Exploring Social Media for Event Attendance

Vinicius Monteiro de Lira[1,3,4], Craig Macdonald[2], Iadh Ounis[2],
Raffaele Perego[1], Chiara Renso[1] and Valeria Cesario Times[3]
[1]ISTI-CNR, Italy – [2]University of Glasgow, UK – [3]UFPE, Brazil – [4]University of Pisa, Italy
{vcml, vct}@cin.ufpe.br, {craig.macdonald, iadh.ounis}@glasgow.ac.uk, {renso, perego}@isti.cnr.it

***Abstract*—Large popular events are nowadays well reflected in social media fora (e.g. Twitter), where people discuss their interest in participating in the events. In this paper we propose to exploit the content of non-geotagged posts in social media to build machine-learned classifiers able to infer users' attendance of large events in three temporal periods: before, during and after an event. The categories of features used to train the classifier reflect four different dimensions of social media: textual, temporal, social, and multimedia content. We detail the approach followed to design the feature space and report on experiments conducted on two large music festivals in the UK, namely the VFestival and Creamfields events. Our attendance classifier attains very high accuracy with the highest result observed for the Creamfields dataset ∼87% accuracy to classify users that will participate in the event.**

I. Introduction

Large events like music festivals or religious celebrations attract thousands of participants, and hence they are usually well reflected in social media, where interested users express, through *posts*, their feelings, experiences or opinions about such events. One interesting analysis that can benefit several applications like advertising or mobility management is to detect, from the analysis of these posts, the actual user attendance to the event. We propose a machine learning approach for analysing social media posts of users discussing the event to infer their actual attendance.

This task could be simple when only using the geotagged posts since the "check-in" or the user position can be trivially associated to the event location. However, very few posts of social media make use of accurate geotagging (on Twitter it is about 2% on average [1]) and this would therefore result in a very sparse dataset. The challenge we address here is to predict the actual attendance of users to the mentioned event by using non-geotagged posts.

We distinguish three temporal intervals when the posts have been shared on social media in reference to the event: *before*, *during* or *after* the event. The posts shared before the event may express the interest of the users in the upcoming event and their intention to attend. During the event, people may express their feelings about the event, may report issues with the provided services or may also share photos and videos. After the event, users may share photos and videos or report feelings and comments on their past experience about the event.

In deploying a classifier for each of the three temporal intervals of the event, we make use of four categories of features representing various aspects of social media: textual, temporal, social & multimedia content. Hence, the contributions of this paper are two-fold: (1) We propose the task of predicting the attendance of users at specific events, based on non-geotagged posts; (2) We instantiate this task into three classification tasks tailored for posts made before, during and after the event. Our experiments are conducted with Twitter datasets related to two large UK music festivals. We generate our ground truth and use it to train different supervised classification models using four categories of features[1]. The experimental results demonstrate that our classifiers can detect event attendance quite effectively, achieving accuracies ranging from 80% to 87% for the three classification tasks and the two datasets.

In the remainder of this paper, we describe related work in Section II. In Section III, we introduce our approach for classifying attendance and the features used to train suitable classifiers, while in Section IV the accuracy of each classifier is reported and analysed. Section V provides concluding remarks.

II. Related Work

Many papers tackle the problem of estimating the current location of users or their home from non geo-located tweets [2], [3], [4], [5], [6]. Compared to these proposals, we have a different objective as we do not want to estimate the user's exact location at the time of the post, but classify the single posts on the basis of the user's future, current and past attendance to a given event. Events in social media have been extensively studied. The main aspects investigated in the literature are: (1) prediction of events attendance in Event-Based Social Networks (EBSN) [7], [8], [9]; (2) recommendation of events to users [10], [11]; and, (3) estimation of the number of attendees in a given event [12]. Du et al. [7] analyse an EBSN to predict users' attendance by taking into account the content, the spatial and temporal context, the users' preferences and their social influence. Zhang et al. [8] propose a supervised learning model to predict event attendance based on semantic, temporal, and spatial features. [9] address the extent to which geospatial, temporal, and social factors influence the users' preferences towards events formulating a predictive modeling task trying to match a user's mobility profile against the collective past Foursquare check-in activity of potential event attendees. Compared to these approaches we do not specifically deal with ESBN but instead focus on popular social media where events can have an "echo". We do not use users' history or preferences as we aim at classyfing single posts by disregarding the user profiles and specific event information.

Within the second category, event recommendation, [10], [11] and [13] address the challenge of recommending events

[1]downloadable from: https://github.com/viniciusmonteiro/asonam2017

within event-based social networks (EBSNs). Each of these approaches is challenged by the cold-start problem, and recommendation evidence may resort to the events that are geographically closest [10]. Our work is complementary with respect to these approaches since we are interested in identifying the posts related to event attendance rather than in making recommendations. In any case our approach could allow to identify more precisely the target users for recommendations.

Finally, within the third category of related works [12] investigates whether mobile phone usage and the geolocated Twitter data can be used to estimate the number of people in a specific area at a given time. In [14], [15], the authors describe a methodology for identifying the user behavior and mobility patterns of Instagram social network users visiting the EXPO 2015 world fair in Milan, Italy and the FIFA World Cup 2014. A key difference of our approach is that we do not use geotagged information to infer attendance.

III. Classifying Event Attendance

In the real world, an *event* is something that occurs in a certain place during a particular interval of time. The place where the event occurs can be associated with its geographical coordinates (<lat, long>), and the temporal duration, which may vary from minutes to days or weeks. A social media *post* by a user u, may contain text, links, emoticons, photos and/or videos (depending on the specific social network), as well as the timestamp at which the post was created and a social component representing the relations of u with other users (likes, followers, retweets, etc). We define an *event-related post* p as any post that mentions one or more event identifiers and is thus possibly related to the specific event considered. We distinguish these event-related posts as occurring *before the event* when posted in a date before the starting date, *during the event* when posted between during the event, and *after the event* when posted after the event.

Our work aims at understanding if these weak and noisy expressions of interest occurring in event-related posts can be exploited to identify the users who are likely to attend an event and discriminate them ones who are not going to attend it. Specifically, we propose to use supervised machine learning approaches to train binary classifiers that can automatically distinguish between posts of attendees and non-attendees. We instantiate our classifier in three different tasks referred to posts published *before, during,* or *after* the date of the event.

We exploit four different categories of features: textual, temporal, social, and multimedia dimensions.

Textual features model the textual content of the post. We used a Bag of Words (BoW) model with unigrams, bigrams and trigrams occurring in the post. We apply lemmatization to group together the different inflected forms of a word. Thus each lemma and each sequence of two and three adjacent lemmas are considered as features.

Temporal features represent the time of the post with respect to the event.

Social features characterize the social profile of the posting user. Our social features are the number of followers, the number of followees and the ratio between them. We notice that users with a high number of followers and a relatively low number of followees are typically sponsors, organizers or VIPs who not necessarily attend the event.

Finally, the *multimedia content* features identify when a post has any multimedia content, such as a photo, video or a link to visual content posted in other social network like Facebook or Instagram.

IV. Experimental Results

We instantiate our attendance classifier in a scenario that considers large, popular music festivals. In particular, our experiments are organized into two research questions:

RQ1: How accurate is our event attendance prediction classifier? (Sections IV-B)

RQ2: What features groups help most to attain high prediction accuracy? (Section IV-C).

Before addressing RQ1 and RQ2, we first discuss the setup for our experiments.

A. Experimental Setup

Our experiments are conducted using Twitter posts about two premier UK music festivals: Creamfields 2016 (held in Daresbury, UK, on August 25th-28th), and VFestival 2016 (held in Chelmsford/South Staffordshire, UK, on August 20th-21st). We collected tweets using both the Streaming API and REST API provided by Twitter. To obtain tweets related to the VFestival, we used the terms 'vfest' and 'v21st' as identifiers, while for the Creamfields event we used its own name. We used the Streaming API from August 10th to September 15th 2016, while the REST API was used to also collect the available past tweets from March 1st to September 15th 2016. Tweets generated by the official accounts of the events (@vfestival and @Creamfields) were removed as not potential attendees.

Aligned with our proposed three attendance classifier tasks, for each respective event the collected tweets are split on the basis of their timestamp into three different disjoint sets: *posts made before, during or after the event*. To generate our training set, we randomly sample (without replacement) 460 distinct tweets for each task from each dataset, thus 1,380 tweets in total for each festival. Then, for each of the three tasks, a binary label is assigned to each tweet (positive class: a user who intends/is/has attended, and vice versa for the negative class). This human assessment is based on the textual or visual content of the tweet that provide any explicit evidence of attendance at the event to be established. Any other kind of interpretation (advertisement, sale of tickets, general information, regrets or impossibility, etc) were labelled as negative. Specifically, for each dataset and task, Table I reports the total number of labeled tweets, the respective percentage of positive and negative labeled cases, the total number of tweets collected, and the number of distinct active users.

Our experiments are conducted using a 5-fold cross validation, while preserving the proportion of positive and negative instances in each fold.

TABLE I
DATASET STATISTICS.

Dataset	Task	Labeled	pos%	neg%	Tweets	Users
Creamfields	**Before**	460	48.3	51.7	24,963	11,700
	During	460	39.1	60.9	25,625	15,884
	After	460	69.3	30.7	29,801	17,850
VFestival	**Before**	460	47.6	52.4	10,754	6,513
	During	460	37.4	62.6	4,873	3,285
	After	460	67.2	32.8	26,027	14,744

For each task and dataset we trained three different classification models using: Gradient Boosting Decision Trees (GBDT), Logistic Regression (LR) and Random Forest (RF). All these algorithms, chosen among those consistently delivering state-of-the-art performance in text classification tasks [16], are available in the scikit-learn library[2] used to train our classifiers.

B. Results: RQ1

In this section we address RQ1 by comparing the effectiveness of our classifiers with those of two baselines: (a) *Naive Bayes*, a Naive Bayes classifier trained on textual features only); (b) *Occurrences*, a baseline constructed by considering positive and negative lists of words. The positive list contains the words occurring in positive tweets of our training set. Instead, the negative list considers the negative tweets. This baseline assigns the positive label to a tweet if its text contains more words matching the positive list of words, negative otherwise. The most frequent class is assigned to ties.

Table II reports the performance of the classifiers for each dataset and classification task (*before*, *during*, *after*). For the classifiers reported in this table, all features groups are used, with the textual content of posts represented according to the BoW model. Classification performance is measured in terms of accuracy, precision, recall, and F1 measures.

On analysing the results in Table II, we find that at least one of our classifiers attain effectiveness higher than the baselines for all tasks, with the highest performance observed on the Creamfields dataset ($\sim$87% accuracy at classifying users that will participate in the event) using GBDT. For posts made during the event, GBDT obtained an accuracy of $\sim$82% when classifying the attendance of the users at the Creamfields and also when inferring past attendance at VFestival. We highlight the high precision ($\sim$87%) achieved by LR on classifying tweets of attendees posted before the Creamfields festival. The performance achieved with RF on the VFestival dataset for the during task is also impressive: nearly every positive attendance case is correctly identified (recall $\sim$99%).

In summary, for RQ1, the accuracy results reported in Table II show that our approach is effective at classifying user attendance. We observe that GBDT on average outperforms the other algorithms and LR achieves the best accuracy in one of the six cases.

[2]http://scikit-learn.org/

C. Results: RQ2

In this section, we address our second research question, concerned with the contribution of the feature groups defined in Section III. In addressing this research question, we aim to understand which feature groups provide the most benefit to performance, across the three attendance classification tasks.

In answering RQ2, we consider only the GBDT classifier, which, according to the results reported in Section IV-B, achieves, on average, the highest performance. To evaluate the contribution of each group of features, we conduct an ablation study, i.e. remove each group of features from the learned classifier model on all features. Table III reports the results of the ablation study sorted by accuracy for each of the *before*, *during* and *after* classification tasks. The representation 'all-text' means we exclude the textual features, and so on.

On analysing Table III, we find, for the *during* task, that the multimedia features are important for attaining high accuracy, particularly for the VFestival, where a $\sim$5% drop in accuracy is observed when the multimedia feature group is ablated ($0.802 \rightarrow 0.757$). Indeed, in this dataset, for example, we have that around 0.85%, 22% and 27% of the tweets posted, respectively before, during and after the event has some multimedia content. For the Creamfields the corresponding percentages containing multimedia content are: 0.4%, 8% and 20%, respectively.

Next, we note that social features exhibit usefulness for the before task in Creamfields and for the after tasks on VFestival, where the exclusion implies loss of accuracy. We postulate that these features allow to identify (negative) advertisement posts coming from event sponsors or news providers, all of whom have high number of followers.

Temporal features are important when classifying attendance after the completion of the event. Indeed, we note that low values for this feature (i.e. shorter difference between the dates before or after the event) are indicative for identifying the actual attendees of the event, while higher values (distant from the event) are indicative for identifying non-attendees.

The users express their attendance in an event through the post text in different ways depending on the period (before, during, after). Hence, the textual features extracted from posts vary depending on the task. As we can see from the table, textual features are the most important for the *before* and *after* tasks. For these tasks, in both datasets, once we exclude those features, the accuracy drops tightly. Before the event, the users mention often their participation by posting about the purchase and delivery of their tickets. After the event, the users express their experience, how they feel after the event and state willingness to come back to next edition.

Lastly, the meta textual content (number of *words, hashtags, mentions, URLs* and *emoticons*) only exhibit importance for attaining accurate classifications for the before task of the VFestival. For the same festival and for the after task, these features introduce noise into the GBDT model, since exclusion of this set of features marginally improves the accuracy of the model.

Finally, and to summarise our findings for RQ2, we find that while each of the features groups has some impact for at least one of the tasks, we highlight again the usefulness

TABLE II
CLASSIFICATION EFFECTIVENESS USING BOW COMPARED TO NAIVE BAYES AND OCCURRENCES BASELINES.

	Dataset: Creamfields					Dataset: VFestival				
Task	**Model**	**Accuracy**	**Precision**	**Recall**	**F1**	**Model**	**Accuracy**	**Precision**	**Recall**	**F1**
Before	Naive Bayes	0.868	0.824	**0.934**	0.876	Naive Bayes	0.779	0.730	**0.858**	**0.786**
	Occurrences	0.752	0.764	0.732	0.746	Occurrences	0.672	0.676	0.604	0.636
	$GBDT_{BoW}$	**0.874**	0.846	0.912	**0.878**	$GBDT_{BoW}$	**0.809**	**0.802**	0.768	0.784
	LR_{BoW}	0.868	**0.870**	0.870	0.868	LR_{BoW}	0.761	0.744	0.762	0.748
	RF_{BoW}	0.819	0.830	0.808	0.816	RF_{BoW}	0.746	0.766	0.676	0.716
During	Naive Bayes	0.759	0.780	0.544	0.638	Naive Bayes	0.752	0.778	0.472	0.582
	Occurrences	0.700	0.630	0.616	0.618	Occurrences	0.694	0.618	0.508	0.552
	$GBDT_{BoW}$	**0.817**	**0.830**	**0.616**	**0.708**	$GBDT_{BoW}$	**0.802**	0.850	0.582	**0.688**
	LR_{BoW}	0.741	0.766	0.538	0.602	LR_{BoW}	0.626	0.600	**0.614**	0.494
	RF_{BoW}	0.770	0.812	0.560	0.652	RF_{BoW}	0.783	**0.908**	0.470	0.612
After	Naive Bayes	0.791	0.784	0.968	0.868	Naive Bayes	0.809	0.802	**0.950**	**0.870**
	Occurrences	0.700	0.786	0.782	0.786	Occurrences	0.650	0.768	0.682	0.722
	$GBDT_{BoW}$	0.780	0.792	0.948	0.864	$GBDT_{BoW}$	**0.815**	**0.824**	0.902	0.862
	LR_{BoW}	**0.813**	**0.810**	0.958	**0.880**	LR_{BoW}	0.809	0.812	0.932	0.868
	RF_{BoW}	0.763	0.752	**0.982**	0.852	RF_{BoW}	0.783	0.780	0.948	0.854

TABLE III
ACCURACIES OF GBDT MODELS BY ABLATING GROUPS OF FEATURES.

	Dataset: Creamfields		Dataset: VFestival	
Task	**Group**	**Accuracy**	**Group**	**Accuracy**
Before	all	0.874	all	0.809
	all-temporal	0.874	all-social	0.809
	all-multimedia	0.874	all-textual_meta_feats	0.809
	all-textual_meta_feats	0.865	all-multimedia	0.794
	all-social	0.863	all-temporal	0.792
	all-text	0.606	all-text	0.656
During	all	0.817	all-textual_meta_feats	0.806
	all-textual_meta_feats	0.815	all	0.802
	all-social	0.811	all-text	0.802
	all-multimedia	0.804	all-social	0.791
	all-text	0.667	all-multimedia	0.757
After	all-social	0.793	all	0.815
	all-textual_meta_feats	0.787	all-textual_meta_feats	0.811
	all	0.780	all-temporal	0.809
	all-temporal	0.780	all-social	0.807
	all-multimedia	0.769	all-multimedia	0.781
	all-text	0.689	all-text	0.724

of the textual features for the prediction of attendance before and after the event. Indeed, when this group is ablated from the model, the classification accuracy decreases remarkably on both datasets. This observation suggests, as future work, attempts for improvements of the results by enriching the group of textual features.

V. CONCLUSIONS

In this paper, we proposed a classification approach to infer event attendance from users media posts. A key detail of our proposed approach is that our inference is done by classifying the non-geotagged content of the users' posts.

We trained machine-learned classifiers using tweets related to two large music festivals in the UK, and we evaluated their accuracy and precision in comparison to two classical baselines and we also highlighted the most informative group of features. The results show how our approach performs consistently better than the baselines, exhibiting 87% accuracy at classifying users that have indicated their intention to attend the event. As future work, we aim to improve our results by enriching the group of textual features and by extracting information from the visual content of the published photos or videos.

Acknowledgements. This work is partially supported by EU H2020 project INFRAIA-1-2014-2015 *SoBigData* (G.A. 654024) and *BASMATI* funded by ICT R&D program of the Korean MSI P/IITP (R0115 - 16 - 0001) and EU H2020 programme (G.A. 723131).

REFERENCES

[1] K. Leetaru, S. Wang, G. Cao, A. Padmanabhan, and E. Shook, "Mapping the global twitter heartbeat: The geography of twitter," *First Monday*, vol. 18, no. 5, 2013.

[2] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: A content-based approach to geo-locating twitter users," in *ACM CIKM '10*, 2010, pp. 759–768.

[3] H.-w. Chang, D. Lee, M. Eltaher, and J. Lee, "@phillies tweeting from philly? predicting twitter user locations with spatial word usage," in *ASONAM '12*, 2012.

[4] J. Mahmud, J. Nichols, and C. Drews, "Home location identification of twitter users," *ACM TIST*, vol. 5, no. 3, pp. 47:1–47:21, Jul. 2014.

[5] K. Lee, R. K. Ganti, M. Srivatsa, and L. Liu, "When twitter meets foursquare: Tweet location prediction using foursquare," in *MOBIQUITOUS '14*, 2014.

[6] S. Kinsella, V. Murdock, and N. O'Hare, "I'm eating a sandwich in Glasgow: Modeling locations with tweets," in *SMUC 2011*, pp. 61–68.

[7] R. Du, Z. Yu, T. Mei, Z. Wang, Z. Wang, and B. Guo, "Predicting activity attendance in event-based social networks: Content, context and social influence," in *UbiComp '14*. ACM, 2014, pp. 425–434.

[8] X. Zhang, J. Zhao, and G. Cao, "Who will attend? predicting event attendance in event-based social network," in *IEEE MDM*, 2015.

[9] P. Georgiev, A. Noulas, and C. Mascolo, "The call of the crowd: Event participation in location-based social services," in *ICWSM*, 2014.

[10] D. Quercia, N. Lathia, F. Calabrese, G. D. Lorenzo, and J. Crowcroft, "Recommending social events from mobile phone location data," in *ICDM*, 2010.

[11] A. Q. Macedo, L. B. Marinho, and R. L. Santos, "Context-aware event recommendation in event-based social networks," in *RecSys*, 2015.

[12] F. Botta, H. S. Moat, and T. Preis, "Quantifying crowd size with mobile phone and Twitter data," *Royal Society open science 2.5*, 2015.

[13] S. Wang, Z. Wang, C. Li, K. Zhao, and H. Chen, "Learn to recommend local event using heterogeneous social networks," in *APWeb*, 2016.

[14] E. Cesario, A. R. Iannazzo, F. Marozzo, F. Morello, G. Riotta, A. Spada, D. Talia, and P. Trunfio, "Analyzing social media data to discover mobility patterns at EXPO 2015: Methodology and results," in *HPCS*, 2016.

[15] E. Cesario, C. Congedo, F. Marozzo, G. Riotta, A. Spada, D. Talia, P. Trunfio, and C. Turri, "Following soccer fans from geotagged tweets at FIFA World Cup 2014," in *Proceedings of ICSDM*, 2015, pp. 33–38.

[16] C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in *Mining Text Data*. Springer, 8 2013, pp. 163–222.