Event attendance classification in social media

Vinicius Monteiro de Lira[a,c,d], Craig Macdonald[b], Iadh Ounis[b], Raffaele Perego[c], Chiara Renso[c], Valeria Cesario Times[a]

[a] *Federal University of Pernambuco, Brazil*
[b] *University of Glasgow, UK*
[c] *ISTI-CNR, Italy*
[d] *University of Pisa, Italy*

Abstract

Popular events are well reflected on social media, where people share their feelings and discuss their experiences. In this paper, we investigate the novel problem of exploiting the content of non-geotagged posts on social media to infer the users' attendance of large events in three temporal periods: before, during and after an event. We detail the features used to train event attendance classifiers and report on experiments conducted on data from two large music festivals in the UK, namely the VFestival and Creamfields events. Our classifiers attain very high accuracy with the highest result observed for the Creamfields festival ($\sim$91% accuracy at classifying users that will participate in the event). We study the most informative features for the tasks addressed and the generalization of the learned models across different events. Finally, we discuss an illustrative application of the methodology in the field of transportation.

Keywords: Social media analysis, event attendance prediction, classification.

1. Introduction

An unprecedented amount of user-generated content about human activities has been created through the introduction of popular social media applications on smart-

Email addresses: vcml@cin.ufpe.br (Vinicius Monteiro de Lira), craig.macdonald@glasgow.ac.uk (Craig Macdonald), iadh.ounis@glasgow.ac.uk (Iadh Ounis), raffaele.perego@isti.cnr.it (Raffaele Perego), chiara.renso@isti.cnr.it (Chiara Renso), vct@cin.ufpe.br (Valeria Cesario Times)

phones (e.g. Facebook, Foursquare, Instagram, Twitter) [1, 2]. The diversity and variety of content shared on these online social networks bears witness to the richness of these new forms of social interactions constituting nowadays an important source for an unprecedented outreach, speed and democratization of communication. Due to this vast applicability, social media analytics is a fast growing research area [3]. Social media can be exploited to extract valuable information concerning human dynamics and behaviors [4, 5, 6]. As a consequence, they can play a role in understanding modern life, including transportation [7, 8, 9] and human mobility [4, 10, 11].

Music festivals, like many other popular events (e.g., important religious celebrations or sports matches) attract thousands of participants. Usually, they are well reflected in social media networks, allowing people to connect with "the event", expressing through *posts* their feelings, experiences or opinions well in advance of its planned date.

Given the attention to popular events reflected in social media, this paper addresses a novel challenging problem: *"Is it possible to infer from Twitter posts the actual attendance of the user to the cited event?"*. If we could classify user posts discussing an event on the basis of the actual attendance of the user to the event, we could enable or enhance several practical applications in the fields, for example, of targeted advertising and mobility management.

The simplest way of inferring the presence of users at events is to consider the geotag associated with their posts: the "check-in" or the user location in the event place at the time of the event can indeed be trivially associated with attendance. We observe however that this approach suffers from two drawbacks.

The first drawback is that very few social media users enable the geotagging of their posts (in Twitter the percentage of geotagged posts is about 2% [12, 13]). In fact, geolocation information is geographically accurate, but represents a very sparse data source. Learning attendance prediction classifiers based on sparse data would be extremely difficult and may lead to ineffective predictive models.

The second drawback of only using geolocated data is that they do not represent the intention of the user to participate in the event. Indeed, our aim is to infer the user's intention of participating to the event even before the event takes place.

To overcome these two aforementioned issues, we take a different direction by addressing the novel challenge of inferring the actual attendance of users to a mentioned event by only relying on the content of *non-geotagged posts*, without considering any spatial features. Moreover, we perform our event attendance classification by distinguishing three temporal intervals identifying when the posts have been shared on social media: *before*, *during* or *after* the event. We propose three distinct classification tasks, one for each temporal interval. The analysis of posts shared before the event acts as a prediction for the users' actual attendance, the analysis of posts shared during the event reflects the actual participation of users at the event, while the analysis of posts shared after the event gives a view or a summary of past attendance.

The "before" case is particularly interesting, since an early knowledge of the possible user attendance can be useful to enable innovative services and applications. For example, event organizers or third-party companies could precisely target their advertisement campaigns by offering specific or personalized services to the users predicted to participate in the event. Another example is in the field of transportation planning, where attendance prediction could allow the organizers or the local authorities to push potential attendees to use public transportation or can help bus and shuttle companies to plan and advertise collective transport services to the event [14]. 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 about the event. After the event, users may report feelings and comments on their experience at the event. Knowledge of the social media users who attended (or did not attend) the event can be very useful as well. Their posts can be used for example to understand attendees' profiles and provide insights, allowing to improve the future editions of the event. In addition, this knowledge can also be used to help in estimating crowd sizes and support transportation planning for the future version of the event [15].

This paper extends a previous work where we initially investigated the supervised training of classifiers aimed at inferring the attendance of users to two musical festivals: VFestival and Creamfields, two large popular events in the UK [16].

The preliminary results achieved highlighted that the features extracted from the textual content are those playing the most important role for attaining a good classification accuracy. This interesting insight encouraged us to explore in more in-depth these features and to study, in the present paper, how to enrich the textual features to help generalize across different events the learned classification models. The present work, apart from being built on the previous research, proposes several new original contributions:

- we investigate the use of word-embedding features to improve the prediction accuracy and the robustness of the attendance classifiers when trained in different events;
- we discuss the new experiments conducted for assessing the improved accuracy and the ability of our classifiers to generalize across different datasets;
- we validate the labelling process and the classifiers' performances by using a second ground truth dataset consisting of posts published by users in which they share their position;
- we investigate the expressions most commonly used by users to convey attendance or not at a large event;
- we introduce and discuss a real-world application of the proposed attendance classifiers for organizing the transportation towards the events.

The remainder of this paper is organized as follows. We describe related work in Section 2. In Section 3, we introduce our approach for classifying attendance, we detail the features used to train suitable classifiers and highlight our research questions. Section 4 assesses the performance of our classifiers and answers the research questions by discussing the experimental results achieved. In Section 5, we provide an example application of the deployed classifier for transport planning, while Section 6 provides concluding remarks and plans for future work.

2. Related Work

Many papers tackled the problem of estimating the current location of users or their home from non geo-located tweets [17, 18, 19, 20, 21, 22, 23, 24]. Compared to these proposals, we have a different objective as we do not want to estimate the exact user location at the time of the post. Instead, we aim to classify the single posts on the basis of a user' 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), such as Meetup and Plancast, and Location-Based Social Networks (LBSN), such as Foursquare [25, 26, 27, 28, 29]; (2) recommendation of events to users [5, 30, 31]; (3) estimation of the number of attendees in a given event [32]; and (4) modeling participants' behavior during an event. [33, 4, 34]

Du et al. [25] analyzed an EBSN to predict users' attendance by taking into account the content, the spatial and temporal context, the users' preferences and their social influence. They used a Singular Value Decomposition with Multi-Factor Neighborhood (SVD-MFN) algorithm to predict activity attendance on the Douban Events network. Zhang et al. [26] proposed a supervised learning model to predict event attendance based on semantic, temporal, and spatial features, representing how frequently and when users have attended similar events in the past, the semantic similarity between events, the location preference when attending events and the home location of the user. They trained three classifiers on a Meetup dataset with semantic descriptions of all organized events. Georgiev et al. [27] addressed 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. Zhang and Lv [35] proposed a group-based social influence propagation network to model group-specific influences on events. In [36], the same authors extended the previous work by proposing a group-based event participation prediction framework that embeds and connects group context features and social-related features using historical event attendance logs. They extracted the group-based social features by using a hybrid event-group/category-user network that captures intrinsic social relationships. Their results showed that these features are important for predicting event participation. In [28], the authors proposed a classification task for inferring the response to Facebook event invitations by using data collected by a Facebook soccer application. Their classifiers used not only user-level features but also network-based friendship information. Their experiments suggested that the use of network-based features is very important since it allows to increase the AUC from 0.22% to 0.82%. The approach presented in [29] aims to predict the response to event invitations in the Meetic social network. The authors proposed and evaluated a competing risk methodology for the task showing how their method performs better than the baselines. Note that in the two above cases, there is not evidence showing if a person actually participated in the event or not but the prediction simply infers the response to the invitation on the social media.

Compared to these approaches, we do not specifically address EBSNs and LBSNs, but instead focus on a popular social media platforms where events can have a large "echo". We do not exploit users' history or friendship relations as we aim to classify single posts by their content, completely disregarding the user profile and specific events information. Furthermore, we aim to directly predict the attendance to the event by considering three different time-frames, rather than the user's interest in the event without indication of her actual attendance.

Within the second category, event recommendation, papers [5, 30, 31, 37] and [38] addressed the problem of recommending events 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 to users [5]. The works in [31] and [39] studied the influence of social groups to improve the event recommendation performance. Gao et al. [31] proposed a new Bayesian latent factor model that combines social group influence and individual preferences for event recommendation. In turn, Liu et al. [39] proposed a collective pairwise matrix factorization model to estimate users' pairwise preferences on events, groups and locations. Macedo et al [30] proposed a recommendation approach that leverages multiple context-aware recommendation models for learning to rank events. They exploited features based on group memberships, location signals based on the users geographical preferences, and temporal signals derived from the users time preferences. In [37], the authors considered also the capacity of an event to limit the number of users for the recommendation. Their objectives was to coordinate the user arrangements among the recommended events to attain a balanced event participation. The works in [38] and

[40] focused on an efficient and scalable learning technique for event recommendation to handle large-scale, streaming data. 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 might permit the more precise identification of target users for recommendations.

Within the third category of related works, Botta et al. in [32] investigated 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 considering two case studies of access-restricted areas in Italy: a stadium and an airport (where there were ground truth visitor statistics), they concluded that geolocated tweets with mobile phone data could be a good proxy for estimating the number of users. Sinnott and Wang provided solutions to estimate the population of suburbs and skyscrapers through the use of geo-tagged Twitter data [41]. They constructed linear models for suburbs of four cities and investigated spatial correlation properties between the geo-tagged tweets and the official Census data. Their results showed that Twitter can be used for micro-population estimation with quantifiable degrees of accuracy. In [15], the authors proposed a regression model to estimate the number of attendees from the quantity of geo-tagged tweets posted at an event. They applied the prediction model to estimate the attendance at the Melbourne marathon.

Finally, in the last category of works, the authors of [4, 34, 33] described a methodology for identifying the user behavior and mobility patterns of the Instagram social network users visiting the EXPO 2015 world fair in Milan and the FIFA World Cup 2014. They analyzed how the number of visitors changes over time, identified the most frequent sets of visited pavilions, which countries the visitors came from, and the main destinations of foreign visitors to Italian regions and cities after their visit to EXPO 2015. They also analyzed geotagged tweets of people attending the 2014 FIFA World Cup identifying the most frequent movements of fans, the number of matches attended by groups of fans, the clusters of most attended matches, and the most frequented stadiums.

These latter two groups of works have similar objectives to our aim in studying the social media users' actual participation in events. However, the main differences are threefold: (1) We do not use geotagged information, but we rely on the media posts content to infer users' participation in events. Compared to the related works based on geotagged data, we explore a higher number of posts about the event since a low percentage of the social media posts are geotagged. For example, as mentioned before, on Twitter, around 2% of the tweets are geotagged[1]; (2) we are not interested in estimating the number of participants or crowd, but instead we aim to identify specific social media users who are likely to be – or have been – present at an event. Our approach can thus provide useful and complementary information to support both applications of crowd behavior modeling and crowd size estimation in large events; and (3) we do not recommend participation but instead we infer current, future or past attendance of users based on their media posts. In the next section, we define our classification tasks, as well as the features used for training the attendance classifiers and the research questions we address.

3. Classifying Event Attendance

In the real-world, an *event* is something that occurs in a certain place during a particular interval of time. The location where the event occurs can be associated with its geographical coordinates (<lat, long>), while the temporal duration, which may vary from minutes to days or weeks, can be represented by a time window between a start time t_{start} and an end time t_{end}. In this work, we are interested in large events with thousands of participants. It is customary that such events have an associated entity in the most popular social media platforms (e.g. a Twitter account, a Facebook page), as well as a way of identifying discussions about them through the mentions of one or more *event identifiers* $i_1, \dots i_n$, e.g., the event name, its acronym, some official or popular hashtags, etc.

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

[1] http://firstmonday.org/ojs/index.php/fm/article/view/4366/3654

users (likes, followers, retweets, etc). In addition, some social networks permit the optional enrichment of the post with geotags, giving the <lat, long> position of the user when the post is made.

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 being considered. We distinguish these event-related posts as occurring *before the event* – when posted in a date before t_{start}, *during the event* – when posted between t_{start} and t_{end}, and *after the event* – when posted after t_{end}. Hereinafter, we will simply use the generic term *posts* to refer to event-related posts.

Our intuition is that the nature of event-related posts from attendees differ depending on *when* the posts are created. For instance, posts created before the event may express the users' intention to participate, or their regret for not being able to attend the event or regarding ticket sales. In contrast, posts published during the event may contain brief live reports from the event itself by the participating users, while non-attendees may express regrets for not being there, or comments about the coverage of the event on traditional or social media channels. After the event, attendees may share their opinions about the event, for example wishing to return to the event soon, while non-attendees may hope to participate in the next edition of the event. In Section 3.1, we illustrate these behaviors by providing some real-world examples of event-related posts. Later, in Section 4.4, we validate these behaviors by analyzing the expressions most commonly used by users to positively or negatively convey event attendance.

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 distinguish them from those users that participate actively to the discussion about the event in social media but are not planning to attend it. In this last category we include user accounts directly linked to the event organization, as well as sponsors, advertisers and spammers. We propose to use supervised machine learning approaches to train binary classifiers that can automatically distinguish between posts of attendees and non-attendees. In order to consider the temporal dimension, we instantiate our attendance classification problem in three different tasks for the prediction of user attendance on the basis of posts published *before, during,* or *after* the date of the event.

3.1. Illustrating classification tasks Before/During/After the event

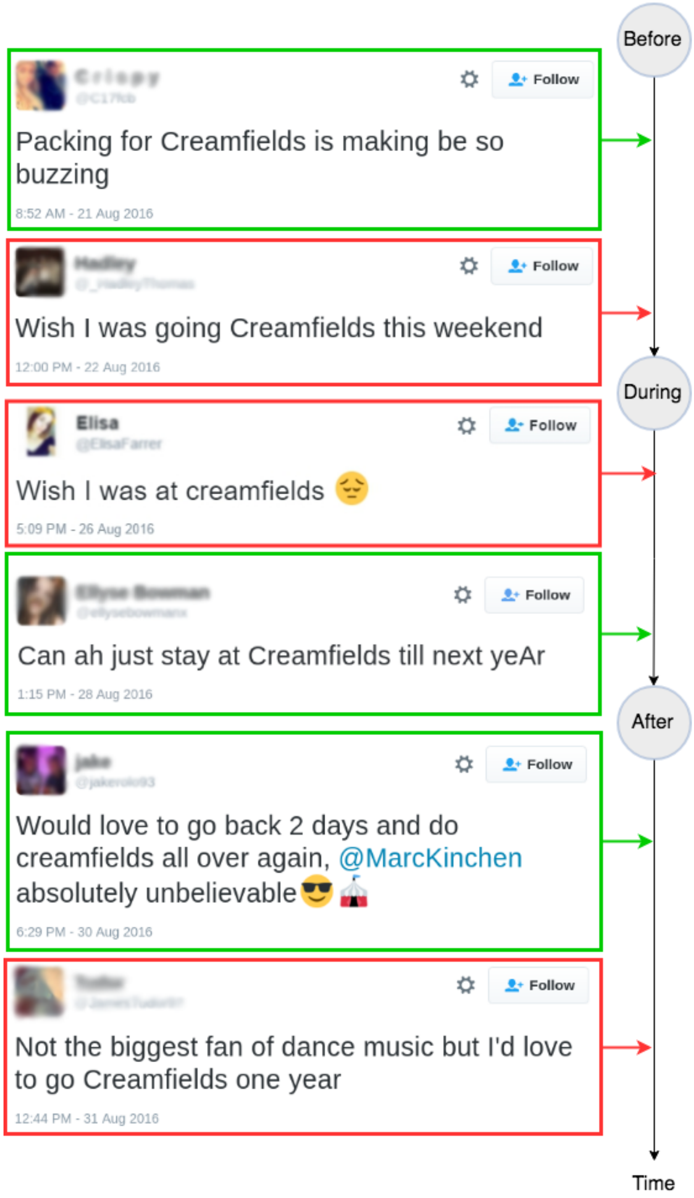
We argue that the types of posts made by users before, during or after an event tend to differ, and different classification models are necessary to attain an accurate classification of these posts.

Before Task: classifying attendance before the event. This task aims at predicting the attendance of a user at the event based on his or her shared posts at a time before the event. The classifier in this case exploits the content of posts where the users implicitly or explicitly express their intention to attend or not the event. Sometimes they explicitly share their intention to go with the words “Go” or “Packing” showing their intention to attend the event. Other common posts that might be considered as members of the negative class are those created by organizers, sponsors, or ticket sellers to provide general information about the event or advertisement and marketing material.

During Task: classifying attendance during the event. The aim of this task is to identify the users who, in the time window of the event, express their presence at the event. Very often, social media users express their actual participation in the event by posting photos or making comments about their experience during the event. On the other hand, non-attendees post general comments about their regrets for not attending or missing the event, or general comments without an explicit attendance meaning.

After Task: classifying attendance after the event. After the event is concluded, people often comment, express their opinions or publish memories and photos on social media. By inspecting such posts, it is often possible to obtain a clear determination of the user’s attendance of the past event (positive) or not (negative).

Figure 1 shows some illustrative examples taken from our dataset related to a large UK music event (the Creamfields festival, see Section 4.1). From the content of the tweets reported in the figure, we can easily distinguish the positive (in green) and negative (in red) attendance cases for the *before*, *during* and *after* tasks.



Packing for Creamfields is making be so buzzing

8:52 AM - 21 Aug 2016

Hadley

Wish I was going Creamfields this weekend

12:00 PM - 22 Aug 2016

Elisa

Wish I was at creamfields 😔

5:09 PM - 26 Aug 2016

Ellyse Bowman

Can ah just stay at Creamfields till next yeAr

1:15 PM - 28 Aug 2016

jake

Would love to go back 2 days and do creamfields all over again, @MarcKinchen absolutely unbelievable😎🎪

6:29 PM - 30 Aug 2016

Not the biggest fan of dance music but I'd love to go Creamfields one year

12:44 PM - 31 Aug 2016

Figure 1: Examples of tweets posted before, during and after the event.

3.2. *Feature space for event attendance classification*

We exploit four different categories of features. Each category reflects a different dimension of social media, namely the: textual, temporal, social, and multimedia dimensions.

- *Textual* features model the textual content of the post. We used two different methods for representing text. The first method uses a Bag of Words (BoW) model. In this case, the textual content is represented as the bag of unigrams, bigrams and trigrams occurring in the post. In order to reduce sparsity, 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. Even if lemmatization reduces sparsity, still the BoW model cannot capture semantic relations among different lemmas. Let us consider for example a post with the words ‘prepared to go’ and another using the words ‘ready to leave’ instead. The same intention to attend the event is expressed in both the posts, but the BoW model does not capture such similarity. Later, we thus propose to encode the text in the posts by exploiting word embedding techniques based on word2vec [42]. These techniques permit to reduce the dimensionality of the textual feature space and, at the same time, to capture text semantics. In addition to the previous features, we consider some additional features modeling textual metadata. Specifically, these features indicate the number of *words, hashtags, mentions, URLs* and *emoticons* occurring in the post. We discuss the text encoding techniques used and study the improvements achieved upon the BoW representation in Section 4.4.

- *Temporal* features represent the time of the post with respect to the event. The temporal dimension is needed to distinguish the classification task (*before*, *during* and *after*), but also to quantify how temporally distant from the event the post has been published. We simply represent time as the number of days separating the posting date from the event date(s). Such temporal feature is obviously meaningful only for the *before* and *after* classification tasks.

- *Social* features characterize the social profile of the posting user. Our social features include the number of followers, the number of followees and the ratio between them. An insight here is that users with a high number of followers and a relatively low number of followees are typically sponsors, organizers or

VIPs that may advertise the event but do not necessarily attend it. Normal users targeted by our attendance classification task are indeed generally characterized by a lower number of followers and a more balanced followers/followee ratio.

- *Multimedia content* features identify whether a post has any multimedia content, such as a photo, video or a link to any visual content posted in other social networks such as Facebook or Instagram. Indeed, this feature group is motivated by the fact that attendees may express their actual or past participation by posting photos or videos during and after the event. In addition, we observe that sponsors commonly use multimedia content before the event as a marketing tool.

It is worth noting that, in order to generalize the classification models learned, we removed the event identifiers $\{i\}$ from the textual content of all of the posts. The generalization aspect of our classifiers is studied in Section 4.2.2.

Table 1: Features used split by category.

Textual	**Temporal**	**Social**	**Multimedia**
unigram			
bi/tri-grams			*num*:photos
num:words	*num*:days before	*num*:followers	*num*:videos
num:hash	*num*:days after	*num*:followees	*bool*:Youtube
num:mentions		*ratio*:(*num*:followers, *num*:followees)	*bool*:Facebook
num:URLs			*bool*:Instagram
num:*emoticons*			*bool*:Foursquare
word embeddings			

Table 1 summarizes the features used by our classifiers grouped using the above four categories. The word embedding features are detailed in Section 4.2.2.

3.3. Research Questions

The overall aim of this paper is to classify social media posts, shared by users before, during and after an event, as indicative of attendance or not attendance. We detail this classification objective into three tasks depending on the temporal aspect of the post: before, during and after. We study the behavior of the approach and, specifically, of the three classifiers, driven by three research questions. These questions will be answered in a number of experiments presented in Section 4. The research questions that we tackle are the following:

RQ1: *How accurate are our event attendance prediction classifiers?* This research question is discussed in details in Section 4.2 where we describe the accuracy results obtained by training supervised machine learning algorithms on an annotated dataset of media event-related posts. We will compare the obtained results with one baseline and discuss the performance achieved on the three different classification tasks. We introduce and discuss three more methods to improve the obtained accuracy. First, in Section 4.2.1, we conduct a *feature ablation* study to identify the feature groups that most contribute to attain high prediction accuracy. We will discover that the textual features are the most important, especially for the before and after tasks. This drives us to the study of word embedding as a way to reduce and enrich the feature space for this group of features. Section 4.2.2 discusses the improvement attained thanks to the word2vec encoding of post texts. Finally, we conclude the study of RQ1 by assessing in Section 4.2.3 the accuracy of the classifiers on a further, objective, ground truth built by considering geo-located tweets.

RQ2: *How do these obtained classifier models generalize across events?* The possibility of deploying an event attendance classifier even when training data for the specific event is not available is highly desirable. In fact, some events do not have a large representation in social media or the cost of building a new training dataset could be unaffordable. The ability of our classifiers to generalize across events is thus of great importance. This research question is discussed in Section 4.3 where we assess how our models generalize across events by applying the model learned on one event to the other and vice-versa.

RQ3: *What are the most meaningful expressions posted by users to express their attendance to a given event?* This question is examined in Section 4.4 where we discuss the results of our analysis of co-occurrence and frequency of the most common terms in the posts classified as attendance or not attendance.

4. Experimental Results

We instantiate our attendance classifiers in a scenario that considers two very popular music festivals held in the UK. Before addressing RQs 1-3, we first describe the setup of our experiments.

4.1. 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). These events are notable in their size, with Creamfields in particular attracting over 70,000 attendees in 2016, and hence likely to be well-reflected in social media. Usually people publish event-related posts using specific hashtags and/or terms that refer to the event. We thus collected tweets related to these events by using the Twitter APIs for selecting tweets including the terms 'vfest' or 'v21st' and 'Creamfields'[2]. Tweets generated by the official accounts of the events (@vfestival and @Creamfields) were removed from the collections, since they are not relevant for our 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). The labelling task has been performed by a single assessor to keep the process consistent. On the other hand, we are aware of the limitations and risks of such human labelling process. In our specific case, we fortunately had the possibility of objectively validating the accuracy of our classifiers and the correctness of the adopted labelling procedure on a

[2] Specifically, in order to cover the time periods before, during and after the considered events, we used the Twitter Streaming APIs from August 10^{th} to September 15^{th} 2016. Moreover, we used the Twitter REST APIs to collect the available tweets related to the events posted from March 1^{st} to September 15^{th} 2016.

Table 2: Creamsfield and VFestival datasets statistics.

Dataset	Task	Labeled tweets	pos%	neg%	Tweets	Users	Geo-located tweets
Creamfields	**Before**	460	48.3	51.7	24,963	11,700	164
	During	460	39.1	60.9	25,625	15,884	309
	After	460	69.3	30.7	29,801	17,850	425
VFestival	**Before**	460	47.6	52.4	10,754	6,513	2
	During	460	37.4	62.6	4,873	3,285	75
	After	460	67.2	32.8	26,027	14,744	58

second, objective ground truth built from posts of georeferenced users. This analysis is reported in Section 4.2.3.

The human assessment is based on the textual or visual content of the tweet, which allows to establish any explicit evidence of attendance at the event. Any other kind of interpretation (advertisement, announcements, newsletter, sponsor's posts, sale of tickets, general information, regrets or impossibility, etc.) is labeled as negative. Table 2 reports for each dataset and task the number of labeled tweets, the respective percentage of positive and negative labeled tweets, the total number of tweets collected, and the number of distinct active users.

Specifically, we collected the tweets by geo-located users posted during the time window of the event and within an area of 3 km radius from the center of the event, gathering a total of 309 tweets from the Creamfields dataset and 75 tweets from the VFestival dataset. These tweets correspond to positive cases of attendance for the *during* task. Starting from these geolocated tweets, we identified a total of 189 distinct users for Creamsfield and 57 unique users for Vfestival who posted those tweets. We also gathered the event-related tweets posted by these users before and after the events. For the Creamfields event, we have 164 tweets before the event and 425 tweets after the event. For the VFestival dataset, we have 2 tweets before the event and 58 tweets after the event. All these tweets are included in a second test set as positive cases of pre- and post-events attendance. Table 2 summarizes in the 'geo-located tweets' column the number of tweets collected for each task by following the above procedure.

Our experiments are conducted using a 5-fold cross validation, while preserving the proportion of positive and negative instances in each fold. For each task and dataset, we train five different classification models, namely: Logistic Regression (LR), Gradient Boosting Decision Trees (GBDT), Random Forest (RF), Support Vector Machine (SVM) and Naive Bayes (NB). All these algorithms, chosen among those consistently delivering state-of-the-art performances in text classification tasks [43], are available in the scikit-learn library[3] used to train our classifiers. We use a grid search to tune the hyperparameters of the algorithms [44]. Specifically: For LR, we consider L1 and L2 regularization and sweep the penalty parameter C in the range of $\{0.01,0.1,1,10,100,1000\}$; For GBDT and RF, we vary the number of trees in the range of $\{50, 80, 100, 120, 150\}$, while the learning rate and maximum tree depth vary in the ranges of $\{0.01, 0.05, 0.1\}$ and $\{2,3,4,5\}$, respectively; For SVM, we use the RBF kernel with γ varying in $\{0.0001, 0.001, 0.01\}$ and C in $\{0.01,0.1,1,10,100,1000\}$.

In the following, we report the performances achieved by our classifiers. Given that the classes are well-balanced in our datasets, and for the peculiarities of the problem addressed both false positives and false negatives have a similar importance, we focus our analysis on classification accuracy values, which directly measure the number of correct predictions made divided by the total number of predictions made. For every classifier, we thus use the setting of hyperparameters that maximizes accuracy by using cross validation. Initially, we report accuracy, precision, recall, F1 and AuC for all classification models trained with the BoW text features. Afterwards, since, as we will show, the LR and GBDT classification models consistently outperform RF, SVM and NB, for the other experiments conducted, we report only the classification accuracy attained using these two classification approaches.

4.2. Results: RQ1

In this section we address RQ1 - studying the accuracy of our event attendance prediction classifiers.

Table 3 reports the accuracy, precision, recall and F1 measure of our 5 classifiers on each dataset and classification task (*before*, *during*, *after*). For the classifiers reported in this table, all feature groups are used, with the textual content of posts represented ac-

[3] `http://scikit-learn.org/`

Table 3: Classification effectiveness using BoW features.

	Dataset: Creamfields						Dataset: VFestival					
Task	**Model**	**Acc.**	**Prec.**	**Recall**	**F1**	**AuC**	**Model**	**Acc.**	**Prec.**	**Recall**	**F1**	**AuC**
Before	LR^{bow}	0.868	**0.870**	0.870	0.868	**0.887**	LR^{bow}	0.761	0.744	0.762	0.748	0.764
	GBDT^{bow}	**0.874**	0.846	0.912	**0.878**	0.873	GBDT^{bow}	**0.809**	0.802	0.768	**0.784**	**0.808**
	NB^{bow}	0.587	0.540	**0.977**	0.696	0.600	NB^{bow}	0.535	0.506	**0.977**	0.667	0.555
	RF^{bow}	0.826	0.760	0.941	0.840	0.830	RF^{bow}	0.778	**0.860**	0.648	0.735	0.772
	SVM^{bow}	0.607	0.591	0.599	0.593	0.606	SVM^{bow}	0.578	0.568	0.471	0.514	0.573
During	LR^{bow}	0.741	0.766	0.538	0.602	0.690	LR^{bow}	0.626	0.600	**0.614**	0.494	0.606
	GBDT^{bow}	**0.817**	**0.830**	**0.616**	**0.708**	**0.790**	GBDT^{bow}	**0.802**	**0.850**	0.582	**0.688**	**0.763**
	NB^{bow}	0.628	0.619	0.117	0.193	0.537	NB^{bow}	0.530	0.429	0.737	0.525	0.571
	RF^{bow}	0.620	0.600	0.028	0.053	0.514	RF^{bow}	0.680	1.000	0.145	0.248	0.573
	SVM^{bow}	0.641	0.584	0.300	0.394	0.580	SVM^{bow}	0.670	0.800	0.157	0.257	0.566
After	LR^{bow}	**0.813**	**0.810**	0.958	**0.880**	**0.762**	LR^{bow}	0.809	0.812	0.932	**0.868**	**0.808**
	GBDT^{bow}	0.780	0.792	0.948	0.864	0.640	GBDT^{bow}	**0.815**	**0.824**	0.902	0.862	0.767
	NB^{bow}	0.702	0.711	0.962	0.818	0.538	NB^{bow}	0.696	0.709	0.929	0.804	0.574
	RF^{bow}	0.713	0.708	**1.000**	0.829	0.532	RF^{bow}	0.689	0.684	**1.000**	0.812	0.527
	SVM^{bow}	0.707	0.706	0.991	0.824	0.527	SVM^{bow}	0.707	0.699	0.994	0.820	0.556

cording to the BoW model. On analysing the results in Table 3, we find that our GBDT classifiers attain the highest performance for all the tasks on the VFestival dataset with an accuracy and precision always greater than 80%. For posts made during the event, GBDT obtained an accuracy of ∼82% when classifying the attendance of the users at the Creamfields event and also when inferring past attendance at VFestival. The performance achieved with GBDT on the VFestival dataset for the after task is also good with an accuracy of nearly ∼82%. LR outperforms GBDT for all metrics on the after task at the Creamfields, while it attains a better recall in other two cases (during and after tasks for VFestival).

In summary, for RQ1, the accuracy results reported in Table 3 show that our approach is reasonably 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.

4.2.1. Feature groups that help the most to attain a high prediction accuracy

In this section, we explore in more details the previous results by analysing the contribution of the feature groups defined in Section 3: multimedia, social, temporal and textual feature groups. Our objective is to understand which feature group deserves further study because it provides the largest benefit to attain a high prediction accuracy.

To evaluate the contribution of each group of features, we conduct an ablation study. Specifically, we remove each group of features one at a time from the datasets used to train and test the classifiers. For such analysis, we use the GBDT classifier, which, according to the results reported in Section 4.2, on average achieves the highest performance. Table 4 reports the results of the ablation study sorted by accuracy for each of the *before*, *during* and *after* classification tasks. In the table, each row denoted with 'All - *feature_group*' indicates that the features of group '*feature_group*' were ablated (removed).

On examination of Table 4, we find that the multimedia features are very important for the *during* task, particularly for 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, we note that 0.85%, 22% and 27% of the tweets posted, respectively before, during and after the event have some multimedia content. For Creamfields, the corresponding percentages tweets containing multimedia content are 0.4%, 8% and 20%, respectively.

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

The temporal features are important when classifying attendance after the completion of the event. We note that low values for this feature (i.e. a shorter difference between the dates before or after the event) are indicative for identifying the actual attendees of the event, while higher values (i.e. more distant from the event) are indicative for identifying non-attendees. This is reasonable when observing the real-world, where people who participated in events discuss them on social media only for a short period of time, usually for a few days before or after the event. Sponsors and news providers, instead, tend to post about the event regularly over a longer time period for marketing purposes. Indeed, by manually inspecting the posts in our training datasets for the tasks before and after, we found that more than 82% of the distribution of posts published by attendees is concentrated in a time interval of 5 days before and 3 days after the event, while the posts of sponsor accounts are more uniformly distributed over time.

Users express their attendance at an event through the post text in different ways depending on the period (before, during, after). Hence, as highlighted in Section 3.1, the textual features extracted from the 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. Furthermore, in all experiments, keeping the textual features allows the models to achieve good accuracies, close to the optimal cases. Before the event, the users mention often their participation by posting about the purchase and delivery of their tickets (feature 'ticket' is among the most important for both Creamfields and VFestival), or when they express their anxiety to attend the festival (e.g. features such as 'wait' and 'excited'). After the event, the users share their experience, how they feel after the event and state willingness to come back to the next edition.

Lastly, the meta textual content (number of *words, hashtags, mentions, URLs* and *emoticons*) only exhibit an 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 the exclusion of this set of features marginally improves the accuracy of the model.

Finally, as a summary of our findings, we observe that while each of the feature groups has some impact for at least one of the tasks, we highlight again the usefulness of the textual features for the prediction of attendance for all the tasks. Indeed, when this group is ablated from the model, the classification accuracy decreases remarkably on both datasets. This observation suggests to attempt improving the results by enriching the group of textual features. This research direction is investigated in the next section.

4.2.2. How to improve classification accuracy with word-embedding features

In the context of RQ1, the analysis in this section aims to investigate new features that could enhance the performance of our classifiers. Thus far, in our models, the tex-

Table 4: Accuracies of the GBDT models by ablating groups of features.

	Creamfields		**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

tual content of posts has been represented as BoW features. One drawback of BoW is that different words have different representations, regardless of their semantic meaning [45, 46]. For example, while the words ‘buy’ and ‘purchase’ have similar meanings (synonyms), in a BoW representation they are as similar as two antonyms. This is not desirable for our attendance classifiers that aim to capture the semantic of the users’ posts. To tackle this problem, we use word2vec, a neural net learning technique that embeds words from a vocabulary into a vector space, which represents the linguistic contexts of words - namely, that words that have similar meanings are represented by close vectors in the embedding space. Specifically, we use the *gensim* [4] implementation of word2vec and a word2vec model trained on part of the Google News dataset (about 100 billion words)[5]. This model contains 300-dimensional vectors for 3 million words and phrases[6]. We represent each post with a single 300-dimensional vector obtained by combining the vectors that represent all the terms occurring in the post. This combination can be done with different aggregation functions. We explore the use of the ‘sum’, ‘mean’ and ‘max’ aggregation functions and also the concatenation of these three representations that we denote by *mix*. The aggregation functions ‘sum’, ‘mean’ and ‘max’ have the intuitive meaning of building a single vector for a post by computing the sum (respectively, mean, max) among the 300 dimensions in the embedding of all the post words. Differently, the *mix* representation of the post consists in simply using the concatenation of the above three aggregated vectors.

Table 5 reports the performances achieved by the GBDT and LR models trained: (a) using BoW features (denoted by *bow*); (b) using word2vec features (denoted by *w2v*) instead of BoW; (c) using both the BoW and w2v features (denoted by *both*). For these experiments the other groups of features (social, temporal and multimedia) are also included in the training sets. For the sake of simplicity, the table reports only the best results achieved by a given algorithm with the corresponding sets of features. For example, the notation $\text{GBDT}^{\text{both}}_{\text{mean}}$ means that the GBDT classifier is trained using *both* BoW and w2v features and that the w2v representation of the post is obtained using the *mean* aggregation function. Similarly, $\text{LR}^{\text{w2v}}_{\text{sum}}$ means that the LR model was trained using w2v features aggregated with *sum*.

We observe that, in general, the use of w2v features improves the classification accuracy compared to the sole use of BoW features (*bow*). Indeed, for the Creamfields dataset, the use of embedding features improves the accuracy and precision figures up to $\sim$91%. It is worth noting that the improvement in Accuracy is higher with LR. In-

[4] `https://radimrehurek.com/gensim/models/word2vec.html`

[5] `https://github.com/mmihaltz/word2vec-GoogleNews-vectors`

[6] We also conducted initial experiments using a word2vec model trained on a large Twitter corpus. However, since the results of the experiments conducted using the Google News model consistently outperformed those with the model trained on the Twitter corpus, we report only the former in the following experiments.

deed, when the w2v textual features are used, either jointly with BoW (*both*) or not (*w2v*), the LR classifiers improve by +4.5%, +7.9% and +2.6% the Accuracy on the before, during, and after tasks on the Creamfields, respectively. Further large improvements are achieved on the VFestival dataset where we observe +5.2%, +16.1%, +4.9% in accuracy for the three tasks. Moroever, the GBDT models attain increased accuracy when using the embedding features, although they are more remarkable for the *after* task. Here, we observe improvements up to $\sim$5% (0.78 $\rightarrow$ 0.833 on Creamfields and 0.815 $\rightarrow$ 0.861 on VFestival) when using only the w2v features. On closer inspection, we see that the w2v features enhance the classification accuracy almost independently of the tasks and algorithm used to train the model. Compared to the results using only the BoW features, the Accuracy is most increased for the *before* (0.874 $\rightarrow$ 0.913 for Creamfields) and *after* (0.815 $\rightarrow$ 0.861 for VFestival) tasks. In these tasks, as discussed above, the textual features have high importance for accurate classification, thus the embedding features provide meaning in a lower-dimensional space that allows for more accurate models compared to the other features.

4.2.3. Assessment of accuracy on the geo-located tweets

As a further evaluation of the classifier accuracy, we test the models with the second ground truth dataset composed by geo-located tweets. Recall that the fraction of geo-located tweets is very low, thus making any approach based on geo-location only not feasible for addressing our event attendance classification problem. However, since the geo-location gives the certainty of the presence of the user at a given place, we can exploit the geo-located tweets in our dataset to further assess the validity of our approach on a second independent test set having no intersection with the training set. In addition, this second experiment permits to indirectly validate the labeling procedure adopted to generate our ground truth.

Table 6 shows the performances of our best performing LR and GBDT models on this second test set. We measured very high accuracies, always higher than 85%, on each classification task. Accuracy reaches the astonishing figures of 96% and 100% on the during task for the Creamfields and VFestival events, respectively. Since the above classification accuracies are higher than those measured on the other test sets, we

Table 5: Accuracy of the GBDT and LR classifiers trained with BoW, w2v and both(BoW+w2v) features. The * indicates statistical significant differences compared to the best classifiers using only BoW features (McNemar's test with 95% confidence interval).

	Creamfields			**VFestival**		
Task	Model	Accuracy		Model	Accuracy	
Before	LR^{bow}	0.868		LR^{bow}	0.761	
	$\text{LR}^{\text{w2v}}_{\text{mix}}$	0.885	(+1.7%)	$\text{LR}^{\text{w2v}}_{\text{mix}}$	0.778	(+1.7%)
	$\text{LR}^{\text{both}}_{\text{max}}$	**0.913***	(+4.5%)	$\text{LR}^{\text{both}}_{\text{sum}}$	0.813*	(+5.2%)
	GBDT^{bow}	0.874		GBDT^{bow}	0.809	
	$\text{GBDT}^{\text{w2v}}_{\text{sum}}$	0.874	(0.0%)	$\text{GBDT}^{\text{w2v}}_{\text{max}}$	0.818	(+0.9%)
	$\text{GBDT}^{\text{both}}_{\text{mean}}$	0.872	(0.0%)	$\text{GBDT}^{\text{both}}_{\text{max}}$	**0.824**	(+1.5%)
During	LR^{bow}	0.741		LR^{bow}	0.626	
	$\text{LR}^{\text{w2v}}_{\text{sum}}$	0.800*	(+5.9%)	$\text{LR}^{\text{w2v}}_{\text{mix}}$	0.772*	(+14.6%)
	$\text{LR}^{\text{both}}_{\text{mix}}$	**0.820***	(+7.5%)	$\text{LR}^{\text{both}}_{\text{mix}}$	0.787*	(+16.1%)
	GBDT^{bow}	0.817		GBDT^{bow}	0.802	
	$\text{GBDT}^{\text{w2v}}_{\text{max}}$	0.789	(0.0%)	$\text{GBDT}^{\text{w2v}}_{\text{max}}$	0.823*	(+2.1%)
	$\text{GBDT}^{\text{both}}_{\text{mix}}$	0.796	(0.0%)	$\text{GBDT}^{\text{both}}_{\text{max}}$	**0.826***	(+2.4%)
After	LR^{bow}	0.813		LR^{bow}	0.809	
	$\text{LR}^{\text{w2v}}_{\text{sum}}$	0.824*	(+1.1%)	$\text{LR}^{\text{w2v}}_{\text{mix}}$	0.850*	(+4.1%)
	$\text{LR}^{\text{both}}_{\text{sum}}$	**0.839***	(+2.6%)	$\text{LR}^{\text{both}}_{\text{sum}}$	0.858*	(+4.9%)
	GBDT^{bow}	0.780		GBDT^{bow}	0.815	
	$\text{GBDT}^{\text{w2v}}_{\text{max}}$	0.830*	(+5.0%)	$\text{GBDT}^{\text{w2v}}_{\text{mix}}$	**0.861***	(+4.6%)
	$\text{GBDT}^{\text{both}}_{\text{max}}$	0.833*	(+5.3%)	$\text{GBDT}^{\text{both}}_{\text{mix}}$	0.854*	(+3.9%)

Table 6: Accuracy of the classifiers on the geo-located tweets.

	Creamfields		**VFestival**	
Task	Model	Accuracy	Model	Accuracy
Before	$\text{LR}^{\text{w2v}}_{\text{mean}}$	**0.854**	$\text{LR}^{\text{both}}_{\text{max}}$	0.500
	GBDT^{bow}	0.726	$\text{GBDT}^{\text{w2v}}_{\text{sum}}$	**1.000**
During	$\text{LR}^{\text{both}}_{\text{mean}}$	0.958	$\text{LR}^{\text{both}}_{\text{sum}}$	1.000
	GBDT^{bow}	**0.964**	$\text{GBDT}^{\text{w2v}}_{\text{sum}}$	**1.000**
After	$\text{LR}^{\text{w2v}}_{\text{sum}}$	0.934	$\text{LR}^{\text{both}}_{\text{mean}}$	0.844
	GBDT^{bow}	**0.960**	$\text{GBDT}^{\text{both}}_{\text{sum}}$	**0.879**

manually inspected the geo-located tweets in these second test sets. We observed that for both festivals, for the during task, about 90% of the geo-located tweets contain some multimedia content. The percentage of during posts including multimedia content in the original ground truth were instead much lower: 8% and 22% for the Creamfields and VFestival events, respectively. As discussed in Section 4.2.1, multimedia features are among the most important for the during task.

The high classification accuracy achieved on this georeferenced posts validates the correctness of the adopted labeling procedure. Finally, it strongly confirms the quality of our attendance prediction classifiers and the validity of our approach based on the content of tweets only.

4.3. Results: RQ2

Our second research question (RQ2) aims to determine how the classifiers can generalize to other similar events (in our case, music festivals). Indeed, while our experiments are conducted over two datasets representing two music festivals, these events have some specific differences. For instance, the VFestival event is a music festival for pop music, while Creamfields is an electronic music festival, with distinctly different genres of performing artists. Therefore, these events may attract different kinds of attendees and may lead to different discussions on social media, reflecting different ways of expressing attendance at the event.

In order to address RQ2, we conduct experiments by applying the model trained on one dataset to classify the labeled samples of the other dataset and vice-versa. The results of these experiments are shown in Table 7. We can observe that our classifiers attain reasonable performances even across different events. The classifiers trained on the VFestival dataset achieve an accuracy $\sim$87% (LR$^{\text{w2v}}$) and $\sim$81% (GBDT$^{\text{both}}$) for the prediction of attendance before and after the Creamfields event respectively. Accuracy however drops to $\sim$75% (GBDT$^{\text{both}}$) for the during task. One possible reason of this drop is that for the VFestival training set the most relevant features for the classification during the event are the posts with photos and Instagram, while for the Creamfields dataset, the textual features were observed to be the most useful (as discussed in Section 4.3).

Table 7 also shows that the LR and GBDT models achieve the highest accuracies when using only w2v features or both BoW and w2v features. The table also shows the improvement of GBDT and LR compared to the use of only BoW features. As expected, we note that the word embedding features substantially boost the performance of cross-event classification with respect to models using BoW features only. Indeed, when training the models with the Creamfields dataset and testing it on VFestival for the *after* task, the GBDT accuracy goes from 71.7% with GBDT$^{\text{bow}}$ to 78.9% with GBDT$^{\text{w2v}}$ (+5.6% improvement compared to the GBDT$^{\text{bow}}$). Moreover, LR reaches 78.7% with LR$^{\text{both}}$ w.r.t. 72.0% with LR$^{\text{bow}}$ (+6.7 %). Answering RQ2, we can conclude that our classifiers, trained on one event and tested on the other, generalize well, particularly benefiting by the abstraction from the specific event provided by the use of w2v features.

4.3.1. Improving the robustness of the classifiers.

We now conduct experiments to understand if the generalizability of our classifiers can be enhanced. In doing so, we use the annotated dataset to understand if a given term occurring in a post is more indicative of attendance or not attendance. To this end, we count the occurrences of all the terms in the positive or negative posts of our gold standard, and consider the normalized frequency of the term in the respective classes as an indicator of whether a word is likely to be more associated with event attendance

Table 7: Generalization ability of the classifiers: models trained on Creamsfields are tested on VFestival and vice-versa. The * indicates statistical significant differences compared to the best classifiers using only BoW features (McNemar's test with 95% confidence interval).

Training/Test	**Creamfields/VFestival**			**VFestival/Creamfields**		
Task	Model	Accuracy		Model	Accuracy	
Before	$\text{LR}^{\text{both}}_{\text{mix}}$	**0.796**	(+1.3%)	$\text{LR}^{\text{w2v}}_{\text{mix}}$	**0.865***	(+1.3%)
	GBDT^{bow}	0.780	(0.0%)	GBDT^{bow}	0.824	(0.0%)
During	$\text{LR}^{\text{both}}_{\text{max}}$	0.702*	(+3.2%)	$\text{LR}^{\text{both}}_{\text{sum}}$	0.741*	(+9.2%)
	$\text{GBDT}^{\text{w2v}}_{\text{max}}$	**0.724***	(+1.3%)	$\text{GBDT}^{\text{both}}_{\text{max}}$	**0.743***	(+3.2%)
After	$\text{LR}^{\text{both}}_{\text{mix}}$	0.787*	(+6.7%)	LR^{bow}	0.787	(0.0%)
	$\text{GBDT}^{\text{w2v}}_{\text{sum}}$	**0.789***	(+5.6%)	$\text{GBDT}^{\text{both}}_{\text{mean}}$	**0.807***	(+3.7%)

or not. For example, a term occurring 10 times in the gold standard, 4 times in posts expressing attendance and 6 times in negative ones, is scored 0.6 for attendance and 0.4 for not attendance. By aggregating (with 'sum', 'mean' and 'max') such values for each term occurring in a post, we can generate two additional features to be used for the classification tasks. Furthermore, the concatenation of the 'sum', 'mean' and 'max' representations is considered (denoted as 'mix'), generating then six additional features (i.e. two features for each aggregation). However, these values are available only for terms occurring in the training set and posts to be classified can include "out-of-vocabulary" (OOV) terms not in this set [47].

The word2vec features provide us with a solution to address the OOV issue. Specifically, given a term t occurring in a post but not present in the training set, we compute its embedding vector v and retrieve the top-k most similar vectors (using Cosine similarity [48]) for which the feature is available from the training set. The feature for t is finally computed as the average of the features associated with the k closest vectors weighted by the cosine similarity. The intuition behind is that terms with similar embedding vectors have also similar semantics. We indicate this approach as *Normalized Frequency Vectors* (*NFV*), and report the results of experiments where we varied the value of k in the range of 1, 3 and 5.

Table 8 reports the accuracy performances for the LR and GBDT classifiers exploiting the NFV features measured across the datasets. In the table, we report the improvement in accuracy achieved over the best results reported in Table 7 and the operators used for aggregating the embedding vectors and the NFV features.

From Table 8, we observe that the NFV features enhance the accuracy of our attendance classifiers up +2.4% and +3.5%, on the VFestival and Creamfields events, respectively. However, the during task still attains the lowest classification accuracies compared to the other two tasks. Furthermore, we see from the table that, for all of the tasks, the accuracy is higher when training uses the VFestival datasets, thus suggesting some over-fitting of the models trained on the Creamfields data. In general however, for most of the tasks and models, we observe statistically significant performance improvements (McNemar's test, $p < 0.05$), corroborating our expectations of the usefulness of the NFV features for the robustness of the classifiers. To better understand how the context and semantic behind the embedding features can help the classification, we investigate in the next section how the semantic similarity among terms actually contribute to the robustness of the models.

Table 8: Robustness of the classifiers exploiting the NFV features. Models trained on Creamsfields are tested on VFestival and vice-versa. The * indicates statistically significant improvements with respect to the best accuracy figures reported in Table 7 (McNemar's test with 95% of confidence interval).

Train/Test	**Creamfields/VFestival**			**VFestival/Creamfields**		
Task	Model$_{\text{aggv,nfv}^{(\text{top})}}$	Accuracy		Model$_{\text{aggv,nfv}^{(\text{top})}}$	Accuracy	
Before	$\text{LR}^{\text{both}}_{\text{max,sum}^{(3)}}$	**0.800**	(+0.4%)	$\text{LR}^{\text{both}}_{\text{max,sum}^{(3)}}$	**0.872***	(+0.7%)
	$\text{GBDT}^{\text{w2v}}_{\text{mix,mean}^{(3)}}$	0.793	(+0.4%)	$\text{GBDT}^{\text{w2v}}_{\text{sum,sum}^{(1)}}$	0.861	(+2.6%)
During	$\text{LR}^{\text{both}}_{\text{max,max}^{(1)}}$	0.707	(+0.5%)	$\text{LR}^{\text{both}}_{\text{sum,max}^{(1)}}$	0.757*	(+1.6%)
	$\text{GBDT}^{\text{w2v}}_{\text{max,max}^{(1)}}$	**0.746***	(+2.2%)	$\text{GBDT}^{\text{both}}_{\text{mean,mix}^{(1)}}$	**0.778***	(+3.5%)
After	$\text{LR}^{\text{both}}_{\text{mix,max}^{(1)}}$	**0.811**	(+2.4%)	$\text{LR}^{\text{w2v}}_{\text{mean,sum}^{(3)}}$	0.811*	(+2.4%)
	$\text{GBDT}^{\text{both}}_{\text{sum,sum}^{(5)}}$	**0.811***	(+2.2%)	$\text{GBDT}^{\text{both}}_{\text{sum,mean}^{(5)}}$	**0.817***	(+1.0%)

4.3.2. Contribution of word embedding features

The experiments above show that the robustness of our classifiers across events is enhanced when word embedding features capturing text semantics for positive and negative attendance are introduced.

To analyze this effect, we consider the twenty five most important terms (BoW features) occurring in the Creamsfields and VFestival datasets and used by the GBDT classifiers trained on the corresponding dataset for each one of the three tasks. Term importance is determined by the gain in the loss function when the node of a decision tree is split on that feature [49]. Then, for each task, the terms occurring in both the datasets are filtered out since they are non-relevant for our analysis. Finally, the Cosine similarity between the embedding vectors of each pair in the Cartesian product of the remaining terms is computed.

The results of this investigation are summarized in Table 9, which reports the top-10 pairs of terms with the highest similarity. From the table, it can be seen that the two datasets include different terms that are likely to be relevant for the classification of the post and whose semantic is captured by the word embedding. For example, the word 'purchase', which appears in some posts of Creamfields but not in the VFestival dataset, has a similar embedding vector to the word 'sell' which, in turn, appears in the VFestival dataset but not in Creamfields: both words are mainly used in posts related to the purchase of the tickets for the events. For the during task, we can observe a high similarity between the embedding of the words 'excite' and 'amaze' and also 'excitement' and 'atmosphere': in both cases, the words mainly represent the attendees' experiences during the event. Similarly for the after task, where we can see the similarity between the words representing periods of time as 'week' and 'weekend' used mainly to refer to the past event.

In summary, in addressing RQ2, we find that the w2v and NFV features introduced allow us to exploit the semantic similarity of text, thus improving the classification accuracy and the robustness across events of our classifiers.

Table 9: Per task top-10 most similar pairs of terms (according to the w2v vectors) in the sets of disjoint terms occurring in the Creamsfields and VFestival datasets.

Before			**During**			**After**		
Creamsfields	VFestival	Sim.	Creamsfields	VFestival	Sim.	Creamsfields	VFestival	Sim.
purchase	sell	0.656	excite	amaze	0.545	week	weekend	0.713
want	go	0.452	wait	watch	0.432	week	day	0.655
ready	wait	0.432	back	rest	0.410	hear	listen	0.649
camp	tent	0.431	excitement	atmosphere	0.382	ago	years	0.505
want	wait	0.419	go	rest	0.347	leeds	justin	0.453
ready	unprepared	0.402	jealous	sick	0.345	week	years	0.433
dj	buzzin	0.391	excitement	experience	0.341	week	time	0.408
work	go	0.354	go	jump	0.318	ago	old	0.393
want	bring	0.315	buzz	atmosphere	0.317	go	miss	0.389
ready	finally	0.300	go	watch	0.317	good	little	0.389

4.4. Results: RQ3

Our last research question (RQ3) asks if it is possible to identify expressions commonly used by users on social media to express attendance (or not) to an event. By using our whole corpus of gathered tweets, we conduct a co-occurrence analysis of the words written in the user’s posts. First, by using our most accurate classifiers for each event and task, we classify all of the unlabeled tweets into (a) attendance and (b) not attendance. Then, for each class, task and event, we compute terms’ co-occurrences to find the set of words most frequently co-occurring in posts of the same class, task and event. The results of this analysis are shown in Table 10 for the positive attendance class, and in Table 11 for the negative one. For the sake of simplicity, in both tables we report only the top-5 sets of 3 words ordered by their co-occurrence frequency. Note that to compute the co-occurrence frequencies, we do not consider the order in which the words occur. It is also worth noting that in this analysis all of the numeric values have been replaced with the symbol ‘#’.

Looking at Table 10, for the before task, we clearly notice the user’s expectation to attend the event when they count down the days, reflected by a high occurrence of the set “{#, days, until}”, or when they mention future participation, supported by the high frequency of the set “{be, next, week}”. This is illustrated for example in the following posts found in the Creamfields dataset: (a) *“I’ll be at Creamfields this* *time next week and I cannot wait"*; (b) *"This time next week I'll be in Creamfields, what an absolute blinding feeling"*; (c) *"Can't believe Creamfields is next week"*. For the during task, the co-occurrence of the words have a much lower frequency. This is justified by the slightly lower amount of tweets in this temporal slot and also by the higher diversity of manners in which people express their current attendance: for example, they sometimes post photos with very few words to describe their personal experience. For the after task, we can see a similar style of posts for both events, expressing pleasure and happiness for attending the event: *"weekend, best, had"*, and desires to relive such experience, commonly written by using the expression *"take me back"*.

On the other hand, the sets of words reported in Table 11 help us to devise common expressions for the negative attendance case. In particular, for the before task and in both datasets, we notice a high correlation among the words 'pounds' and 'ticket' associated with the ticket cost and time periods like month, weekend or day. Indeed, these words are mostly used in advertisements tweets of sponsors and ticket sellers, which are not considered to be actual attendees. For the during task, we notice in the Creamfields dataset common expressions of people regretting not being able to attend the event: (a) *"Couldn't be anymore gutted that I'm not going to Creamfields, cry cry cry"*; (b) *"gutted not to be back at Creamfields this year"*; (c) *"A part of me is very gutted not to be heading to Creamfields tomorrow"*. For the during and after tasks, we observe that many non-attendance posts contain terms related to the performance of famous artists. Those posts are, in general, written by sponsors, newspapers and fans not necessarily attending the festival.

Table 10: Top-5 most frequent 3-grams in the positive attendance class.

	Creamfields		**VFestival**	
Task	Words	Freq.	Words	Freq.
Before	{be, next, week}	253	{#, days, until}	414
	{next, week, time}	214	{#, days, till}	59
	{be, next, time}	178	{ be, so, excited}	59
	{be, week, time}	173	{#, only, hours}	50
	{#, days, work}	170	{weekend, so, excited}	44
During	{#, more, sleep}	68	{park, chelmsford, highlands}	45
	{up, line, great}	40	{you, so, proud}	21
	{#, uk, kingdom}	31	{you, much, thank}	14
	{#, uk, united}	31	{so, park, hylands}	14
	{we, here, come}	29	{down, via, chilling}	13
After	{my, best, life}	377	{weekend, best, had}	302
	{me, back, take}	317	{me, back, take}	207
	{my, weekend, best}	312	{my, weekend, best}	174
	{last, time, week}	283	{my, best, life}	147
	{was, last, time}	233	{weekend, good, such}	134

Table 11: Top-5 most frequent 3-grams in the negative attendance class.

	Creamfields		**VFestival**	
Task	Words	Freq.	Words	Freq.
Before	{#, day, pounds}	7894	{#, tickets, pounds}	141
	{#, pounds, monthdate}	3328	{#, weekend, pounds}	140
	{#, pounds, warrington}	3316	{#, ticket, pounds}	127
	{#, monthdate, warrington}	3316	{#, pounds, sale}	118
	{#, day, camping}	3122	{#, camping, pounds}	118
During	{festival, man, dies}	345	{justin, is, performing}	174
	{be, not, going}	283	{great, john, newman}	120
	{be, not, gutted}	234	{justin, bieber, staffordshire}	116
	{festival, music, dance}	223	{justin, not, bieber}	112
	{going, was, wish}	196	{you, so, love}	101
After	{#, mix, essential}	288	{justin - monthdate - performing}	400
	{#, cirez, essential}	233	{justin, performing, staffordshire}	271
	{#, cirez, mix}	232	{justin, united, kingdom}	262
	{cirez, mix, essential}	209	{justin, monthdate, staffordshire}	260
	{festival, man, dies}	177	{justin, performing, united}	259

5. Example Application: Transport Planning

As an example use case for our proposed classifiers, we aim to evaluate the geographic areas with a higher potential demand for transportation services to an event. We analyse the hometown of users who have been predicted to attend a given festival by our classifiers. This analysis can be useful to support strategies for the allocation of shuttle buses or ride-sharing services to the event, or to forecast possible traffic congestions towards the event. We conduct this analysis upon our Creamfields dataset, the largest in terms of users, thereby allowing for a more realistic analysis compared to the VFestival dataset.

Starting from the event-related posts, we aim to infer the users who participated in the festival. For this purpose, it is important to note that often users on social media share more than one post related to a given event. Each post can be classified as attendance or not attendance depending on the content. There is no guarantee that all event-related posts of the same user will be consistently classified as attending or not attending. We therefore need to infer, given a number of posts of the same user, possibly not uniformly classified as attendance or not attendance, if the user is actually attending or not the event.

For the purpose of this example application, we trained our attendance classifiers on the Creamfields labeled data. We applied the best model for each task according to Table 5 to classify the whole dataset of about 90k tweets. We were able to predict as positive a total of 35,239 tweets. Distinguishing users attending or not attending from a number of - possible discordant - posts can be done in several ways, for example, through majority voting. We propose here a slightly more sophisticated method taking into account the confidence of the used classifier in labelling each post. Intuitively, a more confident attendance prediction should count more than a less confident one. Therefore, during the classification process, for each post, we keep the difference of the confidence scores between the attendance and non-attendance classes. Notice that this value ranges from -1 to 1, where 1 means a higher confidence score for the attendance class, and -1 means the lowest attendance score. Then, taking all the classified posts or users, we compute the mean of the difference of the confidence scores. Our intuition is to capture the most discordant users regarding attendance. As a final decision, we have two cases: (a) users with a positive mean have attended the event (b) users with zero or negative mean have not attended the event.

We perform two kinds of analysis. The first analysis is aimed at inferring future participation in the event based on the posts shared **before** the event. In the second analysis we also consider the posts shared during and after the event. The idea here is to use historical data to identify cities with high amount of attendees to support future strategies in transportation and advertisement for the next editions of the event. The first analysis is based only on the posts published before the event. The idea here is to predict which are the geographical areas with the highest quantity of attendees who may potentially be needing transportation services to reach the event location. We recall that Creamfields is held in Daresbury, England, located between Liverpool and Manchester. We apply the above approach considering only the posts published at least one day before the event. From the quantity of inferred attendees, we collected, using the Twitter REST API, a total of 3856 users' profiles containing details of the users' hometown within their Twitter profiles. Figure 2 shows the spatial distribution of the inferred attendees of the Creamfields festival.

As expected, the results indicate a highest amount of participants in the surroundings of the event location as in the cities of Manchester and Liverpool. However, we can also identify other considerable amount of predicted attendees located in further cities such as London, Newcastle, Peterborough, Glasgow and Edinburgh. Intuitively, the higher the quantity of attendees, the higher the potential demand for transportation services in that area. Therefore, such information could be useful for generating an optimized planning of bus routes across cities and this can provide efficient transportation services to the event. Ride-sharing applications could also take advantage from the identification of groups of predicted attendees. However, we leave such applications as possible future work.

For our second analysis, we run the best classifiers obtained in the generalization experiment for each of the three tasks on the relative sets of posts, namely before, during and after Our intention here is to identify cities with high amount of attendees to support future transportation and marketing strategies for the next editions of the

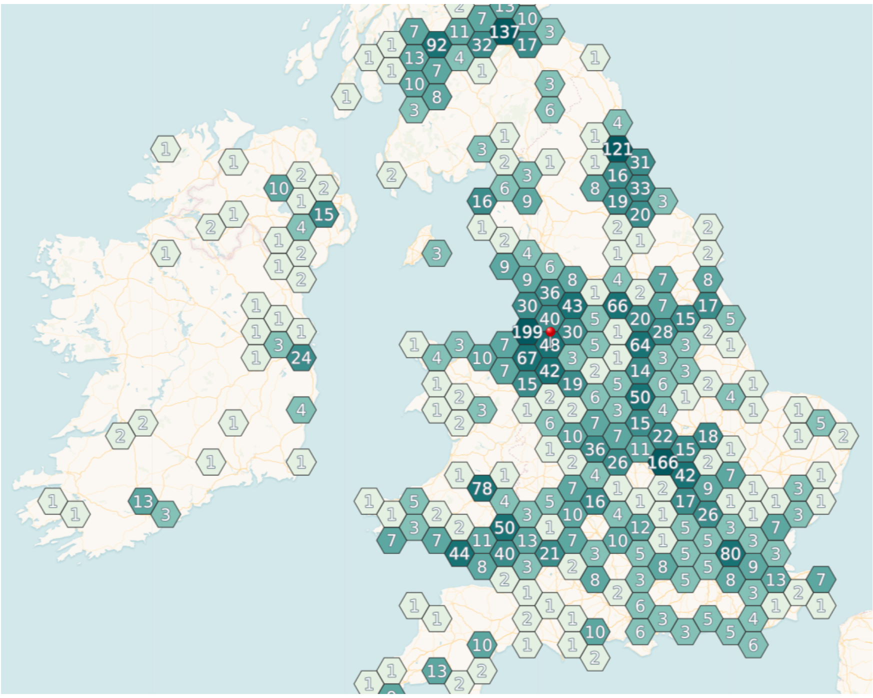


Figure 2: Spatial distribution of inferred participants to the Creamfields festival (red point) from posts published before the event.

event. Here, we use the approach described above to label a user as attendee or not, based on his/her posts. Table 12 summarizes the amount of inferred attendees by city. We have identified a total of 10788 inferred participants to the event that have also their hometown information displayed on their public Twitter profiles. Through the results, we can observe that the previous analysis, predicting the most transportation demanding areas, approximates well the final distribution of attendees by city. We note that Liverpool, Manchester and the surrounding area of the ”North of England” present a high number of attendees. The Scottish cites of Edinburgh and Glasgow might require long-distance transportation services due to the distance of these cities to the event location.

We provide a visualization of the results on a heat map in Figure 3. We also visualized the airports that connect cities from Ireland and The Netherlands to the UK, in blue. Looking at this visualization, we observe red areas (i.e. the hot regions) with

Table 12: Distribution of the before, during and after inferred attendees at the Creamfields festival by hometown.

City	# attendees	City	# attendees	City	# attendees
Aberdeen	57	**Edinburgh**	**263**	Northampton	53
Birmingham	96	**Glasgow**	**221**	Nottingham	112
Bristol	114	Hull	62	Plymouth	86
Cambridge	53	Leeds	160	Sheffield	125
Cardiff	85	Leicester	88	South Wales	109
Coventry	67	**Liverpool**	**732**	Sunderland	55
Derby	53	**London**	**456**	Swansea	106
Doncaster	83	**Newcastle**	312	Warrington	150

a higher density of hometowns of the inferred festival attendees. We observe that, as expected, most of the dense areas are close to the event location. However, we also note some small dense areas located in cities outside the UK, such as the Irish cities of Dublin, Cork and Belfast and the Amsterdam and The Hague Dutch cities. The attendees from these areas might first fly to airports in the UK.

6. Conclusions

In this paper, we proposed a classification approach to infer event attendance from the 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. By not relying on the geotagged posts we can analyze a much larger number of posts to predict user attendance to a given event. The large base of users covered by our approach makes it a good and realistic candidate to enable innovative services and applications in the field, for example, of transportation planning and crowd safety management. We structured the attendance inference into three distinct classification tasks to identify the attendance from the posts published before, during and after the event.

We trained machine-learned classifiers using tweets related to two large music festivals in the UK, and we evaluated their accuracy, precision and recall. The results discussed in Section 4.2 show that our approach provides a remarkably good performance,

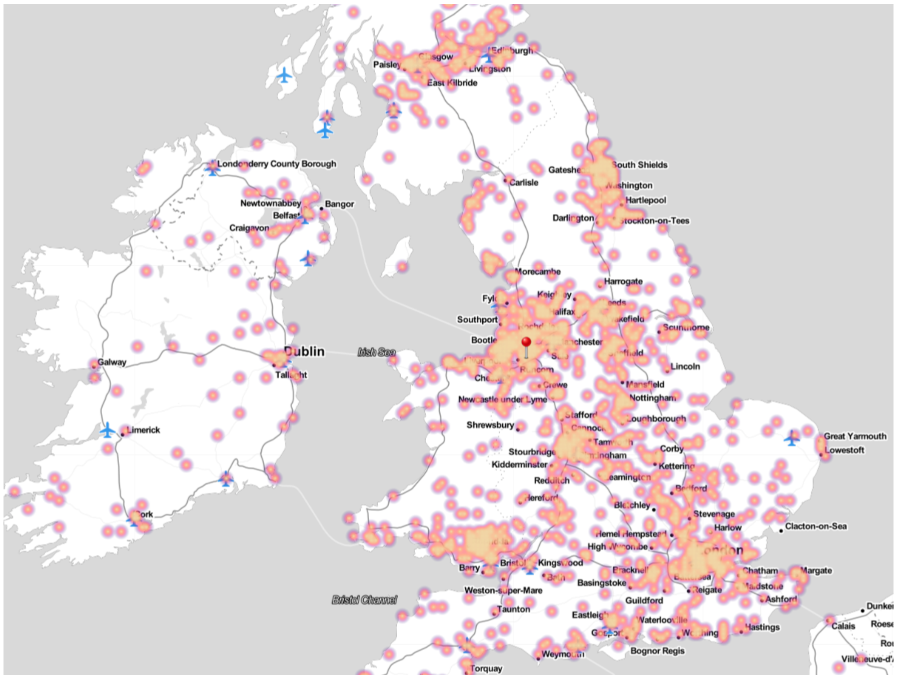


Figure 3: Heatmap with distribution by hometown of the inferred attendees at the Creamfields festival (red point).

exhibiting ∼91% accuracy at classifying users that have indicated their intention to attend the event. Our analysis showed that word embedding features contribute saliently to the performance. Additionally, we highlighted the most informative group of features and assessed the accuracy of our classifier even on an objective test set constituted by geo-tagged tweets. In Section 4.3, we analyzed the generalization of the learned models across the datasets and proposed additional word embedding features to improve cross-dataset performances. For example, when classifying the posts published after the event, by including both the embedding and *NFV* features, the GBDT has increased up to +7.8% (from 73.3% to 81.1%) its generalization ability when trained on Creamfields dataset and tested on VFestival dataset. Furthermore, in Section 4.4, we investigated the common expressions used by social media users to express (or not) attendance to an event. Finally, in Section 5, we proposed an example of application of our methodology in event-related transportation.

As future work, we aim to improve our results using information extracted from the visual content of the published photos or videos. The analysis of visual content is a growing trend in social media and could be better explored in our classification process through the use of deep learning techniques. Furthermore, we aim to further explore our methodology in the context of smart transportation applications.

References

[1] E. Cho, S. A. Myers, J. Leskovec, Friendship and mobility: User movement in location-based social networks, in: Proc. of ACM SIGKDD 2011.

[2] D. Ruths, J. Pfeffer, Social media for large studies of behavior, Science 346 (6213) (2014) 1063–1064.

[3] A. Gal-Tzur, S. M. Grant-Muller, T. Kuflik, E. Minkov, S. Nocera, I. Shoor, The potential of social media in delivering transport policy goals, Transport Policy 32 (2014) 115–123.

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

[5] D. Quercia, N. Lathia, F. Calabrese, G. D. Lorenzo, J. Crowcroft, Recommending social events from mobile phone location data, in: Proc. of ICDM 2010.

[6] C. Chen, J. Ma, Y. Susilo, Y. Liu, M. Wang, The promises of big data and small data for travel behavior (aka human mobility) analysis, Transportation Research Part C: Emerging Technologies 68 (2016) 285 – 299.

[7] D. Efthymiou, C. Antoniou, Use of social media for transport data collection, Procedia - Social and Behavioral Sciences 48 (2012) 775 – 785.

[8] A. Gal-Tzur, S. M. Grant-Muller, E. Minkov, S. Nocera, The impact of social media usage on transport policy: Issues, challenges and recommendations, Procedia - Social and Behavioral Sciences 111 (2014) 937 – 946.

[9] E. D'Andrea, P. Ducange, B. Lazzerini, F. Marcelloni, Real-time detection of traffic from twitter stream analysis, IEEE TITS 16 (4) (2015) 2269–2283.

[10] M. Mikusz, O. Bates, S. Clinch, N. Davies, A. Friday, A. Noulas, Understanding mobile user interactions with the IoT, in: Proc. of MobiSys 2016.

[11] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, C. Ratti, Geo-located Twitter as proxy for global mobility patterns, Cartography and Geographic Information Science 41 (3) (2014) 260–271.

[12] K. Leetaru, S. Wang, G. Cao, A. Padmanabhan, E. Shook, Mapping the global twitter heartbeat: The geography of twitter, First Monday 18 (5).

[13] L. Sloan, J. Morgan, W. Housley, M. Williams, A. Edwards, P. Burnap, O. Rana, Knowing the tweeters: Deriving sociologically relevant demographics from twitter, Sociological Research Online 18 (3) (2013) 1–11.

[14] M. S. Kaiser, K. T. Lwin, M. Mahmud, D. Hajializadeh, T. Chaipimonplin, A. Sarhan, M. A. Hossain, Advances in crowd analysis for urban applications through urban event detection, IEEE TITS (2017) 1–21.

[15] R. O. Sinnott, W. Chen, Estimating crowd sizes through social media, in: Proc. of IEEE PerCom Workshops 2016.

[16] V. M. de Lira, C. Macdonald, I. Ounis, R. Perego, C. Renso, V. C. Times, Exploring social media for event attendance, in: Proc. of IEEE/ACM ASONAM 2017.

[17] Z. Cheng, J. Caverlee, K. Lee, You are where you tweet: A content-based approach to geo-locating twitter users, in: Proc of ACM CIKM 2010.

[18] H.-w. Chang, D. Lee, M. Eltaher, J. Lee, @phillies tweeting from philly? predicting twitter user locations with spatial word usage, in: Proc. of IEEE/ACM ASONAM 2012.

[19] J. Mahmud, J. Nichols, C. Drews, Home location identification of twitter users, ACM TIST 5 (3) (2014) 47:1–47:21.

[20] K. Lee, R. K. Ganti, M. Srivatsa, L. Liu, When twitter meets foursquare: Tweet location prediction using foursquare, in: Proc. of MOBIQUITOUS 2014.

[21] S. Kinsella, V. Murdock, N. O'Hare, I'm eating a sandwich in Glasgow: Modeling locations with tweets, in: Proc of SMUC 2011.

[22] A. Onan, A machine learning based approach to identify geo-location of twitter users, in: Proc. of ACM ICC 2017.

[23] J. Bakerman, K. Pazdernik, A. Wilson, G. Fairchild, R. Bahran, Twitter geolocation: A hybrid approach, ACM TKDD.

[24] H. Efstathiades, D. Antoniades, G. Pallis, M. D. Dikaiakos, Users key locations in online social networks: identification and applications, Social Network Analysis and Mining 6 (1) (2016) 66.

[25] R. Du, Z. Yu, T. Mei, Z. Wang, Z. Wang, B. Guo, Predicting activity attendance in event-based social networks: Content, context and social influence, in: Proc. of UbiComp 2014.

[26] X. Zhang, J. Zhao, G. Cao, Who will attend? predicting event attendance in event-based social network, in: Proc. of IEEE MDM 2015.

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

[28] M. Bogaert, M. Ballings, D. V. den Poel, The added value of facebook friends data in event attendance prediction, Decision Support Systems 82 (2016) 26–34.

[29] L. Li, Predicting online invitation responses with a competing risk model using privacy-friendly social event data, European Journal of Operational Research 270 (2) (2018) 698–708.

[30] A. Q. Macedo, L. B. Marinho, R. L. Santos, Context-aware event recommendation in event-based social networks, in: Proc. of ACM RecSys 2015.

[31] L. Gao, J. Wu, Z. Qiao, C. Zhou, H. Yang, Y. Hu, Collaborative social group influence for event recommendation, in: Proc. of ACM CIKM 2016.

[32] F. Botta, H. S. Moat, T. Preis, Quantifying crowd size with mobile phone and twitter data, Royal Society open science 2 (5) (2015) 150–162.

[33] E. Cesario, F. Marozzo, D. Talia, P. Trunfio, Sma4td: A social media analysis methodology for trajectory discovery in large-scale events, Online Social Networks and Media 3-4 (2017) 49 – 62.

[34] E. Cesario, C. Congedo, F. Marozzo, G. Riotta, A. Spada, D. Talia, P. Trunfio, C. Turri, Following soccer fans from geotagged tweets at FIFA World Cup 2014, in: Proc. of IEEE ICSDM 2015.

[35] S. Zhang, Q. Lv, Event organization 101: Understanding latent factors of event popularity, in: Proc. of ICWSM 2017.

[36] S. Zhang, Q. Lv, Hybrid egu-based group event participation prediction in event-based social networks, Knowledge-Based Systems 143 (2018) 19 – 29.

[37] Y. Mo, B. Li, B. Wang, L. T. Yang, M. Xu, Event recommendation in social networks based on reverse random walk and participant scale control, FGCS 79 (P1) (2018) 383–395.

[38] S. Wang, Z. Wang, C. Li, K. Zhao, H. Chen, Learn to recommend local event using heterogeneous social networks, in: Proc. of APWeb 2016.

[39] C. Y. Liu, C. Zhou, J. Wu, H. Xie, Y. Hu, L. Guo, Cpmf: A collective pairwise matrix factorization model for upcoming event recommendation, in: Proc. of IJCNN 2017.

[40] Z. Qin, I. Rishabh, J. Carnahan, A scalable approach for periodical personalized recommendations, in: Proc. of ACM RecSys 2016.

[41] R. O. Sinnott, W. Wang, Estimating micro-populations through social media analytics, Social Network Analysis and Mining 7 (1) (2017) 13.

[42] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proc. of NIPS 2013.

[43] C. C. Aggarwal, C. Zhai, A Survey of Text Classification Algorithms, Springer, 2012, pp. 163–222.

[44] J. Bergstra, R. Bardenet, Y. Bengio, B. Kégl, Algorithms for hyper-parameter optimization, in: Proc. of NIPS, 2011, pp. 2546–2554.

[45] G. Balikas, M.-R. Amini, An empirical study on large scale text classification with skip-gram embeddings, arXiv preprint arXiv:1606.06623.

[46] G. McDonald, C. Macdonald, I. Ounis, Enhancing sensitivity classification with semantic features using word embeddings, in: European Conference on Information Retrieval, Springer, 2017, pp. 450–463.

[47] Y. Kaewpitakkun, K. Shirai, M. Mohd, Sentiment lexicon interpolation and polarity estimation of objective and out-of-vocabulary words to improve sentiment classification on microblogging, in: Proceedings of the 28th Pacific Asia conference on language, information and computing, 2014.

[48] O. Levy, Y. Goldberg, I. Dagan, Improving distributional similarity with lessons learned from word embeddings, Transactions of the Association for Computational Linguistics 3 (2015) 211–225.

[49] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference and prediction, 2nd Edition, Springer, 2009.

1. Appendix

Table 13: Complement to Table 5 for Creamfields: accuracy achieved by all classifiers trained with BoW, w2v and both BoW+w2v features. The * indicates statistical significant differences compared to the best classifiers using only BoW features (McNemar's test with 95% confidence interval).

Dataset	**Creamfields**						
Task	Model	Accuracy		Precision	Recall	F1-score	AuC
Before	GBDT^{bow}	0.874		0.846	0.912	0.878	0.873
	$\text{GBDT}^{\text{w2v}}_{\text{sum}}$	0.874	(0.0%)	0.869	0.874	0.871	0.874
	$\text{GBDT}^{\text{bow}}_{\text{mean}}$	0.872	(0.0%)	0.865	0.869	0.867	0.872
	LR^{bow}	0.868		0.870	0.870	0.868	0.887
	$\text{LR}^{\text{w2v}}_{\text{mix}}$	0.885	(+1.7%)	0.895	0.905	0.900	0.902
	$\text{LR}^{\text{both}}_{\text{max}}$	**0.913***	(+4.5%)	**0.927**	0.905	**0.916**	**0.919**
	NB^{bow}	0.587		0.540	0.977	0.696	0.600
	$\text{NB}^{\text{w2v}}_{\text{max}}$	0.585	(0.0%)	0.538	**0.982**	0.695	0.598
	$\text{NB}^{\text{both}}_{\text{mean}}$	0.583	(0.0%)	0.537	0.977	0.693	0.596
	RF^{bow}	0.826		0.760	0.941	0.840	0.830
	$\text{RF}^{\text{w2v}}_{\text{mix}}$	0.865*	(+3.9%)	0.842	0.897	0.867	0.866
	$\text{RF}^{\text{both}}_{\text{mix}}$	0.859*	(+3.3%)	0.834	0.892	0.861	0.860
	SVM^{bow}	0.607		0.591	0.599	0.593	0.606
	$\text{SVM}^{\text{w2v}}_{\text{sum}}$	0.637*	(+3.0%)	0.613	0.676	0.642	0.638
	$\text{SVM}^{\text{both}}_{\text{mix}}$	0.654*	(+4.8%)	0.628	0.698	0.661	0.656
During	GBDT^{bow}	0.817		0.830	0.616	0.708	0.790
	$\text{GBDT}^{\text{w2v}}_{\text{max}}$	0.789	(0.0%)	0.791	0.661	0.714	0.768
	$\text{GBDT}^{\text{both}}_{\text{max}}$	0.796	(0.0%)	0.796	0.667	0.720	0.773
	LR^{bow}	0.741		0.766	0.538	0.602	0.690
	$\text{LR}^{\text{w2v}}_{\text{sum}}$	0.804*	(+5.9%)	0.803	0.678	0.730	0.782
	$\text{LR}^{\text{both}}_{\text{mix}}$	**0.815***	(+7.0%)	0.811	**0.706**	**0.751**	**0.796**
	NB^{bow}	0.628		0.619	0.117	0.193	0.537
	$\text{NB}^{\text{w2v}}_{\text{mix}}$	0.637	(+0.9%)	0.816	0.117	0.195	0.544
	$\text{NB}^{\text{bow}}_{\text{mix}}$	0.637	(+0.9%)	0.816	0.117	0.195	0.544
	RF^{bow}	0.620		0.600	0.028	0.053	0.514
	$\text{RF}^{\text{w2v}}_{\text{mean}}$	0.780*	(+16.1%)	**0.855**	0.539	0.656	0.737
	$\text{RF}^{\text{both}}_{\text{mean}}$	0.752*	(+13.3%)	**0.885**	0.428	0.571	0.694
	SVM^{bow}	0.641		0.584	0.300	0.394	0.580
	$\text{SVM}^{\text{w2v}}_{\text{mix}}$	0.641	(0.0%)	0.584	0.289	0.383	0.578
	$\text{SVM}^{\text{both}}_{\text{mean}}$	0.643	(+0.2%)	0.591	0.300	0.396	0.582
After	GBDT^{bow}	0.780		0.792	0.948	0.864	0.640
	$\text{GBDT}^{\text{w2v}}_{\text{max}}$	0.830*	(+5.0%)	0.831	0.953	0.887	0.753
	$\text{GBDT}^{\text{both}}_{\text{max}}$	0.833*	(+5.3%)	0.836	0.947	0.888	0.760
	LR^{bow}	0.813		0.810	0.958	0.880	0.762
	$\text{LR}^{\text{w2v}}_{\text{sum}}$	0.824*	(+1.1%)	0.831	0.937	0.880	0.748
	$\text{LR}^{\text{both}}_{\text{sum}}$	**0.839***	(+2.6%)	**0.847**	0.937	**0.890**	**0.777**
	NB^{bow}	0.702		0.711	0.962	0.818	0.538
	$\text{NB}^{\text{w2v}}_{\text{mean}}$	0.704	(+0.2%)	0.710	0.969	0.820	0.537
	$\text{NB}^{\text{both}}_{\text{mean}}$	0.707	(+0.5%)	0.712	0.969	0.821	0.541
	RF^{bow}	0.713		0.708	**1.000**	0.829	0.532
	$\text{RF}^{\text{w2v}}_{\text{max}}$	0.780*	(+6.7%)	0.763	0.994	0.863	0.646
	$\text{RF}^{\text{both}}_{\text{mix}}$	0.770*	(+5.7%)	0.753	0.997	0.858	0.626
	SVM^{bow}	0.707		0.706	0.991	0.824	0.527
	$\text{SVM}^{\text{w2v}}_{\text{mix}}$	0.713	(+0.6%)	0.709	0.997	0.828	0.533
	$\text{SVM}^{\text{bow}}_{\text{mix}}$	0.713	(+0.6%)	0.708	**1.000**	0.829	0.532

Table 14: Complement to Table 5 for VFestival: accuracy of all classifiers trained with BoW, w2v and both BoW+w2v features. The * indicates statistical significant differences compared to the best classifiers using only BoW features (McNemar's test with 95% confidence interval).

Dataset	**VFestival**						
Task	Model	Accuracy		Precision	Recall	F1	AuC
Before	GBDT^{bow}	0.809		0.802	0.768	0.784	0.808
	$\text{GBDT}^{\text{w2v}}_{\text{max}}$	0.818	(+0.9%)	0.832	0.776	0.802	0.816
	$\text{GBDT}^{\text{both}}_{\text{max}}$	**0.824**	(+1.5%)	0.804	0.835	**0.819**	**0.824**
	LR^{bow}	0.761		0.744	0.762	0.748	0.764
	$\text{LR}^{\text{w2v}}_{\text{mix}}$	0.778	(+1.3%)	0.793	0.826	0.807	0.814
	$\text{LR}^{\text{both}}_{\text{sum}}$	0.813*	(+4.8%)	0.792	0.826	0.807	0.813
	NB^{bow}	0.535		0.506	0.977	0.667	0.555
	$\text{NB}^{\text{w2v}}_{\text{mean}}$	0.535	(0.0%)	0.506	**0.982**	0.668	0.555
	$\text{NB}^{\text{both}}_{\text{sum}}$	0.535	(0.0%)	0.506	**0.982**	0.668	0.555
	RF^{bow}	0.778		**0.860**	0.648	0.735	0.772
	$\text{RF}^{\text{w2v}}_{\text{max}}$	0.804*	(+2.6%)	0.796	0.794	0.794	0.804
	$\text{RF}^{\text{both}}_{\text{mix}}$	0.798	(+2.0%)	0.787	0.799	0.790	0.798
	SVM^{bow}	0.578		0.568	0.471	0.514	0.573
	$\text{SVM}^{\text{w2v}}_{\text{mix}}$	0.609*	(+3.0%)	0.610	0.493	0.545	0.603
	$\text{SVM}^{\text{both}}_{\text{sum}}$	0.602*	(+2.4%)	0.603	0.484	0.537	0.597
During	GBDT^{bow}	0.802		0.850	0.582	0.688	0.763
	$\text{GBDT}^{\text{w2v}}_{\text{max}}$	0.823*	(+2.1%)	0.867	0.633	**0.727**	0.785
	$\text{GBDT}^{\text{both}}_{\text{max}}$	**0.826***	(+2.4%)	0.893	0.622	**0.727**	**0.787**
	LR^{bow}	0.626		0.600	0.614	0.494	0.606
	$\text{LR}^{\text{w2v}}_{\text{mix}}$	0.772*	(+14.6%)	0.855	0.500	0.626	0.722
	$\text{LR}^{\text{both}}_{\text{mix}}$	0.787*	(+16.1%)	0.887	0.505	0.639	0.732
	NB^{bow}	0.530		0.429	0.737	0.525	0.571
	$\text{NB}^{\text{w2v}}_{\text{mix}}$	0.433	(0.0%)	0.388	**0.895**	0.540	0.526
	$\text{NB}^{\text{both}}_{\text{mix}}$	0.446	(0.0%)	0.390	0.866	0.537	0.530
	RF^{bow}	0.680		**1.000**	0.145	0.248	0.573
	$\text{RF}^{\text{w2v}}_{\text{sum}}$	0.796*	(+11.5%)	0.907	0.512	0.651	0.740
	$\text{RF}^{\text{both}}_{\text{max}}$	0.754*	(+7.4%)	0.845	0.442	0.576	0.695
	SVM^{bow}	0.670		0.800	0.157	0.257	0.566
	$\text{SVM}^{\text{w2v}}_{\text{sum}}$	0.676	(0.7%)	0.790	0.175	0.281	0.575
	$\text{SVM}^{\text{both}}_{\text{mean}}$	0.670	(0.00%)	0.800	0.157	0.257	0.566
After	GBDT^{bow}	0.815		0.824	0.902	0.862	0.767
	$\text{GBDT}^{\text{w2v}}_{\text{mix}}$	**0.861***	(+4.6%)	0.862	0.945	**0.901**	**0.817**
	$\text{GBDT}^{\text{both}}_{\text{mix}}$	0.854*	(+3.9%)	0.848	0.948	0.894	0.799
	LR^{bow}	0.809		0.812	0.932	0.868	0.808
	$\text{LR}^{\text{w2v}}_{\text{mix}}$	0.850*	(+4.1%)	0.858	0.932	0.893	0.807
	$\text{LR}^{\text{both}}_{\text{sum}}$	0.858*	(+4.9%)	**0.877**	0.919	0.897	0.827
	NB^{bow}	0.696		0.709	0.929	0.804	0.574
	$\text{NB}^{\text{w2v}}_{\text{mix}}$	0.717*	(+2.1%)	0.717	0.958	0.820	0.592
	$\text{NB}^{\text{both}}_{\text{mix}}$	0.717*	(+2.1%)	0.717	0.958	0.820	0.592
	RF^{bow}	0.689		0.684	**1.000**	0.812	0.527
	$\text{RF}^{\text{w2v}}_{\text{sum}}$	0.789*	(+10.0%)	0.782	0.951	0.858	0.704
	$\text{RF}^{\text{both}}_{\text{mix}}$	0.774*	(+8.5%)	0.763	0.964	0.851	0.674
	SVM^{bow}	0.707		0.699	0.994	0.820	0.556
	$\text{SVM}^{\text{w2v}}_{\text{mean}}$	0.709	(+0.2%)	0.705	0.977	0.819	0.568
	$\text{SVM}^{\text{both}}_{\text{mean}}$	0.709	(+0.2%)	0.699	0.997	0.822	0.558

Table 15: Complement to Table 7 on generalization ability of the various classifiers: models trained on Creamsfields are tested on VFestival. The * indicates statistical significant differences compared to the best classifiers using only BoW features (McNemar's test with 95% confidence interval).

Training/Test	**Creamfields/VFestival**						
Task	Model	Accuracy		Precision	Recall	F1	AuC
Before	GBDT^{bow}	0.780	(0.0%)	0.757	0.795	0.775	**0.862**
	$\text{LR}^{\text{both}}_{\text{mix}}$	**0.796**	(+1.3%)	**0.783**	0.790	**0.786**	0.861
	$\text{NB}^{\text{bow}}_{\text{mean}}$	0.546	(+0.3%)	0.512	**0.945**	0.665	0.565
	RF^{bow}	0.778	(0.0%)	0.748	0.758	0.753	0.843
	$\text{SVM}^{\text{both}}_{\text{mix}}$	0.526	(+0.7%)	0.502	0.470	0.486	0.497
During	$\text{GBDT}^{\text{w2v}}_{\text{max}}$	**0.724***	(+1.3%)	**0.619**	**0.680**	**0.648**	**0.797**
	$\text{LR}^{\text{both}}_{\text{max}}$	0.702*	(+3.2%)	0.607	0.576	0.591	0.732
	$\text{NB}^{\text{both}}_{\text{mix}}$	0.524*	(+4.1%)	0.247	0.134	0.174	0.419
	$\text{RF}^{\text{both}}_{\text{max}}$	0.693*	(+7.8%)	0.604	0.523	0.561	0.679
	$\text{SVM}^{\text{w2v}}_{\text{mix}}$	0.643*	(+1.7%)	0.583	0.163	0.255	0.520
After	$\text{GBDT}^{\text{w2v}}_{\text{sum}}$	**0.789***	(+5.6%)	0.769	0.981	**0.862**	**0.845**
	$\text{LR}^{\text{both}}_{\text{mix}}$	0.787*	(+6.7%)	**0.773**	0.968	0.859	0.817
	$\text{NB}^{\text{both}}_{\text{mean}}$	0.698	(0.0%)	0.699	0.968	0.811	0.559
	$\text{RF}^{\text{w2v}}_{\text{sum}}$	0.735*	(+4.4%)	0.747	0.916	0.823	0.754
	$\text{SVM}^{\text{w2v}}_{\text{mix}}$	0.667*	(+1.5%)	0.671	**0.990**	0.800	0.549

Table 16: Complement to Table 7 on generalization ability of the various classifiers: models trained on VFestival are tested on Creamsfields and vice versa. The * indicates statistical significant differences compared to the best classifiers using only BoW features (McNemar's test with 95% confidence interval).

Training/Test	**VFestival/Creamfields**						
Task	Model	Accuracy		Precision	Recall	F1	AuC
Before	GBDT^{bow}	0.824	(0.0%)	0.844	0.779	0.810	0.912
	LR$^{w2v}_{mix}$	**0.865***	(+1.3%)	**0.867**	0.851	**0.859**	**0.920**
	NB^{bow}	0.570	(0.0%)	0.529	**0.991**	0.690	0.586
	RF ^{bow}	0.808	(0.0%)	0.876	0.667	0.757	0.886
	SVM$^{both}_{mix}$	0.546	(+0.4%)	0.541	0.387	0.451	0.486
During	GBDT$^{both}_{max}$	**0.743***	(+3.2%)	**0.810**	0.450	**0.579**	**0.796**
	LR$^{both}_{sum}$	0.741*	(+9.2%)	0.802	0.450	0.577	0.803
	NB$^{both}_{sum}$	0.370	(+0.3%)	0.377	**0.933**	0.537	0.468
	RF^{bow}	0.678	(0.0%)	0.686	0.328	0.444	0.677
	SVM$^{w2v}_{sum}$	0.593*	(+5.7%)	0.370	0.056	0.097	0.507
After	GBDT$^{both}_{mean}$	**0.807***	(+3.7%)	0.844	0.884	**0.864**	**0.863**
	LR^{bow}	0.787	(0.0%)	**0.862**	0.824	0.843	0.857
	NB^{bow}	0.709	(+0.2%)	0.717	0.959	0.820	0.550
	RF $^{w2v}_{sum}$	0.726*	(+3.7%)	0.818	0.777	0.797	0.757
	SVM^{bow}	0.689	(0.0%)	0.699	**0.969**	0.812	0.582

Table 17: Complement to Table 8: robustness of the GBDT, LR and RF classifiers exploiting NFV features. Models trained on Creamsfields are tested on VFestival. The * indicates statistically significant improvements with respect to the best accuracy figures reported in Table 15 (McNemar's test with 95% of confidence interval). Results of NB and SVM classifiers are not reported since they do not improve by using the NFV features.

Train/Test	**Creamfields/VFestival**						
Task	$\text{Model}_{aggv,nfv^{(top)}}$	Accuracy		Precision	Recall	F1	AuC
Before	$\text{GBDT}^{w2v}_{mix,mean^{(3)}}$	0.793	(+0.4%)	0.772	**0.804**	0.787	**0.867**
	$\text{LR}^{both}_{max,sum^{(3)}}$	**0.800**	(+0.4%)	**0.787**	0.795	**0.791**	0.861
	RF^{bow}	0.763	(0.00%)	0.752	0.749	0.751	0.831
During	$\text{GBDT}^{w2v}_{max,max^{(1)}}$	**0.746***	(+2.2%)	0.646	**0.709**	**0.676**	**0.817**
	$\text{LR}^{both}_{max,max^{(1)}}$	0.707	(+0.5%)	0.612	0.587	0.599	0.732
	$\text{RF}^{both}_{max,sum^{(1)}}$	0.713	(+2.0%)	**0.647**	0.512	0.571	0.733
After	$\text{GBDT}^{both}_{sum,sum^{(5)}}$	**0.811***	(+2.2%)	**0.792**	0.974	**0.874**	**0.872**
	$\text{LR}^{both}_{mix,max^{(1)}}$	**0.811**	(+2.4%)	0.789	**0.981**	0.874	0.866
	$\text{RF}^{both}_{sum,sum^{(1)}}$	0.759	(+2.4%)	0.783	0.887	0.832	0.794

Table 18: Complement to Table 8: robustness of the GBDT, LR and RF classifiers exploiting NFV features. Models trained on VFestival are tested on Creamsfields. The * indicates statistically significant improvements with respect to the best accuracy figures reported in Table 16 (McNemar's test with 95% of confidence interval). Results of NB and SVM classifiers are not reported since they do not improve by using the NFV features.

Train/Test	**Creamfields/VFestival**						
Task	$\text{Model}_{aggv,nfv^{(top)}}$	Accuracy		Precision	Recall	F1-Score	AuC
Before	$\text{GBDT}^{w2v}_{sum,sum^{(1)}}$	0.861	(+2.6%)	0.891	0.811	0.849	0.917
	$\text{LR}^{both}_{max,sum^{(3)}}$	**0.872***	(+0.7%)	0.860	**0.860**	**0.860**	0.915
	$\text{RF}^{w2v}_{none,mix^{(1)}}$	0.833	(+2.4%)	**0.919**	0.716	0.805	**0.921**
During	$\text{GBDT}^{both}_{mean,mix^{(1)}}$	**0.778***	(+3.5%)	**0.792**	0.550	**0.649**	**0.828**
	$\text{LR}^{both}_{sum,max^{(1)}}$	0.757*	(+1.6%)	0.758	**0.556**	0.641	0.797
	$\text{RF}^{w2v}_{mix,max^{(1)}}$	0.702	(+2.8%)	0.717	0.394	0.509	0.725
After	$\text{GBDT}^{both}_{sum,mean^{(5)}}$	**0.817***	(+1.0%)	0.840	0.903	**0.870**	0.839
	$\text{LR}^{w2v}_{mean,sum^{(3)}}$	0.811*	(+2.4%)	**0.847**	0.868	0.858	**0.855**
	$\text{RF}^{both}_{sum,mean^{(5)}}$	0.765*	(+7.6%)	0.771	**0.940**	0.847	0.788