns of Table 2. We do not present results for applying sequential LSTMs on the word-level bin representation, because of slow training on the long word sequences.

Benefit of Chronological LSTM: The bottom part of Table 2 presents the results of the same models followed by a chronological LSTM to capture the natural flow of the stream as illustrated in Fig. 1. We report results as described in Section 4, using the micro F_1 score with the two evaluation strategies (*bin-level* and *relaxed*). We observe that when using the chronological LSTM, the performance in terms of *bin-level* F_1 score is substantially improved for almost every model. Note that the best model using the chronological LSTM (Tweet-AVG) achieves 2.4% better F_1 than the best performing model without the use of chronological LSTM (Word-CNN-AVG). In most cases there is also a consistent improvement for both the precision and the recall metrics, which is thanks to the sequential nature of the upper level LSTM capturing the flow of the text.

Settings	Macro			Micro		
	P	R	F_1	P	R	F_1
Burst	78.00	54.00	64.00	72.00	54.00	62.00
Meladianos et al. (2018)	76.00	75.00	75.00	73.00	74.00	73.00
Our binary classif. baseline	89.70	69.99	**76.16**	83.65	69.05	**75.65**

Table 1: Comparing our neural network binary classification baseline model to state-of-the-art (P = precision, R = recall).

	Model	Bin-level				Relaxed			
		TL	P	R	F_1	TL	P	R	F_1
without chronol. LSTM	Word-tf-idf	-	49.40	52.06	50.69	-	56.10	56.10	56.10
	Word-AVG	-	51.40	45.96	48.53	-	56.10	56.10	56.10
	Word-CNN-AVG	-	56.93	56.01	56.47	-	75.60	75.60	75.60
	Word-attention	-	52.92	58.71	55.66	-	86.59	86.59	**86.59**
	Tweet-AVG	✓	49.04	45.96	47.45	✓	62.19	62.19	62.19
	Tweet-attention	✓	51.99	42.37	46.68	✗	80.48	80.48	80.48
	Tweet-CNN	✗	58.88	51.17	54.75	✗	70.73	70.73	70.73
with chronol. LSTM	Word-AVG	-	58.14	58.35	58.24	-	71.95	71.95	71.95
	Word-CNN-AVG	-	60.89	56.19	58.45	-	60.97	60.97	60.97
	Word-attention	-	52.99	42.90	47.42	-	60.97	60.97	60.97
	Tweet-AVG	✗	57.43	60.32	**58.84**	✗	64.63	64.63	64.63
	Tweet-attention	✓	48.26	52.24	50.17	✗	67.07	67.07	67.07
	Tweet-CNN	✗	65.33	49.73	56.47	✗	60.97	60.97	60.97

Table 2: Comparison of our baseline methods in terms of micro *bin-level* and *relaxed* F_1 score with and without chronological LSTM (see Fig. 1). The ✓and ✗ indicate whether the model uses a tweet-level LSTM (TL).

Limitations of *Relaxed* Evaluation: On the other hand, using the *relaxed* evaluation strategy, we observe that the best models are those without the chronological LSTM layer. Yet, we consider the *relaxed* evaluation strategy flawed for our scenario, despite the fact that it has been used for entity classification tasks (Bekoulis et al., 2018a; Adel and Schütze, 2017). Indeed, it is not able to properly capture sub-events which are characterized by duration: e.g., if a model assigns a different label to each of the bins that together constitute a single sub-event, then this sub-event counts as a true positive based on the *relaxed* evaluation strategy (similar to the evaluation proposed by Meladianos et al. (2018) and followed in Table 1). Thus, in this work, we propose to use the *bin-level* evaluation, since it is a more natural way to measure the duration of a sub-event in a supervised sequence labeling setting. Note that due to the noisy nature of Twitter streams, a tweet sequence spanning a particular sub-event is likely to contain also tweets that are not related to the given sub-event: a given bin inside the event may contain only a minority of tweets discussing the event. Therefore, we consider the standard sequence labeling evaluation (requiring to have types as well as boundaries correct) to be not applicable in sub-event detection.

Performance Comparison of the Top-3 Models: Figure 2 shows the performance of our three best performing models in terms of *bin-level* F_1 score on the validation set. The best performing model is the Tweet-AVG model since it attains its maximum performance even from the first training epochs. The Word-AVG model performs well

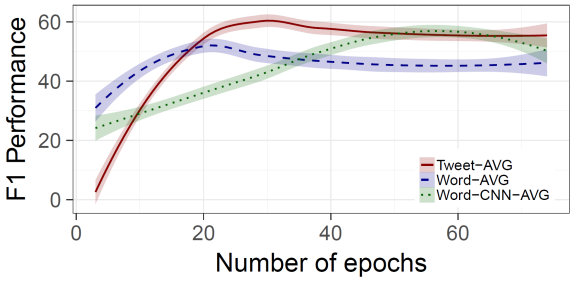

Figure 2: *Bin-level* F_1 performance of the three best performing models on the validation set with respect to the number of epochs. The smoothed lines (obtained by LOWESS smoothing) model the trends and the 95% confidence intervals.

from the first epochs, showing similar behavior to the Tweet-AVG model. This can be explained by the similar nature of the two models. The word-level CNN model attains maximum performance compared to the other two models in later epochs. Overall, we propose the use of the chronological LSTM with the Tweet-AVG model since this model does not rely on complex architectures and it gives consistent results.

6 Conclusion

In this work, we frame the problem of sub-event detection in Twitter streams as a sequence labeling task. Specifically, we (i) propose a binary classification baseline model that outperforms state-of-the-art approaches for sub-event detection (presence/absence), (ii) establish a strong baseline that additionally predicts sub-event *types*, and then (iii) extend this baseline model with the idea of exchanging chronological information between sequential posts, and (iv) prove it to be beneficial in almost all examined architectures.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive feedback. Moreover, we would like to thank Christos Xypolopoulos and Giannis Nikolentzos for providing (i) the Twitter dataset (tweet ids) and (ii) instructions to reproduce the results of their graph-based approach.

References

Dhekar Abhik and Durga Toshniwal. 2013. Sub-event detection during natural hazards using features of social media data. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 783–788, New York, NY, USA. ACM.

Heike Adel and Hinrich Schütze. 2017. Global normalization of convolutional neural networks for joint entity and relation classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1723–1729. Association for Computational Linguistics.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018a. Adversarial training for multi-context joint entity and relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836. Association for Computational Linguistics.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018b. An attentive neural architecture for joint segmentation and parsing and its application to real estate ads. *Expert Systems with Applications*, 102:100 – 112.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018c. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34 – 45.

Carlos Castillo. 2016. *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*. Cambridge University Press.

Guandan Chen, Nan Xu, and Weiji Mao. 2018. An encoder-memory-decoder framework for sub-event detection in social media. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1575–1578, New York, NY, USA. ACM.

Eleonora D'Andrea, Pietro Ducange, Beatrice Lazzerini, and Francesco Marcelloni. 2015. Real-time detection of traffic from Twitter stream analysis. *IEEE Transactions on Intelligent Transportation Systems*, 16(4):2269–2283.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good

the bad and the omg! In *Proceedings of the Fifth International AAAI conference on weblogs and social media*, pages 538–541.

Polykarpos Meladianos, Giannis Nikolentzos, Franois Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. 2015. Degeneracy-based real-time sub-event detection in Twitter stream. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, pages 248–257. AAAI Press.

Polykarpos Meladianos, Christos Xypolopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. 2018. An optimization approach for sub-event detection and summarization in Twitter. In *Proceedings of the 40th European Conference in Information Retrieval*, pages 481–493. Springer International Publishing.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1354–1364. Association for Computational Linguistics.

Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. 2012. Summarizing sporting events using Twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, pages 189–198, New York, NY, USA. ACM.

Daniela Pohl, Abdelhamid Bouchachia, and Hermann Hellwagner. 2012. Automatic sub-event detection in emergency management using social media. In *Proceedings of the 21st International Conference on World Wide Web*, pages 683–686, New York, NY, USA. ACM.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. 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.

P.K. Srijith, Mark Hepple, Kalina Bontcheva, and Daniel Preotiuc-Pietro. 2017. Sub-story detection in Twitter with hierarchical Dirichlet processes. *Information Processing & Management*, 53(4):989 – 1003.

Zhongqing Wang and Yue Zhang. 2017. A neural model for joint event detection and summarization. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4158–4164. AAAI Press.

Chen Xing, Yuan Wang, Jie Liu, Yalou Huang, and Wei-Ying Ma. 2016. Hashtag-based sub-event discovery using mutually generative LDA in Twitter. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2666–2672. AAAI Press.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489. Association for Computational Linguistics.

Siqi Zhao, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. 2011. Human as real-time sensors of social and physical events: A case study of Twitter and sports games. *arXiv preprint arXiv:1106.4300*.

Arkaitz Zubiaga, Damiano Spina, Enrique Amigó, and Julio Gonzalo. 2012. Towards real-time summarization of scheduled events from Twitter streams. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, pages 319–320, New York, NY, USA. ACM.