Event-Based Clustering for Reducing Labeling Costs of Incident-Related Microposts

Axel Schulz SCHULZ.AXEL@GMX.NET
DB Mobility Logistics AG, Germany and Telecooperation Lab, Technische Universität Darmstadt, Germany

Petar Ristoski PETAR.RISTOSKI@INFORMATIK.UNI-MANNHEIM.DE
Data and Web Science Group, University of Mannheim, Germany

Johannes Fürnkranz JUFFI@KE.INFORMATIK.TU-DARMSTADT.DE
Knowledge Engineering Group, Technische Universität Darmstadt, Germany

Frederik Janssen JANSSEN@KE.TU-DARMSTADT.DE
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. Though, current approaches focus on the thematic dimension, i.e., the event type, for selecting instances to label; other metadata such as spatial and temporal information that is helpful for achieving a more fine-grained clustering is currently not taken into account. Also, labeling quality is always assumed to be perfect as currently no qualitative information is present for manual event type labeling.

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. Furthermore, we also inspect the quality of the manual labeling in a crowdsourcing study by comparing experts and non-experts. An evaluation on incident-related tweets shows that (i) labels provided by crowdsourcing are of acceptable quality and (ii) 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 (Schulz, 2014). However, solving the actual problem of classifying the incident type in this domain requires labeled data. One of the main problems with microposts is acquiring ground truth for utilizing supervised learning. Thus, we deal with two major issues: (1) The costs for labeling a single instance, and (2) the number of instances to label.

On the one hand, to actually build a classifier that is able to accurately predict the type of the incident mentioned in a tweet, usually experts are deployed for labeling as they have enough domain knowledge to create ground truth. However, as often several hundreds of examples have to be labeled until the classifier is able to reach sufficient quality, relying on experts for labeling is not always possible and it is costly. In contrast, labels can also be derived from non-experts, i.e., by making use of crowdsourcing. Given that the labels obtained in this way are of sufficient quality, the costs for such a process would be acceptable as crowdsourcing is rather cheap. But up to now there is no information about labeling quality for incident-related tweets. Hence, first we proceeded by comparing the labeling quality of experts and non-experts.

On the other hand, the number of instances to label has to be kept as low as possible. Due to the huge number of tweets, labeling all instances is not possible as even with cheap labeling the costs would explode. Keeping the number of instances to label low while maintaining accurate

Proceedings of the 2^{nd} International Workshop on Mining Urban Data, Lille, France, 2015.

classifiers is a typical *active learning* (Settles, 2012) 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, in general, but also specifically for classifying microposts, there are two 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* and *representativeness* (Huang et al., 2010). *Informativeness* measures the usefulness of an instance to reduce the uncertainty of the model, whereas *representativeness* measures how good an instance represents the overall input of unlabeled data. The latter usually is solved by employing clustering approaches where then from each cluster the representative instances are drawn. Indeed, for event-type classification the number of clusters to build is not known in advance, as it is unknown how often an event occurred. Hence, most often it is set to the number of distinct event types, 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 the "fire" event type, state-of-the-art approaches would not distinguish these two events and thus could not yield an optimal selection of instances to label. For better distinguishing events, a straightforward approach is to characterize an event not only by its type, but also by spatial and temporal information. Proceeding this way, the two example events are inherently assigned to different clusters and hence instances to be labeled are drawn from both of them.

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 on incident-related tweets 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.

In summary, the contributions of this paper are: (1) A study comparing the labeling quality of experts and non-experts showing no significant difference of error rates. (2) A novel event-based clustering approach that makes use of spatial, temporal, and thematic information present in microposts. The clustering benefits strongly from these additional dimensions. (3) A comparison of our approach using different number of annotators and different levels of noise. Even with a classifier that was not explicitly build to be robust, noise does not hinder the classifier much.

We begin with summarizing related approaches. Next, we show how the ground truth data was developed. Then we summarize the results of the study on crowdsourced labels (Section 4) followed by a description of the event-based clustering for active learning. After, the results are shown and discussed (Section 6) and the paper is concluded.

2. Related Work

Although active learning has been studied extensively for text classification (Hoi et al., 2006; Tong & Koller, 2002), it was used for tweets only by a few previous works. (Thongsuk et al., 2010) presented a technique for classifying tweets into three business types. They showed that using active learning outperforms simple supervised learning approaches in terms of labeling costs.

(Hu et al., 2013) presented the ActNeT approach, which takes the relations between tweets into account for identifying representative as well as informative instances. Based on a social network, the topology is used to detect representative instances using the PageRank algorithm. Informative instances are chosen using an entropy-based uncertainty sampling. However, as building the social network is time consuming and not always possible due to API restrictions, their approach is not applicable for our problem. Also, they do not use event-related metadata.

Several selection strategies were presented that propose to select informative as well as representative instances. (Tang et al., 2002) used k-means clustering and proposed to select the most uncertain instance for each cluster. Information density was then used to weight instances. (Shen et al., 2004) applied k-means clustering and uncertainty sampling and used the information density calculated within a cluster. (Donmez et al., 2007) combined uncertainty sampling and k-Medoid to identify representative as well as informative instances and showed that this combination is indeed beneficial.

The approach of (Zhu et al., 2008) is the most advanced related approach when it comes to combining representativeness and informativeness, thus, we used it as a foundation for our technique. The authors employed clustering for the initial selection. Uncertainty sampling is combined with estimating a density for each iteration. Unlike their work, we apply our event-based clustering also for the iterations. (Huang et al., 2010) followed a similar approach. Instances are selected based on clustering and on confidence in predicting a class label as informativeness measure. Though their approach is quite promising, the authors stated that it is restricted to binary classification, whereas we are able to classify multiple classes.

Taking labeling quality into account is still open to research. Up to now, there is no study of labeling quality of event-related tweets, but only studies on structured texts such as the work of (Hsueh et al., 2009). Since 2008, the active learning community also tackled the problem of different reliabilities of oracles (Donmez & Carbonell, 2008; Zhao et al., 2011; Wallace et al., 2011). These approaches have been proposed to take labeling uncertainty into account and show that repeated re-labeling of wrongly labeled tweets could improve label quality and model quality. Nevertheless, most often synthetic error rates have been assumed.

To sum up, some works tried to combine informativeness and representative for selecting instances and showed promising results. Nevertheless, none of these 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.

3. Developing Ground Truth Data

In this section, we present our dataset used for our evaluation. We focus on incident-related tweets as a specific type of event-related data. We differentiate between three incident types in order to classify microposts. These have been chosen because we identified them as the most common incident types in the Seattle Fire Calls dataset[1], which is a frequently updated source for official incident information. We also add one neutral class, thus, our final classes are: *car crash*, *fire*, *shooting*, and *no incident*.

As there are no publicly available labeled datasets for event-related microposts, we needed to create our own high-quality ground truth data. For this, we collected English microposts using the Twitter Search API. For the collection, we used a 15km radius around the city centers of Seattle, WA and Memphis, TN. We focused on only two cities, as for our analysis we were interested in a large stream of tweets for a specific time period of certain areas instead of a world-wide scattered sample. This gave us a set of 7.5M microposts from Nov. 19th, 2012 until Feb. 7th, 2013. Although the datasets have been collected in different time periods, we do not expect any difference in the way people post about incidents.

As this initial set was used for conducting our experiments, we had to further reduce the size of the datasets following our approach as described in (Schulz et al., 2013b). The resulting 2,000 tweets were 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 328 fire, 309 crash, 334 shooting, and 1029 not incident related tweets.[2] For our evaluation, we used 1,200 tweets 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.

Table 1. Results for the random error evaluated in a study on quality of crowdsourced labels. Means (μ) and standard deviation (SD) of the error rates are displayed for each user group.

	Random Error	
	Crowd	Expert
μ	0.0338	0.0323
SD	0.0006	0.0002

4. Study on Quality of Crowdsourced Labels

In active learning, most often a perfect oracle is assumed for labeling instances. As this might not hold true in a real-world environment, we conducted a study on labeling accuracy. When it comes to labeling accuracy, the general assumption is that labeling quality in crowdsourcing environments might be dependent on the domain knowledge of the annotators (Zhao et al., 2011). Thus, one of the goals of the study is to analyze if the labeling quality of non-experts differs significantly from domain experts. To answer this question, we evaluated two user groups in our study: *domain experts* and *regular crowd users* with no or limited domain knowledge. Second, there is no work describing error rates for labeling of incident-related microposts. Thus, we want to quantify the error rates, so we can use them for our simulations. For this, we evaluated the *random error*, i.e., the error that results from the annotator carelessness. E.g., a wrong label is occasionally assigned. The random error is regarded as i.i.d. noise on each label, thus, we assume a fixed probability $RE \in [0, 1]$.

We assume a different labeling quality for crowd users (CU) and domain experts (EX) and test the following hypothesis H: The means (μ) of the *random error* are different across both user groups ($H_0 : \mu_{RE,CU} = \mu_{RE,EX}, H_A : \mu_{RE,CU} \neq \mu_{RE,EX}$).

We created a survey to conduct the labeling of our complete ground truth dataset according to the incident types. Fourteen users participated in the study. Eight participants were crowd users with no or low experience in the crisis management domain and six users were domain experts with more than three years experience in the domain. At least three crowd users and at least two domain experts labeled each tweet. Based on the results, we calculate the random error (cf. Table 1) compared to the ground truth labels.

[1] http://data.seattle.gov

[2] All datasets will be published at http://www.doc.gold.ac.uk/~cguck001/IncidentTweets/

For evaluating our hypothesis, we first confirmed normal distribution for all error types and both user groups using the *Anderson-Darling* as well as the *Shapiro-Wilk Normality* test. Furthermore, we conducted a two-sample F-test for variances to verify same variances for all combinations with $p < 0.01$. For each combination we conducted the two-sample t-test assuming equal variances. For all combinations the null hypotheses could not be rejected with $p < 0.01$. Thus, for all error types, we cannot assume a difference between both user groups. This means that in our study there is no conceivable difference between domain experts and common crowd users.

One reason might be the rather low sample size. Others might be found in the nature of microposts as they are short and the amount of available information per tweet is limited. Thus, the complexity of the information is low and it is possible to understand the content even as a non-expert. Furthermore, as tweets are send by lots of different individuals, the number of domain specific terms could be rather low compared to specialized texts. Also, as incident-related tweets are common topics compared to physics or medicine, people are somehow used to the vocabulary.

To reflect a real-world situation best, we combined the results of both groups as in typical crowdsourcing studies both groups might be present. Also, the labels of the experts are available anyway for our dataset. This gave us a final error rate of 0.0331 for the random error.

5. Event-Based Clustering

In this section, we show how active learning can be utilized to classify the incident type of microposts. We also introduce our approach and present how we cope with the initial selection problem, i.e., how to select the initial training set, as well as with the query selection problem, i.e., how to choose appropriate instances for labeling in each iteration.

5.1. 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, a large number of microposts are collected as an initial pool of unlabeled data U. From this information base, a set of training examples L is chosen for learning an initial model. 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 (Kang et al., 2004).

For training a classifier using this initial set, we reuse the classification approach presented in (Schulz et al., 2013b). Here, microposts are processed with standard Natural Language Processing (NLP) techniques such as stopword removal, POS-tagging, and lemmatization. Afterwards, several features are extracted from the preprocessed instances such as word-3-grams after POS-filtering, TF-IDF scores, syntactic features as well as semantic features. The syntactic features are the number of exclamation and question marks as well as the number of upper case characters. The semantic features are a feature group derived using different means of Semantic Abstraction (Schulz et al., 2015). Furthermore, the existing approach allows us to extract a likely date of an event mentioned in a micropost. To identify the temporal information in a tweet, we adapted the HeidelTime framework for temporal extraction as presented in (Schulz et al., 2013b).

As the number of geotagged microposts is rather low (about 1-2%), we reuse an extension of our approach for geolocalization (Schulz et al., 2013a) of microposts as well as for extracting location mentions as features used in the classification. For geolocalization an estimation of the city and the country where a tweet was send from was used and additionally location mentions extracted from the tweet message were considered. First, we use a Stanford NER[3] model to identify all location mentions. Then, the discovered locations are geocoded using the geographical database GeoNames[4], and the MapQuest Nominatim API[5] for more fine-grained locations, like streets. The intersection of all locations extracted from the tweet is used as an estimation of the location where an event mentioned in a tweet has happened.

After the initial training, the 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 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 strategies can be chosen based on informativeness and representativeness (Huang et al., 2010).

For informativeness as selection criteria, uncertainty sampling (Lewis & Catlett, 1994) is commonly applied that selects particularly those examples for labeling for which the learner is most uncertain. However, the main issue with the informativeness approach is that only a single instance is considered at a time (Settles, 2012). Thus, outliers could be selected erroneously as the context is not taken into account. In contrary, clustering helps to identify representative instances. According to Nguyen and Smeulders (Nguyen & Smeulders, 2004), the most representative

[3] http://nlp.stanford.edu

[4] http://www.geonames.org/

[5] http://developer.mapquest.com

examples are those in the center of cluster, which are the instances most similar to all other instances in the cluster. Nevertheless, selecting always the centers of the clusters might result in selecting always very similar instances for each iteration, thus, the model might not improve very much. Furthermore, it remains unclear how many clusters have to be built. Also, the resulting clusters not necessarily correlate to the real-world events as spatial and temporal information is neglected.

To overcome the individual problems of each approach, related work proposes to select the most informative *and* representative instances. This results in selecting the instances that are representative for the whole dataset as well as have the highest chance to improve the model. In our approach, 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 stopping criteria is met, e.g., a maximum number of iterations is reached or when the model does not improve any more.

5.2. 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 event-related information such as temporal and spatial information in combination with the event type to perform an *event-based clustering* to take the properties of real-world events into account. This way, we are directly able to find a number of clusters without the need of specifying the number beforehand. Furthermore, our event-based clustering is based on both selection criteria, so we overcome the limitations of each individual one.

The design of our approach follows the assumption that every event-related information is either related to a real-world event or not. Thus, we propose to cluster all instances based on the three dimensions that define an event: temporal, spatial and thematic extent. As a result, each instance is aggregated to a cluster.

If a micropost lies within the spatial, temporal, and thematic extent of another micropost, it is assumed to provide information about the same event. This assertion can be formalized as a triple of the form $\{event\_type, radius, time\}$. The spatial extent is a radius in meters drawn around the spatial location of the event. The temporal extent is a timespan in minutes calculated from the creation time of the initial event. The thematic extent is the type of an event. For example, for our approach we use the rule $\{Car\_Crash, 200m, 20min\}$, which asserts that each incoming micropost of the event type *Car Crash* is aggregated to a previously reported incident if it is of the same type, within a range of 200 meters, and within a time of 20 minutes. Clearly, altering the radius or the time will have a strong effect on the final clustering. However, as emergency management experts suggested to use these values, we did not change them. Inspecting the effects of different parameterizations remains subject for future work, however, we are confident that our proposed approach is not affected negatively by a change of these parameters. With the help of these three assertion types, a rule engine computes whether microposts are clustered as they describe the same event or not.

Algorithm 1 Algorithm for initial selection strategy.

```
Data: Unlabeled instances $U$, Clusters $C$ generated by event-based clustering,
Size of initial training set $b_i$
Result: Instances to label $L$

for all $clusters\ c \in C$ do
    for all $instances\ i \in c$ do
        Calculate information density $DS(i)$
    end for
end for
for all $clusters\ c \in C$ do
    Calculate average information density $DSC(c)$
end for
Order clusters in $C$ based on $DSC$
while $|L| \leq b_i$ do
    for $cluster\ c \in C$ do
        Add one instance from $c$ to $L$
    end for
end while
```

Microposts containing no thematic information are assigned the $unknown\_event$ type. Missing spatial information is replaced with a common spatial center (the center of a city). Missing temporal information is replaced with the creation date of the micropost. Thus, even with one or two missing dimensions, we are still able to build clusters. 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, microposts not related to events are assigned to larger clusters, containing lots of noise and being less valuable for the learning process.

5.3. Initial Selection Strategy

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

Our approach for selecting the initial dataset is shown in Algorithm 1. 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 training the initial model. For this, we use the event clusters ordered by information density of their containing instances to obtain a good initial set. 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 as this is not present yet. Thus, the $unknown\_event$ type is used. Second, for all instances in each cluster the information density is calculated. This is done based on how many instances are similar or near to each other, thus, outliers are regarded as less valuable. We used a k-Nearest-Neighbor-based density estimation (Zhu et al., 2008): $DS\left(x\right) = \frac{\sum_{s \in S(x)} \text{Similarity}(x,s)}{k}$

The density $DS(x)$ of instance x is estimated based on the k most similar instances in the same cluster[6] $S(x) = \{s_1, s_2, ..., s_k\}$. 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: $DSC\left(c\right) = \frac{\sum_{x \in c} DS(x)}{k}$

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 (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.

5.4. Query Selection Strategy

For the query selection strategy we choose representative using clustering as well as informative instances using uncertainty-based sampling. The pseudo-code is shown in Algorithm 2. 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. Based on this, the event clustering is applied using the spatial, temporal, and thematic information resulting in 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 an uncertainty measure the entropy calculated for each instance x and each class $y \in Y = \{y_1, y_2, ..., y_i\}$ was employed: $H(x) = -\sum_{y \epsilon Y} P(y|x) \log P(y|x)$

Algorithm 2 Algorithm for one iteration of the query selection strategy.

```
Data: Unlabeled instances $U$, Labeled instances $L$, Clusters $C$ generated by
event-based clustering, Number of instances to label per iteration $b_i$, Trained
Model for iteration $M$, Mean average size of all cluster in iteration $ms$
Result: Instances to label $L$

Use $L$ to train classifier $M$
for all $clusters\ c \in C$ do
  for all $instances\ i \in c$ do
    Calculate information density $DS(i)$
    Calculate entropy $H(i)$ using $M$
    Calculate density×entropy measure $DSH(i)$
  end for
end for
for all $clusters\ c \in C$ do
  Calculate $DSHC(c)$
end for
Order clusters based on $DSHC$
while $|L| \leq b_i$ do
  for all $clusters\ c \in C$ do
    $n = log_{ms}(|c|)$
    Add $n$ instances from $c$ to $L$
  end for
end while
```

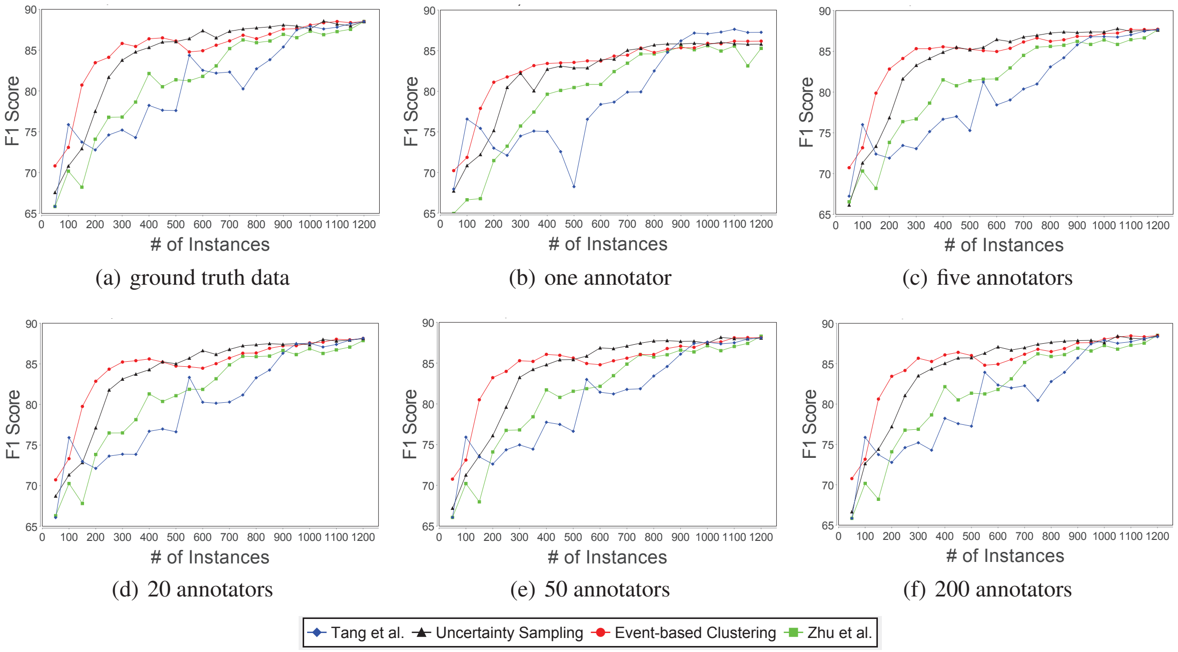
Based on the information density and the entropy, the density×entropy measure $DSH(x) = DS(x) \times H(x)$ (Zhu et al., 2008) is calculated for each instance x. The informativeness and representativeness of each cluster is then computed based on the mean average of DSH of each instance i in the cluster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

For selecting the 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 logarithmic scale by using a logarithm at basis ms to avoid drawing too many instances from larger clusters as would be the case with a linear approach. We assume that drawing only small numbers per cluster is sufficient, as at some point additional instances will not yield any additional information, as the instances will be too similar to each other. Instances are selected 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.

6. Experiments

We conducted two experiments regarding incident type classification. First, we compared related approaches to show that event-based clustering outperforms other clustering-based active learning approaches. Additionally, the effect of the number of labeled instances on the classifier performance is examined. In the second experiment,

[6] k is equal to the number of instances in the cluster.



(a) ground truth data

(b) one annotator

(c) five annotators

(d) 20 annotators

(e) 50 annotators

(f) 200 annotators

Figure 1. Evaluation results of state-of-the-art selection strategies and our approach. The graphs for different number of annotators (regular crowd users) are shown. Note that the more annotators labeled an instance the lower is the probability for a noisy labeling.

the influence of noise is inspected. Noise results from a low number of non-expert labelers. For keeping costs low a classifier should not be affected by a bad labeling.

6.1. Classification and Metrics

The active learning algorithms select instances from the training set to query for labels. Based on these, a classifier was trained and evaluated on the test set. As classifier we used Weka's implementation of John Platt's sequential minimal optimization (SMO) algorithm for training a support vector machine (Platt, 1998). Due to the complexity of determining best parameters for each iteration and each approach, we followed related approaches (see (Huang et al., 2010) and (Donmez et al., 2007)), and decided to compare all algorithms on fixed parameters. Consequently, the SVM was used with standard settings.

For comparison, the deficiency metric (Raghavan et al., 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 largest F1 score and the learning curve of the reference algorithm REF. Thus, the measure is nonnegative and values smaller than 1 indicate more efficient algorithms compared to the baseline strategy, whereas a value larger than 1 indicates a performance decrease compared to the baseline strategy.

6.2. Algorithms and Parameters

In order to evaluate the performance of our approach, we re-implemented the following related approaches:

(Tang et al., 2002): For initial sampling a k-means clustering is used. For query selection, first the most uncertain instances for each cluster are selected. Then, information density is used to weight the examples. We set $k = 4$, because we have four different event types.

(Zhu et al., 2008): For initial sampling a k-means clustering is used ($k = 4$). During the iterations, the entropy $\times$ density measure is used as selection criteria and no clustering is applied.

Uncertainty: Random instances (initial) and the entropy-based uncertainty sampling (iterations) is used.

Event-based clustering: Our event-based clustering is applied with a spatial extent of 200m and a temporal extent of 20min.[7][8]

Following the experimental settings of (Hu et al., 2013) and (Huang et al., 2010), we set the size of the initial training set and the size during the iterations to 50. No further tuning or parameterization was applied. Each iteration for

[7] As a result, the 1,200 tweets of the training set are divided into 438 distinct event clusters.

[8] The spatial and temporal extent are a result of discussions with emergency managers.

Table 2. Deficiencies with Tang et al. as a baseline strategy.

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

each algorithm was repeated 10 times, as for instance, the uncertainty approach is highly dependent on the selected instances. We used averaged F1 based on the repetitions.

6.3. Comparison to state-of-the-art approaches

The performance graph for the ground truth data is shown in Figure 1 (a). Note that the x-axis shows the *total* number of instances combining the 50 instances of the initial training set and 50 instances drawn per iteration. Thus, iterations from 0 to 23 are depicted. As shown, 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 2 shows the deficiencies. 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 that 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 events 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 for instance only a sixth of all instances are needed to achieve a F1 score of about 84%. 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 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.

Influence of noise in the labels In Figure 1 (b) and 1 (c), the learning curves for the very error-prone cases with one respective five annotators are shown. As can be seen in the curve of the approach of Tang et al., the influence of noise is notable in the big drop with 500 instances. Also Zhu et al.'s approach has a much lower initial F1 score compared to all others, which is an indicator for an inappropriate initial selection strategy. The results indicate that even with noisy labels, our approach outperforms the state-of-the-art as the situation in the graphs of the lower part of Figure 1 does not change much. In all these cases, the learning curves are quite similar, which is a result of the decreased number of wrongly labeled instances. Clearly, the performance of all approaches increases with a lower number of errors.

As we showed, our approach outperforms related work also if noise is taken into account. Not surprisingly, we found that with an increasing number of annotators, noise is negligible. With only one annotator, the deficiency is worse by 57% and with five annotators still worse by 26%. Even with 50 annotators, the deficiency still is worse by 10%. For more than ten annotators, an F1 score of 85% is reached comparably fast. With a maximum of five annotators, this level is only reached at the end of the simulation. For one annotator, this maximum is never achieved. These results indicate that a minimum number of annotators is needed for achieving good results by crowdsourced labeling tasks. In our experiments, ten annotators seem to be sufficient, while in other domains with different error rates, there might be a need for much more annotators.

7. Conclusion

We presented an event-based clustering strategy for event type classification of microposts and coped with several problems of active learning in the emergency management domain. First, it was shown that domain experts do not differ significantly from regular crowd users when it comes to labeling quality. Second, we presented a novel selection strategy for active learning based on temporal, spatial, and thematic information. Our event-based clustering that identifies representative as well as informative instances outperforms state-of-the-art clustering approaches. On incident-related microposts we showed that a better initial training set is selected as well as to appropriate instances for labeling in each iteration are chosen. The learning curve indicated that only a sixth of all instances are needed to achieve a F1 score of about 84%, which is especially important when it comes to labeling costs, as only a limited amount of data would need to be labeled to achieve good classification results.

In the future, we aim at using our active learning framework in addition to labeling of single features. Furthermore, though our framework follows a general approach, we only evaluated it on incident-related data, thus, we also want so show the applicability on other types of events.

References

Donmez, Pinar and Carbonell, Jaime G. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *CIKM'08*, pp. 619–628, 2008.

Donmez, Pinar, Carbonell, Jaime G., and Bennett, Paul N. Dual strategy active learning. In *ECML'07*, pp. 116–127, 2007.

Hoi, Steven C. H., Jin, Rong, and Lyu, Michael R. Large-scale text categorization by batch mode active learning. In *WWW'06*, pp. 633–642, 2006.

Hsueh, Pei-Yun, Melville, Prem, and Sindhwani, Vikas. Data quality from crowdsourcing: A study of annotation selection criteria. In *NAACL HLT'09*, pp. 27–35, 2009.

Hu, Xia, Tang, Jiliang, Gao, Huiji, and Liu, Huan. Actnet: Active learning for networked texts in microblogging. In *SIAM'13*, 2013.

Huang, Sheng-Jun, Jin, Rong, and Zhou, Zhi-Hua. Active learning by querying informative and representative examples. In *NIPS*, pp. 892–900, 2010.

Kang, Jaeho, Ryu, Kwang R., and Kwon, Hyuk C. Using Cluster-Based Sampling to Select Initial Training Set for Active Learning in Text Classification. In *PAKDD'04*, pp. 384–388, 2004.

Lewis, David D. and Catlett, Jason. Heterogeneous uncertainty sampling for supervised learning. In *ICML'94*, pp. 148–156, 1994.

Nguyen, Hieu T. and Smeulders, Arnold. Active learning using pre-clustering. In *ICML'04*, pp. 79–86, 2004.

Platt, J. Fast training of support vector machines using sequential minimal optimization. In Schoelkopf, B., Burges, C., and Smola, A. (eds.), *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.

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

Schulz, Axel. *Mining User-Generated Content for Incidents*. PhD thesis, TU Darmstadt, 2014. URL http://tuprints.ulb.tu-darmstadt.de/4107/.

Schulz, Axel, Hadjakos, Aristotelis, Paulheim, Heiko, Nachtwey, Johannes, and Mühlhäuser, Max. A multi-indicator approach for geolocalization of tweets. In *Proceedings of the Eight International Conference on Weblogs and Social Media (ICWSM)*, pp. 573–582, Menlo Park, California, USA, 2013a. AAAI Press.

Schulz, Axel, Ristoski, Petar, and Paulheim, Heiko. I see a car crash: Real-time detection of small scale incidents in microblogs. In *The Semantic Web: ESWC 2013 Satellite Events*, volume 7955 of *Lecture Notes in Computer Science*, pp. 22–33. Springer Berlin Heidelberg, 2013b.

Schulz, Axel, Guckelsberger, Christian, and Janssen, Frederik. Semantic abstraction for generalization of tweet classification: An evaluation on incident-related tweets. In *Semantic Web Journal: Special Issue on The Role of Semantics in Smart Cities (to appear)*. IOS Press, 2015.

Settles, Burr. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012.

Shen, Dan, Zhang, Jie, Su, Jian, Zhou, Guodong, and Tan, Chew-Lim. Multi-criteria-based active learning for named entity recognition. In *ACL'04*, 2004.

Tang, Min, Luo, Xiaoqiang, and Roukos, Salim. Active learning for statistical natural language parsing. In *ACL'02*, pp. 120–127, 2002.

Thongsuk, Chanattha, Haruechaiyasak, Choochart, and Meesad, Phayung. Classifying business types from twitter posts using active learning. In *I2CS'10*, pp. 180–189, 2010.

Tong, Simon and Koller, Daphne. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, 2002.

Wallace, Byron C., Small, Kevin, Brodley, Carla E., and Trikalinos, Thomas A. Who should label what? instance allocation in multiple expert active learning. In *SDM'11*, 2011.

Zhao, Liyue, Sukthankar, G., and Sukthankar, R. Incremental Relabeling for Active Learning with Noisy Crowdsourced Annotations. In *SocialCom'11*, pp. 728–733, 2011.

Zhu, Jingbo, Wang, Huizhen, Yao, Tianshun, and Tsou, Benjamin K. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *COLING'08*, pp. 1137–1144, 2008.