# ReSEED: Social Event dEtection Dataset

Timo Reuter
Universität Bielefeld, CITEC
Bielefeld, Germany
treuter@cit-ec.uni-bielefeld.de

Symeon Papadopoulos
CERTH-ITI
Thermi, Greece
papadop@iti.gr

Vasilios Mezaris
CERTH-ITI
Thermi, Greece
bmezaris@iti.gr

Philipp Cimiano
Universität Bielefeld, CITEC
Bielefeld, Germany
cimiano@cit-ec.uni-bielefeld.de

## ABSTRACT

Nowadays, digital cameras are very popular among people and quite every mobile phone has a build-in camera. Social events have a prominent role in people's life. Thus, people take pictures of events they take part in and more and more of them upload these to well-known online photo community sites like Flickr. The number of pictures uploaded to these sites is still proliferating and there is a great interest in automatizing the process of event clustering so that every incoming (picture) document can be assigned to the corresponding event without the need of human interaction. These social events are defined as events that are planned by people, attended by people and for which the social multimedia are also captured by people. There is an urgent need to develop algorithms which are capable of grouping media by the social events they depict or are related to. In order to train, test, and evaluate such algorithms and frameworks, we present a dataset that consists of about 430,000 photos from Flickr together with the underlying ground truth consisting of about 21,000 social events. All the photos are accompanied by their textual metadata. The ground truth for the event groupings has been derived from event calendars on the Web that have been created collaboratively by people. The dataset has been used in the Social Event Detection (SED) task that was part of the MediaEval Benchmark for Multimedia Evaluation 2013. This task required participants to discover social events and organize the related media items in event-specific clusters within a collection of Web multimedia documents. In this paper we describe how the dataset has been collected and the creation of the ground truth together with a proposed evaluation methodology and a brief description of the corresponding task challenge as applied in the context of the Social Event Detection task.


*MMSys '14,* March 19 - 21 2014, Singapore, Singapore

ACM 978-1-4503-2705-3/14/03 ...$15.00.
http://dx.doi.org/10.1145/2557642.2563674

## Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

## General Terms

Experimentation

## Keywords

Event Detection, Dataset, Content Analysis, Clustering, Classification

## 1. INTRODUCTION

As social media applications proliferate, an ever-increasing amount of web and multimedia content available on the Web is being created. More and more people are using digital cameras to take pictures from important and interesting happenings in their life which they upload regularly to social networks on the Web. In fact, the number of media uploaded to those sites is still increasing every year. A lot of this content is related to social events. According to our definition, a *social event* is an event that is organized and attended by people and illustrated by social media content created by people.

Finding digital content related to a certain social event is very challenging for users. It requires to search large volumes of data, possibly at different sources and sites. It is obvious that algorithms supporting humans in this task are urgently needed. Thus, an important task consists in developing algorithms that can detect event-related media and group them by the events they illustrate or are related to. Such a grouping would provide the basis for aggregation and search applications that foster easier discovery, browsing and querying of social events.

In order to create, test, and validate developments in the field of social event detection [7], it is necessary to have a non-toy and real-world dataset that supports the development and comparison of different approaches on the task. In this paper we propose such a dataset. It has been used in the MediaEval 2013 Social Event Detection task and is thus already established in the community.

In comparison to the datasets from earlier MediaEval SED tasks [4] and TrecVID MED [3] for videos, the dataset presented in this paper features real events and not only event

types. Every picture is assigned to a single and unique event (like the baseball match between the San Francisco Giants vs. New York Mets in April, 2008). In addition, the number of pictures has also increased in comparison to the SED dataset.

# 2. APPLICATIONS OF THE DATASET

The dataset has been created to support researchers in the field of social event detection with a freely available and comparable dataset. While the dataset has already been used in the Social Event Detection task (see Section 2.1), it is also useful for other research questions in the area of event detection and searching.

## 2.1 MediaEval 2013

The dataset was already successfully used in the 2013 edition of the Social Event Detection (SED) task at MediaEval. There were 11 teams with about 65 people altogether who participated in the task. The task description for the challenge was as follows: "Produce a complete clustering of the image dataset according to events.".

The task consisted in the automatic induction of event-related clusters for all images in the dataset, determining the number of events in the dataset automatically. The challenge was thus a completely data-driven task involving the analysis of a large-scale dataset, requiring the production of a complete clustering of the image dataset according to events (see Figure 1). The task might be regarded as some sort of supervised clustering task [6, 5] where a set of training events (groupings of images) was provided. It might also be regarded as a unsupervised clustering task where a classical algorithm like k-means can be used. The event clusters in the training set were disjoint, i. e. participants were told to assume that one image belongs to exactly one event.

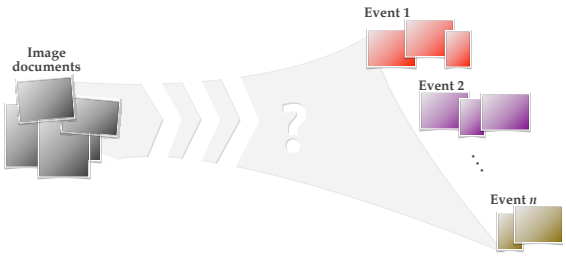


**Figure 1: Clustering of image documents into event clusters**

There was a required run for the challenge which involved using only the metadata. The use of additional data for this run was forbidden (e. g. visual information from the images). For the other runs, additional data could be used (including the images). It was allowed to use generic external resources like Wikipedia, WordNet, or visual concept detectors trained on other data. However, it was not allowed at all to use external data that was directly related to the individual images included in the dataset, such as machine tags[1].

Participants were allowed to submit up to five runs per task, where each run contained a different set of results. This could be produced by either a different approach or a variant of the same approach. Each run was evaluated separately.

[1]A special triple tag to define extra semantic information for interpretation by computer systems

## 2.2 Further applications

A subchallenge of the MediaEval 2013 SED task has consisted in classifying images into a set of nine event types: concert, conference, exhibition, fashion, protest, sports, theatre & dance, other, or classifying it as not depicting any event (non-event). This task thus represents a standard classification task that can be addressed using standard supervised classification approaches and exploiting both visual features and metadata. More specifically, a set of eight event types were pre-defined, and methods were expected to to assign pictures to one of these event types.

Overall, the datasets associated to the MediaEval 2014 SED challenge can be used also for other applications. It is for instance possible to rely on the dataset as a basis to enrich the data via other types of social media information items coming from Twitter or other social network sites. Further, it can be used in tasks related to the semantic search/information retrieval of pictures and/or social media events.

# 3. DATASET

The dataset consists of pictures from Flickr. These are assigned to individual social events. The events include sport events, protest marches, debates, expositions, festivals, concerts, and more. All pictures in the dataset, together with their associated metadata, were downloaded from Flickr using their official API[2]. Furthermore, they are all published under a Creative Commons license allowing for their free distribution. In the following section, we describe the dataset in more detail and also provide some figures. We then proceed to give information on how we obtained the data. The creation process of the event clusters is described in Section 3.3. We end with an exact description of the dataset format in Section 3.4.

## 3.1 Dataset Statistics

The dataset as a whole contains 437,370 pictures from the Flickr photo community site. For the dataset we only considered pictures available under a Creative Common license with an upload time between January 2006 and December 2012, yielding a dataset of 437,370 pictures assigned to 21,169 events in total. The events are heterogeneous with respect to type and length, e. g. it includes festivals which last for several days as well as protest marches with only a few hours of duration. The dataset includes the pictures themselves together with metadata about each picture. The data itself has not been been post-processed in any fashion. As it is a real-world dataset, there are some features like upload time and uploader information that are available for every picture, but there are also features that are available for only a subset of the images. In particular, we fetch the following metadata and information about the pictures directly from Flickr:

- Unique ID for the picture. This is an integer value.
- Information about the person who uploaded the picture to Flickr

[2]http://www.flickr.com/services/api/

**Table 1: Availability of features**

| | |
|---|---|
| Upload time | 100.0% |
| Capture time | 98.3% |
| Geographic Information | 45.9% |
| Tags | 95.6% |
| Title | 97.6% |
| Description | 37.8% |
| Uploader Information | 100.0% |

**Table 2: Distribution per year**

| | |
|---|---|
| 2006 | 4.6% |
| 2007 | 21,0% |
| 2008 | 25,7% |
| 2009 | 21,4% |
| 2010 | 13,5% |
| 2011 | 8,8% |
| 2012 | 5,0% |

- URL where the picture can be downloaded from
- Timestamp when the picture was taken
- Timestamp when the user uploaded the picture to Flickr
- Geographic location (specified by longitude and latitude)
- All tags assigned to the photo
- Title of the photo as chosen by the uploader
- A description of the photo as chosen by the uploader
- The number of people who have viewed that pictures from date of upload till 2013-03-30
- Type of Creative Commons license (as indicated)

In spite of the fact that some pictures include EXIF data from the camera, this data is not used directly in the creation process of the metadata. Nevertheless, the EXIF information is used by Flickr to create the metadata. If the capture time is unknown, Flickr uses the upload time for both timestamps; to ensure a good quality we removed the capture time information if it equalled the upload time. While time, geographic, and tag information is usually of good quality, we discovered that the title often contains the filename; this is problematic in the sense that it is named by the camera (i. e. DSCFxxxx, IMGxxxx, etc.). It is also mentionable that the description field is used by some uploaders to advertise themselves and not to give a discription of the shown picture. Exact statistics for each metadata feature are given in Table 1. We also report the relative number of pictures per year in Table 2 as the distribution over time is not constant.

As already mentioned above, the pictures are licensed under a Creative Commons license. We only included pictures in the dataset which allow for free distribution, remixing, and tweaking. The following subtypes of Creative Commons allow the use of the image along the lines mentioned above as long as the owner is credited for his photograph: Attribution (CC BY), Attribution Share Alike (CC BY-SA), Attribution Non-Commercial (CC BY-NC), and Attribution Non-Commercial Share Alike (CC BY-NC-SA). The two licenses Attribution No Derivatives (CC BY-ND) and Attribution Non-Commercial No Derivatives (CC BY-NC-ND) were not used to enable scientists to also change pictures for research purposes. Exact figures of the distribution of the used licenses are given in Table 3.

**Table 3: Use of license**

| | |
|---|---|
| CC BY | 20.1% |
| CC BY-SA | 15.4% |
| CC BY-NC | 17.2% |
| CC BY-NC-SA | 47.3% |

The distribution of pictures per event is not uniform (see Figure 2). The size of the events varies a lot. While 3,598 events include only one single picture and 1,799 events include 2 pictures, there is a small number of events which include over 1,000 pictures. This is very challenging as not only the natural numbers of cluster has to be determined but also the cluster size is unknown beforehand.

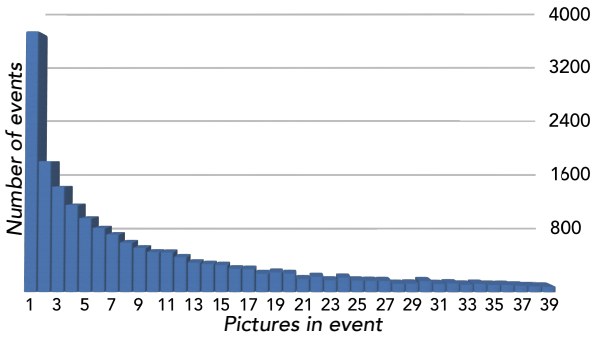


**Figure 2: Distribution of pictures per event**

The 437,370 pictures were uploaded by 4,926 people, thus corresponding roughly to 89 pictures uploaded per person. Here we observe that 2,418 user uploaded only one event which is not surprising.

The total number of tags assigned to the photographs is 3,444,612; there are 91,219 unique tags. About 32.7% of all tags have only one single occurrence in the whole dataset. This is not surprising as the number of single picture events is also high. This leads to the assumption that there is a high correlation between both.

## 3.2 Collection of the Dataset

For the collection of the dataset we relied on the official API from Flickr. The retrieval of the photos and their accompanying metadata was done in four steps:

1. The metadata for the photos was fetched using the `flickr.photos.search` function of the Flickr API. (see Section 3.2.1)

2. All available information about the uploaders of the photos was fetched using the `flickr.people.getInfo` function. (see Section 3.2.2)

3. The photos itself were downloaded via HTTP requests using the addresses fetched in step 1. (see Section 3.2.3)

4. Fetching of additional event information from Upcoming and last.fm (see Section 3.2.4)

Flickr has an essential key function which enabled us to create the gold standard. Every user of the Flickr services has the possibility to assign tags to the pictures which are previously uploaded. Tags are meaningful keywords which describe the picture by important and relevant (key-)words. Usually, these tags contain human-readable information. In 2007, Flickr introduced a special type of tags to be consumed by machines. These special tags are called *machine tags* or *triple tags*. The main difference to the normal user-readable tags is that machine tags use a fixed schema. This schema uses a namespace, predicate, as well as a value. Such a machine tag is denoted in the following form:

`namespace:predicate=value`.

### 3.2.1 Fetching Metadata

In the first step we exploited an API function allowing us to search for photos meeting certain criteria: `flickr.photos.search`. The result of this API call is a set of textual metadata about the photos including the address where the photo itself can be downloaded. Our aim was to download all photos uploaded between January 2006 and December 2012 which fulfill the following criteria: a) there is a special tag assigned to the photo which is either in the namespace of `upcoming:event=` or `lastfm:event=` and b) the license of the picture is one of the following: CC BY, CC BY-SA, CC BY-NC, and CC BY-NC-SA.

Flickr's API returns a maximum of 500 results per call. Even there is no official limit for the number of retrieved photo data using the `flickr.photos.search`-function, after a certain amount of photos retrieved, the API repeatedly returns the last 500 photos as duplicates. In order to circumvent this behavior, we call the API function to retrieve the pictures for one single day. The final API call thus uses three constraints: a) the machine tag is in one of the following namespaces: `upcoming:event=` and `lastfm:event=`, b) the picture has been taken on the specified day, c) the picture is under a redistributable Creative Commons license. The exact approach for the whole process is shown in Algorithm 1.

The result of Algorithm 1 is data for about 450,000 pictures. This data is used as a basis for the next steps.

### 3.2.2 Fetching Uploader Information

As step 2 we use the API again to fetch information about the uploaders. This is necessary to credit the original author as requested by the license. We employ the following function of the API to fetch the information:
`flickr.people.getInfo`.

We request the uploader information for each picture entry in our local database using the unique user ID provided by Flickr. As a result we get the following information about the people:

- Unique ID on Flickr
- Username chosen by the user
- The real name (if the user has provided that name)
- A link to the user profile on Flickr
- The location where the user is situated (if provided by the user)

**Input** : Two namespaces: `upcoming:event=` and `lastfm:event=`
**Output**: Retrieved pictures matching query

```
Day ← '2006-01-01';
while Day ⩽ '2012-12-31' do
    foreach Namespace do
        PageNumber ← photoSearch (Day, Namespace);
        foreach PageNumber do
            CurrentPicture ← photoSearch (Day,
            Namespace, PageNumber);
            foreach CurrentPicture do
                if CurrentPicture is new then
                    Store to local database;
                else
                    Discard picture;
                end
            end
        end
    end
    Day ← Day + 1;
end
```

**Algorithm 1:** Retrieval algorithm

- A self-description of the user

At the end of this step, we have stored the uploader information together with the metadata for the 450,000 pictures.

### 3.2.3 Fetching the Picture Files

Using the addresses from the local database, we downloaded all pictures using a HTTP connection. As about 15,000 of the pictures are not available for download for several reasons (like unavailability of Flickr server, removal by the uploader or copyright holder, changed privacy settings), the process eventually lead to a total of 437,370 pictures downloaded.

### 3.2.4 Fetching event information from Upcoming and last.fm

We employed the APIs of Upcoming and last.fm[3] in order to fetch more information about the events.

We used the `Event.getInfo` function from last.fm's API and the no more existing equivalent from Upcoming to fetch the information shown in Table 4.

At a glance, we obtained detailed information for 5,236 out of 5,384 events from Upcoming and 15,043 out of 16,334 events from last.fm.

## 3.3 Creation of the Gold Standard

As the creation of the gold standard for such high amount of pictures by hand is very time-consuming and expensive, we exploit existing manually created and readily available social events defined collaboratively by a community of users. In fact, there are websites on the Web that host social event calendars. A social event calendar is defined as a repository of social events which can be searched and browsed by events. In such a social event calendar, humans create an entry for a distinct event. These services are often managed professionally and only social events that have been validated by the community are added. The advantage of this

[3]http://www.lastfm.de/api

**Table 4: Available information from two social event calendars, last.fm and Upcoming**

| Information | last.fm | Upcoming |
|---|---|---|
| Unique ID | ✓ | ✓ |
| Event title | ✓ | ✓ |
| Event tags | | ✓ |
| Event description | | ✓ |
| Start time | ✓ | ✓ |
| End time | | ✓ |
| Number of attendees | ✓ | |
| Venue name | ✓ | ✓ |
| Venue address | ✓ | ✓ |
| URL to venue | ✓ | ✓ |
| Exact geolocation | ✓ | ✓ |

is that the events extracted in this way have a rather small noise level. It is notable that these sites provide more information about the social event than the event name. Usually the entries contain more information about the event type but also provide a more detailed description about the event itself; this includes the date and time when the event takes place, the location, a detailed description, and a lot more (a full list of available information is provided in Table 4).

In this paper, we focus on two services which provide a social event calendar: 1) Upcoming and 2) last.fm[4]. The first service, Upcoming, was a public Internet page by Yahoo Inc which provided event information since 2003. The site was unfortunately taken down on April, 30, 2013. The second service, last.fm, provides a calendar since 2007. While last.fm focuses on events related to music, Upcoming provided also entries for all other varieties of events like sport events, conferences, protests, etc. The information of the events available from both services is comparable. In recent years less people were using these services which explains why the number of pictures for recent years are lower (see Table 2).

An example for a social event presented on last.fm is shown in Figure 3.


**Figure 3: Example from last.fm**

Both services provide an API which enabled us to easily access and download the information about the events. There is usually more information available using the API than what is shown in Figure 3. For us, the most interesting information is the unique event identifier (*Event-ID*) for each event; this is a simple integer value but is the key used to uniquely identify an event and can thus be used as the basis for creating a gold standard (as originally proposed by Becker et al. [1]).

The introduction of machine tags in Flickr enabled users to tag their pictures with information in that form. It is automatically recognized as a machine tag by Flickr if the user enters a tag using the special schema for the triple tag `namespace:predicate=value`. These machine tags can be used in several ways. On the one hand, the tags can thus be used to link data, e.g. pictures, across sites. This is where the unique event identifiers from Upcoming and last.fm come into place. The users of both sites were attending these events and took pictures with their cameras, uploading these pictures to Flickr thereafter. Both sites began to ask their users to add a well-formed machine tag to these pictures so that other users could use this unique ID when uploading their pictures to some service. The machine tags for these pictures are in the following form: `upcoming:event=#eventid` or `lastfm:event=#eventid`. For us, this provides important information about the pictures: a) we know that the picture with such a tag belongs to a social event in the database of the social event calendar provider and b) obtain the corresponding unique #eventid for this event. Therefore, we can make the assumption that all pictures which were marked with the same ID in the machine tag belong to the same event [6] (see also Becker et al. [1] for a previous approach exploiting this same principle that inspired this approach).

Having that information, we extracted the machine tags for each picture. Here we discard all machine tags which are not in the namespace we are looking for. We then extract the *value* of the triple tag for the following machine tags: `upcoming:event=`*value* and `lastfm:event=`*value*. These values are then stored together with the Flickr picture ID. We end up with 3 different types of relations for the pictures: a) having only a relation to the Upcoming event calendar, b) having only a relation to the last.fm event calendar, or c) having a relation to both services.

We finally merge an event listed on both services to one event, so that the number of events for the dataset is 21,169.

## 3.4 Dataset Format

The dataset is made available in the following formats:

- Compressed archive of pictures in JPEG format with a maximal optical resolution of 1024px for the longer and 768px for the shorter side.
- CSV files for the textual metadata of the pictures as this facilitates the direct usage and easily allows an import into diverse database systems.
- CSV files for the textual metadata of the events from Upcoming and last.fm.

The files come together with a ReadMe file introducing each file and describing the database schema and its fields in detail. An overview is shown in Figure 4 for the picture metadata.

The dataset is available for download from

*http://greententacle.techfak.uni-bielefeld.de/reseed/*[5] or *http://umass/*.

In Figure 5 we show the database schema for the metadata from the Upcoming events.

# 4. EVALUATION PROPOSAL

For the evaluation and comparability, we split the dataset in a training and an evaluation part. We also propose certain evaluation measures.

[4]http://www.last.fm

[5]http://dx.doi.org/10.4119/unibi/citec.2014.10

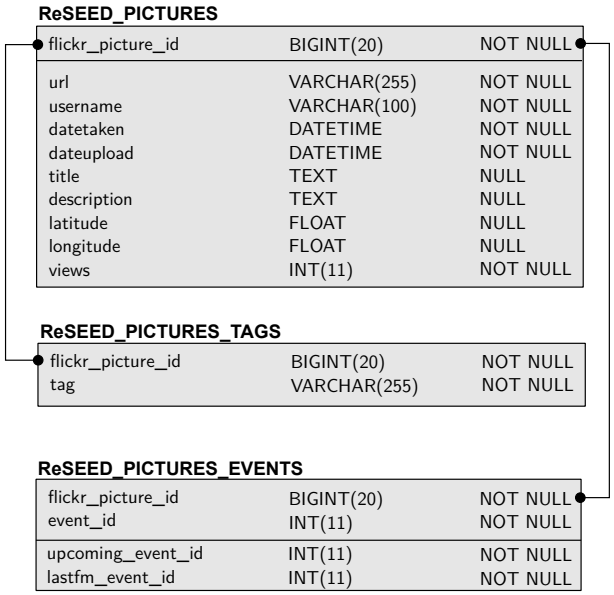


**Figure 4: Database schema for dataset (picture metadata)**

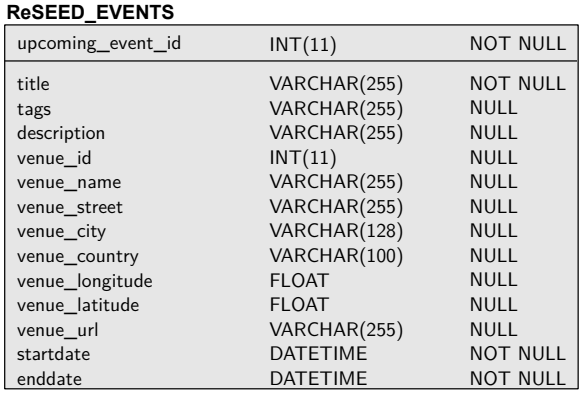


**Figure 5: Database schema for dataset (event metadata from Upcoming)**

## 4.1 Training and Test Dataset Split

The dataset is split into two parts: 70% of the dataset is declared to constitute the training set and the rest is supposed to be used for evaluation purposes. The split was made without the overlap of any events.

## 4.2 Evaluation measurements

We evaluated the submissions to the SED challenge by comparing the results to the ground truth information that has been compiled from the event calendars.

The results of event-related media item detection were evaluated using three evaluation measures which return values in the range of $[0, 1]$ (higher values indicate a better agreement with the gold standard):

- $F_1$-score is proposed to be used as the main evaluation measurement. In SED2013, the micro-averaged version was used to calculate the harmonic mean of Precision and Recall (see also [6]). It measures the appropriateness of for the event clusters.
- Normalized Mutual Information (NMI) to compute the overlap between clusters.
- Divergence from a Random Baseline: indicating how much the results diverge from a random baseline as described in De Vries et al. [2]. This is used as a sanity check.

We deliver an evaluation script together with the dataset so that new algorithms can be compared easily to the official results of the Social Event Detection challenge [7].

# 5. CONCLUSION

In this paper we presented a dataset for research in the area of social event identification. It contains user-contributed images and metadata under a Creative Commons license. It is suitable for clustering and classification tasks in that area. Both the scale and the complexity of the dataset make it more challenging and more representative of real-world problems. The dataset is freely available and we see this dataset as an important contribution to the advancement of the field, supporting the development, evaluation and systematic comparison of different social event detection approaches. As it has been used with the MediaEval Social Event Detection task in 2013 it is also well introduced in the community and helps scientists to easier compare their results to those of other approaches.

## Acknowledgments

The work was supported by the Deutsche Forschungsgemeinschaft (DFG) – Excellence Cluster 277 (Cognitive Interaction Technology) and by European Commission under contracts FP7-287911 LinkedTV, FP7-318101 MediaMixer, FP7-287975 SocialSensor and FP7-249008 CHORUS+.

# 6. REFERENCES

[1] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *WSDM*, pages 291–300, 2010.

[2] C. de Vries, S. Geva, and A. Trotman. Document clustering evaluation: Divergence from a random baseline. 2012.

[3] P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. F. Smeaton, W. Kraaij, G. Quénot, et al. An overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID 2012-TREC Video Retrieval Evaluation Online*, 2012.

[4] S. Papadopoulos, E. Schinas, V. Mezaris, R. Troncy, and I. Kompatsiaris. Social event detection at mediaeval 2012: Challenges, dataset and evaluation. In *Proceedings of MediaEval 2012 Workshop*, 2012.

[5] G. Petkos, S. Papadopoulos, and Y. Kompatsiaris. Social event detection using multimodal clustering and integrating supervisory signals. In *Proceedings of the 2nd ACM Intern. Conf. on Multimedia Retrieval*, page 23. ACM, 2012.

[6] T. Reuter and P. Cimiano. Event-based classification of social media streams. In *Proceedings of the 2nd ACM Intern. Conf. on Multimedia Retrieval*, page 22. ACM, 2012.

[7] T. Reuter, S. Papadopoulos, G. Petkos, V. Mezaris, Y. Kompatsiaris, P. Cimiano, C. de Vries, and S. Geva. Social event detection at mediaeval 2013: Challenges, datasets, and evaluation. Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop Barcelona, Spain, October 18-19, 2013, 2013.