Event-based Clustering for Reducing Labeling Costs of Event-related Microposts

Axel Schulz[1,2], **Frederik Janssen**[4],
Petar Ristoski[3], and **Johannes Fürnkranz**[4]

[1]DB Mobility Logistics AG, Germany
[2]Telecooperation Lab, Technische Universität Darmstadt, Germany
[3]Data and Web Science Group, University of Mannheim, Germany
[4]Knowledge Engineering Group, Technische Universität Darmstadt, Germany

Abstract

Automatically identifying the event type of event-related information in the sheer amount of social media data makes machine learning inevitable. However, this is highly dependent on (1) the number of correctly labeled instances and (2) labeling costs. Active learning has been proposed to reduce the number of instances to label. Albeit the thematic dimension is already used, other metadata such as spatial and temporal information that is helpful for achieving a more fine-grained clustering is currently not taken into account.

In this paper, we present a novel event-based clustering strategy that makes use of temporal, spatial, and thematic metadata to determine instances to label. An evaluation on incident-related tweets shows that our selection strategy for active learning outperforms current state-of-the-art approaches even with few labeled instances.

1 Introduction

Detecting event-related information in microposts has shown its value for a variety of domains. Especially in emergency management, different situational information is present that could contribute to understand the situation at hand. However, solving the actual problem of classification of the incident type in this domain requires labeled data that often is hard to acquire. Therefore, we deal with two major issues: (1) The costs for labeling a single instance, and (2) the number of instances to label.

Due to the huge number of tweets, labeling all instances is impossible as even with cheap labeling the costs would explode. Keeping the number of instances to label low while maintaining accurate classifiers is a typical *active learning* problem. Here, labeling costs are reduced by iteratively (1) selecting small subsets of instances to query for labels and (2) re-training a classifier with the newly labeled data. Thus, there are two important issues to solve, namely selecting a good initial training set and the right instances in each iteration.

For selecting appropriate instances, several selection strategies have been proposed based on the two criteria *informativeness* which measures the usefulness of an instance to reduce the uncertainty of the model and *representativeness* where it is measured how good an instance represents the overall input of unlabeled data (Tang, Luo, and Roukos 2002; Huang, Jin, and Zhou 2010). The latter usually is solved by employing clustering approaches where then from each cluster the instances to be labeled are drawn. As it is unknown how often an event occurred, the number of clusters is not known in advance. Hence, most often the number of distinct event types is used, which obviously is not appropriate. For instance, one event might be a tiny fire in a waste bin whereas another is a huge fire in a factory; though microposts for both events need to be classified with ”fire”, state-of-the-art approaches would not distinguish these two events and thus could not yield an optimal selection of instances. For better distinguishing events, an event should be characterized not only by its type, but also by spatial and temporal information resulting in two different clusters.

Except the work of (Hu et al. 2013), which takes the relations between tweets into account, none of the existing approaches has been evaluated on microposts or has taken event-related metadata into account. Also, no information about real-world error rates is present or was used in active learning. Consequently, we contribute an event-based clustering approach that also leverages the temporal and spatial dimension of tweets to allow a more fine-grained clustering. Due to smaller clusters the selection of appropriate instances is easier because one can assume that even with a bad sampling the selected instances will still be of high quality. The evaluation shows that this enhanced clustering indeed improves the selection compared to state-of-the-art approaches. It is also shown that our approach has a good performance even when only few examples are labeled.

2 Event-Based Classification and Clustering

2.1 Active Learning for Event Type Classification

Active learning is an iterative process to build classification models by selecting small subsets of the available instances to label. Two major steps are conducted: (1) a learning step, where a classifier is built and (2) an improvement step, in which the classifier is optimized. We follow a pool-based sampling approach. First, large amounts of microposts are collected as an initial pool of unlabeled data U. From this, a set of training examples L is chosen. It is highly important

how to choose this set, because with a well-selected initial training set, the learner can reach higher performance faster with fewer queries.

For training a classifier using this initial set, we reuse the classification approach for social media data presented in (Schulz, Ristoski, and Paulheim 2013). As the amount of geotagged microposts is rather low, we employ an extension of our approach for geolocalization (Schulz et al. 2013) of microposts and for extracting location mentions as features.

After the initial training, this classifier is retrained in several iterations using newly labeled instances. After each iteration, the labeled instances are removed from the pool of unlabeled instances U and added to the pool of labeled instances L, thus, more instances can be used for learning. A selection strategy is used on U to query labels for a number of instances in each iteration. For coping with this query selection problem, several selection strategies can be chosen based on informativeness and representativeness.

For informativeness, uncertainty sampling (Lewis and Catlett 1994) is commonly applied that selects particularly these examples for labeling for which the learner is most uncertain. However, the main issue with this approach is that only a single instance is considered at a time, often leading to erroneously selecting outliers. In contrary, clustering helps to identify representative instances. The most representative examples are those in the center of the cluster, which are the instances most similar to all other instances. Nevertheless, selecting always the centers of the clusters might result in selecting always very similar instances for each iteration hindering improvement of the model. Furthermore, it remains unclear how many clusters have to be built. Also, and most important in our case, the resulting clusters not necessarily correlate to the real-world events as spatial and temporal information is omitted.

The general idea to overcome these individual problems is to select the most informative *and* representative instances. This results in selecting the instances, which are representative for the whole dataset as well as have the highest chance to improve the model. We use metadata provided in microposts to cluster instances based on both criteria and to choose the most valuable instances for training the classifier. The whole process of active learning continues until a maximum number of iterations is reached or when the model does not improve any more.

2.2 Event-based Clustering

Clustering-based approaches are frequently used for identifying representative instances. However, there might not be an obvious clustering of event-related data, thus, clustering might be performed at various levels of granularity as the optimal number of cluster is unknown.

Consequently, we use a more natural way of clustering by taking the properties of real-world events into account. We use event-related information such as temporal and spatial information in combination with the event type to perform an *event-based clustering*. On the one hand, we are directly able to find a number of clusters without the need of specifying the number beforehand and on the other hand both selection criteria are combinied.

The design of our approach follows the assumption that every event-related information is either related to a specific real-world event or not. Thus, we propose to cluster all instances based on the three dimensions that define an event: temporal and spatial extent as well as the event type. As a result, each instance is aggregated to a cluster. As we use the properties of real-world events, it is much easier to identify those tweets that might be helpful for training.

If a micropost lies within the spatial, temporal, and thematic extent of another micropost, then the new micropost is assumed to provide information about the same event. The spatial extent is given by a radius in meters around the location of the event, the temporal extent is a timespan in minutes, and the thematic extent is the type of the event. To specify the spatial and temporal extent we relied on emergency management staff. In this work we used *200m* and *20 min*. Clearly, altering the radius or the time will have a strong effect on the final clustering. Inspecting the effects of different parametrizations remains subject for future work, however, we are confident that our proposed approach is not affected negatively by a change of these parameters.

To handle missing values, microposts containing no thematic information are assigned the $unknown\_event$ type. Missing spatial information is replaced with a common spatial center, e.g., the center of the city for which the microposts are used. Missing temporal information is replaced with the creation date of the micropost.

Based on this clustering approach, we are able to cluster all microposts related to a specific event. This helps to identify those microposts that might be helpful for better training. Opposed, those not related to events are assigned to larger clusters, containing lots of noise and being less valuable for the learning process.

2.3 Initial Selection Strategy

First, the initial dataset that needs to be labeled is selected. Related approaches rely on random sampling or clustering techniques (Zhu et al. 2008). However, the selection of appropriate instances is not guaranteed, because the initial sample size is rather small, whereas the size of clusters is large. In contrast, event-based clustering uses the properties of real-world events to perform an initial clustering.

Based on the set of clusters resulting from our event-based clustering, the most representative instances for the complete and unlabeled dataset are identified. For this, we use the event clusters ordered by information density of their containing instances. Selecting informative instances clearly is not possible yet, as a classifier cannot be trained at this point. In the following, we describe the algorithm in detail.

First, our clustering approach is applied on the complete unlabeled set U without a thematic specification. Thus, the $unknown\_event$ type is used as a thematic extent. Second, for all instances in each cluster the information density is calculated. This is done based on the similarity of instances, thus, outliers are regarded as less valuable. We used a K-Nearest-Neighbor-based density estimation (Zhu et al. 2008): $DS\left(x\right) = \frac{\sum_{s \in S(x)} \text{Similarity}(x, s_i)}{N}$

The density $DS(x)$ of instance x is estimated based on

the N most similar instances in the same cluster[1] $S(x) = \{s_1, s_2, ..., s_i\}$. As a similarity measure, we use the cosine similarity between two instances. The information density DSC of each cluster C is then calculated based on the average of the information density of each instance as follows: $DSC\,(c) = \frac{\sum_{x \in C} DS(x)}{N}$

Doing this, we are able to avoid noisy clusters with lots of unrelated items, which would typically be clusters not related to an event. Based on $DSC(c)$ the clusters are sorted. Then we iterate over the ordered list and select instances until b_i (number of the initial training size) instances are selected. Proceeding this way, we achieve a good distribution over all valuable event clusters as it is guaranteed that the instances are selected from the most representative clusters. Based on these instances, the initial model is build.

2.4 Query Selection Strategy

The initial selection strategy gives us the most valuable instances for training the initial model. For every following iteration, appropriate instances for improving the classifier have to be chosen. Besides identifying representative instances based on clustering, the goal of our approach is to avoid instances that the learner is already confident about.

In every iteration, the classifier trained on the currently labeled instances is applied to label all unlabeled instances. As a result, every instance is assigned a thematic dimension. Then, the event clustering is applied using the spatial, temporal, and thematic information yielding a set of clusters C.

Next, for the query selection strategy, we calculate the information density DS per instance. For identifying informative instances, we use the instances for which the classifier is most uncertain. As uncertainty measure the entropy calculated for each instance x and each class y was employed.

Based on the information density and the entropy, the density×entropy measure $DSH(x) = DS(x) \times H(x)$ (Zhu et al. 2008) is calculated for each instance x. The informativeness and representativeness of each cluster is then computed based on the mean average of DSH of each instance i in the cluster c: $DSHC\,(c) = \frac{\sum_{i \in C} DSH(x)}{N}$

For selecting appropriate instances to query, the clusters are sorted by the $DSHC$ of each cluster. The number of instances to draw per cluster is calculated as $n = \log_{(ms)} CS$. To determine how many instances have to be selected per cluster (n), we calculate the average size of all clusters ms and the size of the current cluster CS. We decided to use a logarithm at basis ms to avoid drawing too many instances from larger clusters as would be the case with a linear approach. In the latter, large clusters would contribute many more instances compared to small clusters which is avoided by penalizing large clusters with the employed logarithmic scale. We assume that drawing only small numbers per cluster is sufficient, as at some point additional instances will not yield any additional information. Furthermore, we achieve that a limited amount of instances is drawn per cluster, to avoid choosing too similar instances for training, e.g., as it would happen by using $n = CS/ms$.

[1] K is equal to the number of instances in the cluster.

We select instances until the number of instances to label per iteration is reached. Based on the previous and the new instances the model is retrained. The whole process is repeated until all iterations are finished.

3 Experiments

3.1 Methodology

We differentiate between three incident types and a neutral class in order to classify microposts: *car crash*, *fire*, *shooting*, and *no incident*. We collected public microposts in English language using the Twitter Search API in a 15km radius around the city centers of Seattle, WA and Memphis, TN. As this initial set of 7.5M tweets needed to be labeled manually, we had to further reduce the size of the datasets. For this, we identified and extracted microposts containing incident-related keywords as described in (Schulz, Ristoski, and Paulheim 2013).

After applying keyword-filtering, we randomly selected 2,000 microposts. These were then manually labeled by four domain-experts using an online survey. To assign the final coding, at least three coders had to agree on a label. Instances without an agreement were further examined and relabeled during a group discussion. The final dataset consists of 2,000 tweets (328 fire, 309 crash, 334 shooting, 1029 not incident related). For our evaluation, we used 1,200 tweets from dataset for training and 800 tweets for testing (temporal split, i.e., the testing instances are later in time than the training instances). Though this selection might seem arbitrary, all compared algorithms rely on the same sampling, thus, allowing for a fair comparison. However, please note that the absolute numbers in terms of F1 do only reflect the performance for the current train/test split.

As a classifier, we used Weka's support vector machine (Platt 1998). A different classifier could also be used but the primary interest is the difference of the approaches not so much the absolute performance.

The active learning algorithms select instances from the training set to query for labels. Based on these, a classifier was trained and then evaluated on the test set. Due to the complexity of determining best parameter settings for each iteration and each approach, we follow related approaches (see (Huang, Jin, and Zhou 2010)), and decided to compare all algorithms on fixed parameters. Consequently, the SVM was used with standard settings. Clearly, parameter tuning would result in much better performance.

For comparison, the deficiency metric (Raghavan and others 2006) is calculated using the achieved F1 score of all iterations of a reference baseline algorithm (REF) and the compared active learning approach (AL). The result is normalized using the maximal F1 score and the learning curve of the reference algorithm REF. Thus, the measure is non-negative and values smaller than one indicate better performance than the REF algorithm, whereas a value larger than one means worse performance.

We applied different active learning algorithms. In order to evaluate the performance of our approach, we compared it to two state-of-the-art clustering-based approaches that also take representativeness as well as informativeness into ac-

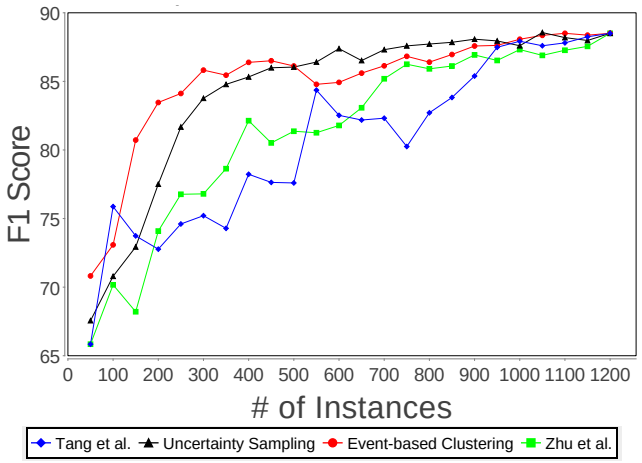

Figure 1: Evaluation results of state-of-the-art selection strategies and our approach.

Table 1: Deficiencies with Tang et al. as a baseline strategy

Approach	Deficiency
(Tang, Luo, and Roukos 2002)	1
Uncertainty Sampling	0.53
(Zhu et al. 2008)	0.90
Event-based Clustering	0.44

count. Furthermore, we compared to an entropy-based uncertainty sampling algorithm.

We reimplemented the approaches of (Tang, Luo, and Roukos 2002) and (Zhu et al. 2008) as well as a simple uncertainty sampling. When using 200m and 20 min for our approach, the 1,200 tweets of the training set are divided into 438 distinct event clusters.

Following the experimental settings of (Huang, Jin, and Zhou 2010) and (Hu et al. 2013) we set the size of the initial training set as well as the size during the iterations to 50. No further tuning or parameterization was applied. Each iteration for each algorithm was repeated 10 times, as for instance, the uncertainty approach is highly dependent on the selected instances. We used the averaged F1 score based on the repetitions.

3.2 Comparison to state-of-the-art approaches

The overall performance graph for the ground truth data is shown in Figure 1. As can be seen in the graph, the performance after selecting the initial training set is superior with our approach. Also, in regions where only a few instances were labeled, the event-based clustering has a higher F1 value. This shows that a high-quality selection of the iteration instances is possible with our method.

Table 1 shows the deficiency. With respect to the performance of the iterations, our approach has a decreased deficiency compared to other clustering approaches (0.44 *vs.* 0.53). The approach of Zhu et al. outperforms the approach of Tang et al. in most iterations and also with respect to the deficiency. We attribute this to the improved strategy for query selection. A surprising result is the performance of uncertainty sampling, which outperforms the other two clustering strategies. Apparently, only focusing on the informativeness seems to be a good strategy for our dataset. In contrast, using the number of distinct event types as the number of clusters might not be the most efficient approach.

The graph also shows that our approach has a steep learning curve as only a sixth of all instances are needed to achieve about 84% F1. This is especially important when it comes to labeling costs, as only a limited amount of data would need to be labeled. One can see that our approach has a drop at 500 instances. This is most likely because with more instances the number of clusters is decreasing, thus, selecting appropriate instances is more difficult.

We can conclude that event-based clustering that takes representative as well as informative instances into account is a promising strategy for active learning. We also showed that our approach outperforms state-of-the-art for selecting an initial training set and for choosing appropriate instances for labeling in each iteration.

4 Conclusion

In this paper, we presented an event-based clustering strategy for event type classification of microposts based on temporal, spatial, and thematic information. The approach that identifies representative as well as informative instances outperforms state-of-the-art clustering methods and was able to select a better initial training set as well as to choose appropriate instances for labeling in each iteration.

References

Hu, X.; Tang, J.; Gao, H.; and Liu, H. 2013. Actnet: Active learning for networked texts in microblogging. In *SIAM'13*.

Huang, S.-J.; Jin, R.; and Zhou, Z.-H. 2010. Active learning by querying informative and representative examples. In *NIPS*, 892–900.

Lewis, D. D., and Catlett, J. 1994. Heterogeneous uncertainty sampling for supervised learning. In *ICML-94*, 148–156.

Platt, J. 1998. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

Raghavan, H., et al. 2006. Active learning with feedback on both features and instances. *J. of Machine Learning Research* 7:1655–1686.

Schulz, A.; Hadjakos, A.; Paulheim, H.; Nachtwey, J.; and Mühlhäuser, M. 2013. A multi-indicator approach for geolocalization of tweets. In *Proc. ICWSM*.

Schulz, A.; Ristoski, P.; and Paulheim, H. 2013. I see a car crash: Real-time detection of small scale incidents in microblogs. In *Proc. ESWC*, 22–33.

Tang, M.; Luo, X.; and Roukos, S. 2002. Active learning for statistical natural language parsing. In *ACL'02*, 120–127.

Zhu, J.; Wang, H.; Yao, T.; and Tsou, B. K. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *COLING'08*, 1137–1144.