# Music Event Detection Leveraging Feature Selection based on Ant Colony Optimization

Jian Xi*†, Michael Spranger† and Dirk Labudde†‡
†University of Applied Sciences Mittweida
Forensic Science Investigation Lab (FoSIL), Germany
Email: {xi, spranger}@hs-mittweida.de ‡Fraunhofer
Cyber Security
Darmstadt, Germany
Email: labudde@hs-mittweida.de

***Abstract*—Announcements of events are regularly spread using the Internet, e.g., via online newspapers or social media. Often, these events involve playing music publicly that is protected by international copyright laws. Authorities entrusted with the protection of the artists' interests have to find unregistered music events in order to fully exercise their duty. As a requirement, they need to find texts in the Internet that are related to such events like announcements or reports. However, event detection is a challenging task in the field of Text Mining due to the enormous variety of information that needs to be considered and the large amount of data that needs to be processed. In this paper, a process chain for the detection of music events incorporating external knowledge is proposed. Furthermore, a feature selection algorithm based on ant colony optimization to find featurse with a high degree of explanatory power is presented. Finally, the performance of five different machine learning algorithms including two learning ensembles is compared using various feature sets and two different datasets. The best performances reach an $F_1$-measure of $0.95$ for music texts and $0.968$ for music event texts, respectively.**

***Keywords–Event Detection; Text Classification; Named Entity Recognition; Feature Selection; Ant Colony Optimization.***

## I. INTRODUCTION

In a highly connected world, in order to assure the rights of artists, it is of utmost importance to develop an automatized solution to retrospectively detect violations against copyrights and exploitation rights related to music events. In [1], a first approach towards such a system for online data was proposed. This paper deepens the discussion and proposes an additional feature selection method based on ant colony optimization (ACO).

Individuals, groups and organizations can infringe artists' copyrights in different ways. Whereas individuals might create unauthorized copies, groups and organizations responsible for public events or artists playing at them might play music or show movies without respecting the copyright interests of artists, either deliberately or due to ignorance of the law. Pursuing individual interests of artists is difficult to realize due to practical reasons (e.g., lack of information). Therefore, authorities or private institutions are entrusted with the artists' interests. Usually, organizers register music or movie titles, which are going to be played at an event together with the expected number of participants with these representatives and by paying for the licenses receive the right to play these titles. The official authorities and institutions of the artists then transfer the license fees to the respective artist. One of the largest private institutions in Germany is the Gesellschaft für musikalische Aufführungs- und mechanische Vervielfältigungsrechte (GEMA, English: Society for musical performing and mechanical reproduction rights) representing the rights of about 2 Million artists all over the world and with a total revenue of 1 Billion Euros a year [2]. So far, finding unregistered events after they have taken place is very difficult and is a process mostly done manually.

Nowadays, the information that an event is taking place is often spread using online newspapers, Facebook, Twitter as well as websites. Additionally, after an event has taken place it is often discussed using the same means of communication. Spreading the information this way is often the first choice, as many people can be reached in a short amount of time. Hence, analyzing these textual data makes it possible to automatically find the information needed to uphold the artists' rights. Text Mining, also referred to as Text Analysis, focuses on the analysis of texts in order to receive high level information and latent patterns. It can be applied to many different areas [3], however, it plays a special role in forensics and predictive policing, where it can be used to detect events with a potential to escalate [4]. Event detection is a specific Text Mining problem in which texts are analyzed in order to mine a set of texts that have a semantic link or share conceptual patterns regarding past or future events [5] [6].

The study specifically addresses the detection of music events announced or talked about in social media and online newspapers, with the intention to find those events where copyrights might be violated. Because of the vast amount of data that needs to be taken into account, the data can only be effectively analyzed using machine learning techniques and methods applied in automatized text classification [7]. Since the selection of features is a critical step due to the "curse of dimensionality" problem [8], the Ant Colony Optimization (ACO) algorithm was used for feature selection. To the best of our knowledge, there have been no studies, so far, using ACO for event detection problems. As shown in Section V-B, when using ACO for the selection of features, the experimental results slightly improve compared to the results achieved using the approach proposed in [1]. Additionally, the number of features representing the documents decrease dramatically.

This paper is organized as follows: In Section II, some re-

lated work is briefly reviewed. Section III describes difficulties in the current domain as well as the development of a gold standard. The proposed concept is explained in Section IV. Details about the experimental evaluation, the results as well as a short discussion can be found in Section V. Finally, in Section VI a short conclusion is given and some aspects of future work are discussed.

## II. Related Work

As mentioned above, event detection is a special text mining problem and can also be seen as a classification problem [9]. However, first it needs to be clarified how an event is defined. In this study, we chose a definition based on the ontology by [10]. Accordingly, an event is defined by the presence of agents at a specific time and place, who are engaged with or in a common matter (product) under concomitant circumstances (factors). Event detection was initiated by the Topic Detection and Tracking research program [11], yet, only focuses three of the original five tasks, namely: tracking, detection and first story detection [12].

In the past years, several approaches have been developed for closed and open domains. For the former manually designed keyword lists can be used to detect specific events in texts [13]. Those keyword lists work effectively, yet need expert knowledge to define the event-specific keywords. Furthermore, keyword lists are limiting the search framework, which is why they will not work for open domains and can only be used as an additional resource for more complex event types, as is the case with the detection of music events. Another example for the detection of events within a specific field is presented by [14] and [15], both working on the detection of economic events that might influence the market, such as mergers. For open domains, [6] proposed a method using machine learning techniques, like clustering and Named Entity Recognition (NER) combined with an ontology (DBpedia) in order to classify Tweets into eight predefined event categories.

Similar to event extraction, the recognition of events might also be categorized as data-driven or knowledge-driven event recognition. In [14] and [15], data-driven approaches were used, both taking mentions of real-world occurrences into account in order to classify their texts into different types of economic events. However, the data-driven approaches fail to consider semantics. In contrast, knowledge-based approaches focus on mining patterns from data to deliver potential rules representing expert knowledge. Depending on the domain or the context, linguistic, lexicographic as well as human knowledge or a combination of these is applied [16].

Much work has been done concerning event detection using different approaches within different fields. Certainly, some of the proposed methods, such as those presented in [13] and [6], can be applied for the detection of music events and our concept is based on the work by [6]. However, in the domain of music event detection, some difficulties appear. For example, events might be announced only using the name of an artist. Some of these difficulties will be discussed in later sections.

One example for a study on music events working with Twitter data is given in [17]. In their study, they identify musical events mentioned in Twitter in order to create a list including sets of artists and venues. The information can be added to an already existing list, for example, a city event calendar [17].

## III. Data Preparation

Since the nature of the data is very heterogeneous – different sources like Facebook and newspapers are considered – its analysis has inherent challenges. Below, some of them are discussed in more detail.

### A. Data Sources

At the beginning of the study, experts, during their work on manually detecting unregistered music events, independently and arbitrarily preselected more than 1000 music event relevant and irrelevant texts from Facebook and online newspapers. This dataset was then annotated as presented below and used as a basis for our gold standard.

### B. Challenges

***Noisy Data:*** In general, texts from social media are inherently characterized by noise. For example, texts often include web addresses, telephone numbers, dates and other characters like hashtags. Furthermore, the texts posted, for example, on Facebook or Twitter are not well written in terms of their grammar and orthography. The application of standard NLP tools to correct such mistakes may lead to incorrectly written names of musicians. As these names are crucial for this study, important events may not be detected.

***Text Length:*** Due to technical restrictions and their intended usage, texts in social media are often very short. Information is compressed as much as possible, for example, by using emoticons or abbreviations or by completely leaving out words. Therefore, the application of standard text analysis methods is often difficult, especially, if the method relies on syntactically correct structures. Considering the following text from Facebook, the application of standard Named Entity Recognition methods fails, because some syntactic features are missing:

"Foo Fighters Eintritt 19. in Hamburg"

***Latent Information:*** Taking the example from above, the crucial information that needs to be found is – even if the text is already classified as an event – that Foo Fighters is a band name and, therefore, the text announces a music event. Typically, such information is extracted by applying methods from the field of NER as discussed in [18]. Traditionally, NER is a subtask in the field of information extraction that focuses on locating structured information in a text and assigning it to predefined categories such as names of persons, organizations and locations. However, distinguishing normal persons from singers or normal organizations from bands is challenging

and presents one of the biggest problems in the selection of appropriate features as no prior information is available that indicates whether what the NER model identified is really music-related. This can be changed by adding additional information in the gazetter. This means, before the classification it is already known that, i. e., Johann Sebastian Bach is a musician. However, a much more challenging task is the identification of entities in a text such as musicians that are unknown, for example, a new band or DJ. Unfortunately, texts including these entities appear more often than texts announcing events with known entities.

***Dynamic Entities:*** Information is always dynamic and changes in meaning depending on the time of production. The latent new NER-entities (e.g., musicians, bands or groups) change over the time. An example would be the singer and songwriter Ed Sheeran. Before he became a known musician, he would need to have been labeled as a normal person. However, now he needs to be labeled as a musician. This means, which named entities are relevant changes depending on the point of time a text was written. This triggers the requirement to simultaneously update the knowledge base of our system.

### C. Gold Standard

Because there are no suitable training data available, it was necessary to create a gold standard as a basis for the training and evaluation of various classification models. As was mentioned above, texts were collected arbitrarily, including 21 texts from online newspapers and 1,097 texts from Facebook. These were manually annotated as music related or music unrelated as well as event related or event unrelated. Both decisions were made independently of each other. Due to text-inherent vagueness, the data was independently labeled by 35 people. In order to ensure the quality of the labeled data, each person was only allowed to work for 2 hours a day.

The final decision regarding what category a text belongs to was made by using a majority criterion. This criterion requires a minimum number of people to agree on a decision in order to provide a confident classification. If the minimum number of agreements was not achieved for a given text, the text was considered ambiguous and removed from the corpus. The minimum number of agreements was derived from a binomial test under the null hypothesis that each decision individually made by every study participant is conducted at random. This hypothesis thus states that $p^+ = p^- = 0.5$, where $p^+$ and $p^-$ are the decision probabilities. With respect to the null hypothesis, for every number of agreements $d$ a probability $P(d|p^+)$ can be derived from the corresponding binomial distribution. The minimum number of agreements $d_{crit}$ is equal to $d$, where the null hypothesis can be rejected according to $P(d \geq d_{crit}|p^+) < \alpha$. Here, $\alpha$ corresponds to the Bonferroni-corrected significance level of $0.05/n$, with $n$ being the number of considered texts. In this study, the minimum number of agreements $d_{crit}$ was 29 for the text corpus.

As a result, the corpus consists of 19 newspaper texts and 867 Facebook texts. 335 out of the 867 Facebook texts and 14 out of the 19 newspaper texts are music relevant. Table I provides some descriptive statistics. When music event classification is considered, the number of texts that meet the Bonferroni constraint drops to 505, whereas 251 Facebook texts and 9 online newspaper texts are music event relevant. Table II provides the descriptive statistics for the music event related data. In summary, at the end, two datasets were created: one for music relevance, including 886 texts and one for music event relevance with 505 texts.

TABLE I. STATISTICS OF THE DATA REGARDING MUSIC DETECTION.

| | *# texts* | $\#_{tot}$ *words* | $\#_{avg}$ *words* | shortest | longest |
|---|---|---|---|---|---|
| newspaper | 19 | 2,071 | 109 | 14 | 387 |
| Facebook | 867 | 85,965 | 99.1 | 1 | 1,238 |
| total | 886 | 87,965 | 99.3 | 1 | 1,238 |

TABLE II. STATISTICS OF THE DATA REGARDING MUSIC EVENT DETECTION.

| | *# texts* | $\#_{tot}$ *words* | $\#_{avg}$ *words* | shortest | longest |
|---|---|---|---|---|---|
| newspaper | 13 | 1,077 | 82.85 | 14 | 277 |
| Facebook | 492 | 59,440 | 120.81 | 1 | 1,238 |
| total | 505 | 60,517 | 119.84 | 1 | 1,238 |

In order to describe the data in the domain of music events, we defined an XML-schema, with which our raw data can be concisely structured in order to serve as a gold standard to train and test models in this field. Even though this work is focused on music event detection, the schema is constructed to contain various types of event data, such as music, theater, or readings. It includes, beside others, the following information:

- raw text
- source (e. g., Facebook)
- event-related ($\{0, 1\}$ and certainty)
- event-type-related ($\{0, 1\}$ and certainty)
- event location
- event-date
- persons
- different types of roles (e. g., musician, actor)
- different types of events (e. g., music, theater)

It needs to be emphasized that the relation between any text and a specific category is described twice: binary and with a numeric value. The binary description refers to the classification and thus serves as a ground truth, whereas the numeric value represents the degree of certainty. With this gold standard the following areas may be addressed:

- classification of texts regarding different event-types
- recognition of event-related entities, i. e., roles of persons, organizations and locations

Named entities are considered because they provide strong features for the classification, as was shown in [19] and [20]. For example, if the name Eric Clapton, an English singer and songwriter, appears in a text, this is a strong indication that the current text is music related. Since classic NER mostly

concentrates on distinguishing between persons, locations, and organizations, a more detailed categorization including some kind of prior knowledge is needed. The entire dataset was annotated and curated manually according to the schema described so far.

## IV. Proposed Concept

The task of detecting texts concerning music events is a typical categorization task. Categorization, as a special case of classification, attempts to categorize a text into a predefined set of conceptual categories using machine learning techniques. Formally, let $T = t_1, ..., t_m$ be a set of texts to be categorized, and $C = c_1, ..., c_n$ a set of categories, then the task of categorization can be described as surjective mapping $f : T \rightarrow C$, where $f(t) = c \in C$ yields the correct category for $t \in T$. In the field of music event detection, texts need to be assigned to one out of two main classes: related to a music event or not. Texts of the former class can be further categorized into different event types, such as public concerts. This might be of great importance as some music, e. g., religious music or classical music concerts, are license free or public music resources.

Currently, institutions responsible for the enforcement of exploitation rights have to detect unannounced music events predominantly manually and with the help of search engines. This leads to various problems. Firstly, the manual search is very inefficient on large-scale data. Secondly, the manual checking process is error-prone and differs depending on the person who judges the data. Furthermore, the current process chain can hardly be deployed in an online mode due to its semi-automated nature.

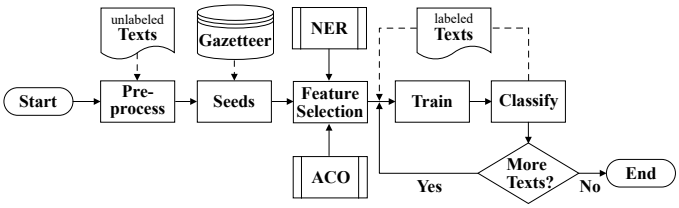


Figure 1. The proposed workflow of music event detection.

To overcome these limitations, a semi-supervised process-chain using a bootstrapping approach, as depicted in Figure 1, is proposed. The advantage of the chosen approach is that the training can start with very few but highly descriptive examples in order to create a first restrictive classifier which will be further improved in upcoming iterations until all texts are classified or no further improvement is possible. Next, each step is discussed in more detail.

### A. Preprocessing

As mentioned in Section III-B, the texts we worked on mostly come from the Internet. Such texts often contain typing errors and are often written in informal language, including dialect. This leads to even noisier data than usual in textual texts. Besides common shallow text preprocessing, including stopword and punctuation removal as well as stemming or lemmatizing, there is a strong need for additional language information. This information can be provided in the form of a knowledge base curated by experts. For instance, preselected terms, such as party or live music, can be used to build the gazetteer. Additional useful information might be venues of interest, such as clubs or cafés, where music events often take place. In short, information directly related to music events can be used as a basis of knowledge. This knowledge base can be a simple gazetteer, as is the case in our study, or can incorporate more complex structures, as in [21].

### B. Collecting Seed Texts

The most crucial task in bootstrapping is finding seed texts which represent the concept of the classes as well as possible. The usage of some kind of highly descriptive key words or phrases collected from experts in this field is one possible way to find seed texts in a highly accurate, but, nevertheless, very restrictive way.

### C. Named Entity Recognition (NER)

As was shown in [18]–[20], named entities might be a useful feature for text classification tasks. In a first step, named entities are identified using any NER method, as discussed in [18]. However, as was already discussed in Section III-B, the named entities detected this way are not specific enough. Hence, domain-specific knowledge resources like MusicBrainz, an open music encyclopedia, and DBpedia can serve as a music database for distinguishing recognized entities further, in order to assign appropriate roles to them, for example, `musician` to a person. The richer and up-to-date this knowledge base, the more accurate is the classification. The entire process of music event related Named Entity Recognition is shown in Figure 2. The influence of using NER with a knowledge base is clearly shown in Section V-B.

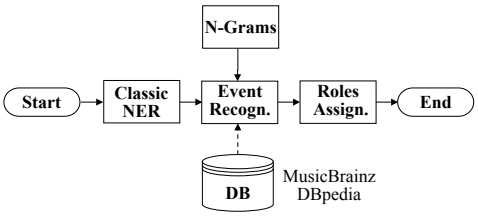


Figure 2. The proposed workflow of detecting music related named entities.

### D. Feature Selection

The next step is the selection of appropriate features to represent the text data. Feature selection is always a critical step in text classification tasks. On the one hand, well selected

features are necessary to achieve highly accurate results. On the other hand, they help reduce the feature space and, as a consequence, minimize the time complexity [22].

Generally, feature selection, as a typical machine learning task, can be distinguished in supervised, unsupervised and semi-supervised approaches depending on whether the data is labeled or not [23]. The approach for music event detection used in this paper relies, at least initially, on the availability of labeled training data. Thus, it belongs to the algorithmic class of supervised methods. Below, three general approaches in supervised feature selection are briefly introduced:

1) **Filter Approach**: Here, the explanatory power of features is measured by using intrinsic properties of data to select the best feature without employing predictive models. The principal components of methods based on this approach are the feature search and selection criterion. The feature can be ranked according to the score evaluated by statistical measures such as the chi-squared test or the information gain [22]. The ranked features can be either removed or selected from feature sets by comparing their explanatory power with a given threshold. Thus, the most challenging part of this method is to select the proper feature candidates [24].
2) **Wrapper Approach**: Methods based on this approach utilize a machine learning model to evaluate the explanatory power of the elements in the feature set. Hence, the selection of proper features can be considered as a search algorithm. In this case, candidate features are found using a search strategy like heuristic search, sequential forward selection, or backward elimination [24]. The selected features are then used to train a predictive model and evaluate the fitness of the selected features by utilizing hold-out data. The overall procedure is repeated until the desired quality criterion is met. In this approach, the trained model can be seen as a black box and its representational bias is decreased by repeatedly applying the feature search process (i.e., by means of cross validation). The feature selection method based on ACO as proposed in this paper is inspired by this approach.
3) **Embedded Approach**: Here, a linear classifier such as SVM or logistic regression is chosen in order to learn a predictive model and, at the same time, select the most appropriate features. An additional regularization function is usually included in order to constrain the learning of the coefficients of this model. The features, whose coefficient is non-zero, are selected [23].

***Ant Colony Optimization Algorithm:*** Heuristic algorithms and especially those using swarm intelligence, such as ACO, are developed to solve combinatorial optimization problems. They are based on the natural principles of self-organization as a mechanism to control the behavior of the individuals of the swarm. In this sense, the ACO is inspired by observing the behavior of ants in an ant colony and was first reported in [25]. An ant population is able to find a solution for a combinatorial problem, which of course might be sub-optimal, by mimicking the ants' forage by performing randomized walks between food resources and their colony.

This highly coordinated behavior between the individuals in an ant colony can be transferred to solve computational tasks. In this case, each ant iteratively tries to find a possible solution to solve the problem taking into account its own current heuristic information as well as additional information propagated by other ants. In this way, the population as a whole is able to find one, though maybe sub-optimal, consensus solution. Whenever an ant finds a candidate solution, the information about the path this ant has walked is spread throughout the population guiding following individuals. This kind self-organization is made possible by the chemical odoriferous redolence, called *pheromone*, which is left by each individual of the colony.

During the search, each ant contributes proportionally to the final optimal solution. The repeated communication between individuals happens in-directional by changing their environment. Together the pheromone paths, left behind by the interaction between the individuals, form a pattern which is bigger than the pattern of each individual [26]. This collaborative behavior pattern is called *stigmergy* [27].

***ACO based Feature Selection:*** In [28], ACO based feature selection was used to select feature subsets for different disease data sets. An artificial neural network was trained as a predictive model in order to evaluate each feature subset. It is also shown that the selection of feature can improve the performance of a classifier. In a different study a variation of ACO, *AntMiner+*, was used to mine rules that provide information about the decision making process, i.e., considering a couple of given attributes such as amount of bank deposits, duration of the deposits, or credit histories [29]. Consequently, a rule-based relation is extracted in order to classify the customer's creditworthiness [29]. Although the time complexity of ACO is the highest in this study, it still provides plausible results [29].

In the field of text classification, the feature space is usually large. ACO-based feature selection works under the assumption that inter-variable relations among a reduced feature subset represents the original data in a way that the predicted results are accurate [28]. As shown in [30], the ACO-based feature selection outperforms genetic algorithms, information gain, and the chi-squared test on the Reuters-21578 data set. Similar results were achieved by [31] for the classification of web pages using a decision tree, Naïve Bayes and $k$-NN each combined with an ACO-based feature selection method. Using the same data set as [28], [32] also effectively applied a ACO-based feature selection method. The main difference between the two studies are the strategies for state transitions and the pheromone updates. Subsequently, the feature selection approach used in this study is based on these two studies.

When applying ACO, in a first step the problem should be represented as a graph, whereas each node in the graph represents a single feature in the feature space [25] [28] [29] [32]. In this study, all the tokens are stemmed by using the same pre-processing method as introduced in Section IV-A. Following the principle of ACO, those features selected

by the ants, which lead to better test results, get a higher pheromone concentration and higher heuristic values and are, consequently, selected more frequently. Hence, there is no additional weighting for features necessary, however, a strategy for initializing the features in feature selection is needed. Here, we experimented with two strategies, one using a constant value and the other Mutual Information (MI). The features are selected and evaluated according to their desirabilities and contributions to the fitness. The details are shown in Section IV-D. This process is repeated until the desired explanatory power of the resulting feature set is reached.

**Algorithm 1** Proposed Feature Selection Algorithm with ACO

**Require:** All features
**Ensure:** The best feature subsets $S^k_{best}$
1: Initial colony information $\tau, \eta$
2: Generate Ants $\mathcal{A}$
3: $i \leftarrow 0$
4: **while** $i \neq K$-Fold **do**
5: Prepare training/Test data in $i$-Fold
6: **for** $l = 0$ to $L$-iteration **do**
7: **for** ant $k \in \mathcal{A}, t \in \mathcal{T}$ **do**
8: Construct feature subsets $S^k(t)$ in $i$-Fold
9: **end for**
10: **if** all constructions are finished **then**
11: Evaluate feature subsets $S^k(t)$ in $i$-Fold
12: **if** stop condition satisfied **then**
13: Return the feature subsets $S^k_{best}$
14: **else**
15: Update colony information $\tau, \eta$
16: Reset ant memory
17: Generate Ants $\mathcal{A}$
18: Go to step 5
19: **end if**
20: **else**
21: Go to step 5
22: **end if**
23: **end for**
24: $i \leftarrow i + 1$
25: **end while**

***Algorithmic Details**:* The feature selection algorithm used in this study is shown in Algorithm 1 and explained in more detail below:

*Initializing the Colony:* The pheromone level $\tau$ and the heuristic information $\rho$ should either be initialized with a constant value as suggested by [28] and [32] or with more informative values like the information gain or the Pearson Correlation Coefficient as suggested by [33]. In this study, the initialization of the pheromone level and the heuristic information is done once assigning a constant value and once assigning MI.

*Generating Ants:* As suggested in [32], the number of ants $\#ants$ should equal the number of features in the feature space $\#features$. Unfortunately, in our case, there are at least 10,000 features. Therefore, in order to reduce the computational complexity only $\#ants = \#features/100$ were created. Note that we only decreased the number of ants in the ant colony, however, the features to be crawled stayed unchanged.

*Constructing the Feature Sets:* Each ant has a limited capacity to hold the features it crawled. In [32], the roulette-wheel schema was used to decide the size of the feature subset $\mathcal{T}$ for each ant given a hyper-parameter $\mu$. In this study, $\mu$ is set to $\mu = 0.35$. This factor influences how many features an ant can take during its trail constructing process. The next problem is how the ants pick a particular feature (i.e., a node in feature graph) in such a way that the classification accuracy is maximized.

The artificial ants move between features under a *pseudo-random proportional rule*. In other words, a probability decision policy guides the ants' walk through the adjacent features in the search space. In [32], the state transition does not consider exploration but only exploitation, while [28] uses a random number drawn from a uniform distribution in order to control exploration and exploitation. In order to get informative, yet still compact features without losing generalization the same state transition policy as the one in [28] was used in this study.

The probability of a feature $i$ at time step $t$ to be selected by ant $k$ is defined in Equation (1):

$$P^k_i(t) = \begin{cases} \arg\max[\tau_i(t)]^\alpha \times [\eta_i(t)]^\beta, & \text{if } (q < q_0) \\ \frac{[\tau_i(t)]^\alpha \times [\eta_i(t)]^\beta}{\sum_{u \in j^k} [\tau_u(t)]^\alpha \times [\eta_u(t)]^\beta}, & \text{if } (u \in j^k) \end{cases}, \quad (1)$$

where factor $\alpha, \eta$ controls the influence of the pheromone and heuristic of feature $i$ at time step $t$ on the transition, respectively. The random number $q \in [0, 1]$ drawn from a uniform distribution, controls the trade-off between exploration and exploitation. The hyper-parameter $q_0$ forms the threshold value for this trade-off and $j^k$ is the set of possible features that can be taken by ant $k$ at time step $t$.

*Checking whether all constructions are finished:* the construction step repeats until all ants have finished constructing their tours.

*Evaluate Feature Subsets:* In this step, each feature subset $S^k(t)$ is used to train an arbitrary machine learning model. Here, the Multi-Layer-Perceptron algorithm is used with the same parameter settings as proposed in former experiments described in [1], where it outperformed all other considered models.

Afterwards, the features are evaluated with hold-out data. The pheromone concentration is proportionally updated depending on the fitness of the features $f(S_k(t))$. Additionally, the fitness of the features also gives information about how good a feature is and, thus, its desirability. In this study, a 3-fold cross validation was used to evaluate the features. Within each fold, out of all the best local feature subsets $S^l(t)$, calculated during $L$ iterations, here $L = 10$, the best global feature subset was determined. During an iteration, if the fitness of the selected best local features does not change anymore, the iteration breaks.

This way, the colony does not only focuses its search for optimal features in a fraction of the dataset, yet in the whole

training dataset, i.e., feature space [32]. The best local and global feature subset contribute to the fitness of features as is explained in the following step.

When a pre-defined stop criterion is reached, the best feature subset will be returned and the procedure is terminated. Otherwise, the colony information such as pheromone and heuristic of features is updated with respect to their contributions.

*Updating the Colony Information:* Based on the fitness of features determined during the previous feature evaluation step, the pheromone values and the heuristic information of each feature are updated. In this study, the same update rules as in [32] were applied (see Equation (2)).

$$\tau_i(t+1) = (1-\rho) \times \tau_i(t) + \frac{1}{m_i} \sum_{k=1}^{\#_{ant}} \Delta_i^k(t) + e \times \Delta_i^g(t), \quad (2)$$

where $\rho$ is the evaporation parameter of the pheromones, $m_i$ the absolute frequency of the feature $i$ considering all feature subsets and $e$ the elitist parameter which decays the contribution of features occurring in the best local feature subsets. $\Delta_i^k(t)$ is defined as the fitness of those features that are selected by the ants during their current tour and is calculated as shown in Equation (3).

$$\Delta_i^k(t) = \begin{cases} f(S^k(t)) & \text{if } i \in S^k(t) \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

$\Delta_i^g(t)$ is defined as the sum of the fitness of those features that occur in the best local feature subsets. Equation (4) shows how it is calculated.

$$\Delta_i^g(t) = \begin{cases} f(S^l(t)) & \text{if } i \in S^l(t) \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

Subsequently, both results are normalized by their occurrence frequency. The decayed pheromone of each feature is incrementally updated by considering local updates from all tours and the penalized global updates from the elitist tours. The heuristic of each feature is updated as shown in Equations (5) and (6).

$$\eta_i(t+1) = (1-\rho) * \eta_i(t) + \Delta_{\eta_i}, \quad (5)$$

where

$$\Delta_{\eta_i} = \frac{1}{m_i} \sum_{k=1}^{\#_{ant}} f(S_i^k(t)) \times \left[1 + \phi \times \exp\left(-\frac{|S_i^k(t)|}{\#_{feature}}\right)\right], \quad (6)$$

and $\phi$ weights the exponential ratio between $|S_i^k(t)|$, the length of selected feature subset by $k$ in step $t$ in which the feature $i$ occurs, and the total size of the feature space. Those features that are associated with a reduced subset and yield a better test performance than other features deserve a higher desirability.

*Reset Ant Memory:* All the features selected by ants, their pheromone concentrations as well as their desirability are released for the next tour.

This procedure can be finished with $\mathcal{O}(ikn)$ time complexity and $\mathcal{O}(kn^2)$ space complexity, where $i$ is the iteration, $k$ the number of ants in the colony and $n$ the number of features. It can be noticed that, in this study the ACO approach includes, in addition to the pheromone level $\tau$ and the heuristic information $\eta$ other hyper-parameters, namely, $\mu, \alpha, \beta, \rho, e, \phi, q_0$. While [27] discusses how the pheromone level and the heuristic information can be determined, so far, there have been no studies reporting how the additional parameters are set. Therefore, the optimal settings for these parameters have to be determined empirically.

### E. Training and Classification

The final step is to train a first classifier using the seed texts and to try to assign categories to the other texts. This step is repeated until no improvement of the classifier can be achieved or no remaining texts are left.

### F. System Complexity

In the following section, the system complexity shall be briefly described on the basis of time and space.

***Time Complexity**:* The time complexity of the system, without considering the training and classification process, can be described as shown in Equation (7),

$$\begin{aligned} T(n) &= T_{pre} + T_{gazetteer} + T_{seeds} + T_{ner} + T_{fea_{sel}} \\ &= \mathcal{O}(2^p) + 2\mathcal{O}(1) + 3\Theta(lp) + \mathcal{O}(L|S|^3) + \mathcal{O}(ikn) \end{aligned} \quad (7)$$

where $L$ is the number of samples and $|S|$ the number of labels in the NER process. Furthermore, $l$ is the length of the string, $p$ the length of the search pattern in the string, $i$ the iteration in feature selection, $k$ the number of ants in the ant colony, and $n$ the number of features.

***Space Complexity**:* Similarly, the space complexity can be measured without considering the training and classification process as shown in Equation (8),

$$\begin{aligned} T(n) &= T_{pre} + T_{gazetteer} + T_{seeds} + T_{ner} + T_{fea_{sel}} \\ &= \mathcal{O}(2^p) + \mathcal{O}(g) + \Theta(lp) + \mathcal{O}(s+l) + 2\Theta(lp) \\ &\quad + \mathcal{O}(r) + \mathcal{O}(kn^2) \end{aligned} \quad (8)$$

where $g$ is the size of the gazetteer, $r$ the number of roles, $s$ the size of the trained NER-model, $k$ the number of ants in the ant colony and $n$ the number of features. After analyzing the time and space complexity, it can be shown that the system requires intensive resources in preprocessing and in identifying named entities with respect to time and space complexity. Thus, the performance of our system, regarding time and space complexity, depends on the methods that are used in these two setups.

## V. Experimental Evaluation

To create first baseline results, the labeled data (see Section III-C) were categorized using the following supervised machine learning methods: Naïve Bayes, SVM, and MLP and two ensemble approaches: AdaBoost and RandomForest [34]. The categorization was done once with each dataset. Firstly,

the dataset with 886 texts was used and categorized as music relevant or not. However, as the ultimate goal is a system for the detection of music events and not just music, secondly, the dataset with only 505 texts was categorized as music event relevant or not.

The performance of each experimental setting was determined using micro-averaged precision (Micro P.) and recall (Micro R.) as well as the F1 measure.

### A. Setup

In this study, only two sources of texts concerning music events are considered: Facebook as well as daily and weekly online newspapers. The raw data were preprocessed as described in Section IV-A. Furthermore, all numbers, for example, telephone numbers and dates, were removed and, therefore, not considered in the categorization. For comparison, two different datasets for each dataset were created. The first dataset contains word tokens that were processed with the Porter stemmer [35], whereas for the second dataset the algorithm proposed in [36] was used. For the detection of named entities a Conditional Random Field approach was applied, as proposed by [37]. As was mentioned in Section IV-D, MusicBrainz und DBpedia were used to assign roles to named entities and were combined in order to increase the number of matches.

In this study, the following four representations of the texts incorporating different features were compared:

- multinomial Bag of Words (BoW),
- Term Frequeny Inverse Document Frequency representation of BoW (TF-IDF(BoW)),
- multinomial BoW and music event related Named Entities (BoW+NE), and
- TF-IDF representation of BoW and NE (TF-IDF(BoW+NE)).

In case of named entities only their type (role) was considered as a feature rather than the entity itself, e. g., song writer or musician were taken as a feature instead of Eric Clapton. Moreover, it was only possible to train the SVM with frequency-based features. In the ensemble approaches, an SVM (using the same setting as reported in [1]) and a decision tree were used as basic classifiers. This is distinguished later by *svm* and *tree*, respectively. As a criterion to measure the quality of split, "Gini" for the Gini impurity and "Entropy" for the information gain were considered, which are labeled with *gini* and *ent*, respectively.

### B. Results

Because the features were investigated by means of ACO by using a 3-fold cross validation due to the computation time, the same experimental settings as in [1] were used. Additionally, for the ACO-based selection of features the following settings for the hyper-parameter were used: $\alpha = 0.5$, $\beta = 0.8$, $\rho = 0.5$, $q_0 = 0.6$, $e = 0.6$, $\phi = 0.5$, $\omega = 0.5$. The original number of features for both datasets used in this study is shown in Table III, while the size of the selected features is shown in Table IV. For initializing the pheromone and heuristic in ACO, a constant value for the first experiments was used and later MI.

TABLE III. NUMBER OF FEATURES OF THE MUSIC AND THE MUSIC EVENT DATASET.

| | # raw | # raw including NEs |
|---|---|---|
| Music Relevance | 14,275 | 10,921 |
| Music-Event Relevance | 12,171 | 9,268 |

TABLE IV. NUMBER OF SELECTED FEATURES OF THE MUSIC AND THE MUSIC EVENT DATASET.

| Dataset | Initialization | # raw | # raw including NEs |
|---|---|---|---|
| Music Relevance | Constant Value | 11,164 | 9,584 |
| | MI | 9,921 | 8,428 |
| Music-Event Relevance | Constant Value | 8,618 | 7,375 |
| | MI | 7,852 | 6,485 |

As can be seen in Table IV, the number of features is considerably reduced in both datasets after using ACO to select the features. The overall results achieved with this reduced feature set when being used in the proposed classification pipeline is shown below:

*Baseline Results:* The baseline results of the music relevance decisions regarding the gold standard dataset described in Section III-C are shown in Table V whereas the results for the categorization of music event relevance are shown in Table VI. The results using stemming were compared with those achieved using lemmatization and it was observed that stemming lead to slightly better results. As can be seen in Table V, the best results for the categorization of music relevance, based on the $F_1$-measure, were achieved using a frequency-based representation of words and named entities (roles) and MLP. Comparable results were achieved by an AdaBoost model based on SVMs as the basic classifier but considering the same feature settings. In comparison, a combination of BoW and named entities (roles) and an MLP model achieved the best results for the categorization of music event relevance. These results are presented in Table VI. Furthermore, it was found that the best performing model and feature combination (MLP and BoW+NE) failed when the features (words) occurred in both, the relevant and non-relevant texts, when the texts were very short or when there were not enough features with a high explanatory power available.

Overall, the classification results of music relevance and music event relevance are clearly improved when named entities are considered as features.

*Considering ACO Feature Selection:* As mentioned in Section IV-D, by using MI to initialize the pheromone and heuristic of the features, 9921 features were found using the 886 raw documents and 8428 features using 886 documents by considering named entities in the documents. This is repeated

TABLE V. RESULTS OF THE 3-FOLD CROSS VALIDATION FOR THE MUSIC DATASET USING STEMMING.

| Model | Feature | Micro P. | Micro R. | $F_1$ |
|---|---|---|---|---|
| Naïve Bayes | BoW | 0.656 | 0.974 | 0.784 |
| | TF-IDF(BoW) | 0.991 | 0.625 | 0.766 |
| | BoW+NE | 0.699 | **0.983** | 0.817 |
| | TF-IDF(BoW+NE) | 0.988 | 0.736 | 0.844 |
| MLP | BoW | 0.886 | 0.874 | 0.880 |
| | TF-IDF(BoW) | 0.884 | 0.897 | 0.890 |
| | Bow+NE | 0.943 | 0.903 | 0.922 |
| | TF-IDF(BoW+NE) | 0.934 | 0.931 | **0.933** |
| SVM | TF-IDF(BoW) | 0.961 | 0.840 | 0.896 |
| | TF-IDF(BoW+NE) | 0.975 | 0.894 | 0.933 |
| AdaBoost | *svm*_TF-IDF(BoW) | 0.961 | 0.840 | 0.896 |
| | *svm*_TF-IDF(BoW+NE) | 0.975 | 0.894 | **0.933** |
| | *tree*_TF-IDF(BoW) | 0.982 | 0.633 | 0.770 |
| | *svm*_TF-IDF(BoW+NE) | **0.996** | 0.719 | 0.835 |
| RandomForest | *gini*_TF-IDF(BoW) | 0.985 | 0.771 | 0.865 |
| | *gini*_TF-IDF(BoW+NE) | 0.990 | 0.851 | 0.915 |
| | *ent*_TF-IDF(BoW) | 0.981 | 0.751 | 0.851 |
| | *ent*_TF-IDF(BoW+NE) | 0.990 | 0.831 | 0.903 |

TABLE VI. RESULTS OF THE 3-FOLD CROSS VALIDATION FOR THE MUSIC EVENT DATASET USING STEMMING.

| Model | Feature | Micro P. | Micro R. | $F_1$ |
|---|---|---|---|---|
| Naïve Bayes | BoW | 0.892 | 0.943 | 0.917 |
| | TF-IDF(BoW) | 0.852 | 0.943 | 0.895 |
| | BoW+NE | 0.907 | **0.966** | 0.936 |
| | TF-IDF(BoW+NE) | 0.873 | **0.966** | 0.917 |
| MLP | BoW | 0.922 | 0.901 | 0.912 |
| | TF-IDF(BoW) | 0.905 | 0.909 | 0.907 |
| | Bow+NE | 0.953 | 0.924 | **0.938** |
| | TF-IDF(BoW+NE) | 0.932 | 0.932 | 0.932 |
| SVM | TF-IDF(BoW) | 0.927 | 0.920 | 0.924 |
| | TF-IDF(BoW+NE) | 0.946 | 0.928 | 0.937 |
| AdaBoost | *svm*_TF-IDF(BoW) | 0.927 | 0.920 | 0.924 |
| | *svm*_TF-IDF(BoW+NE) | 0.946 | 0.928 | 0.937 |
| | *tree*_TF-IDF(BoW) | 0.965 | 0.734 | 0.834 |
| | *svm*_TF-IDF(BoW+NE) | **0.991** | 0.795 | 0.882 |
| RandomForest | *gini*_TF-IDF(BoW) | 0.956 | 0.821 | 0.883 |
| | *gini*_TF-IDF(BoW+NE) | 0.979 | 0.882 | 0.928 |
| | *ent*_TF-IDF(BoW) | 0.948 | 0.829 | 0.884 |
| | *ent*_TF-IDF(BoW+NE) | 0.978 | 0.859 | 0.915 |

by constantly initializing the pheromone and heuristic of the features. Table V shows that with dramatically reduced features, the best $F_1$-measure for the music relevance dataset is improved from 0.933 (in i.e., MLP and TF-IDF (BoW+NE)) to 0.950 (i.e., SVM and TF-IDF (BoW+NE)) when initializing the pheromone and heuristic with MI. Furthermore, all the results are clearly improved by using MI in the initialization, as shown in Table VII. In comparison, the best results do not change dramatically when initializing ACO in a constant way, as can be seen in Table VIII. However, the feature size is clearly reduced in comparison to the original feature size in Table III.

The results in Tables VII and IX show that with the ACO feature selection the performance of all machine learning algorithms are improved: using MI for initializing the pheromone and heuristic, the best baseline result for the music relevance decision is improved from 0.933 to 0.950, and for the music event relevance decision from 0.938 to 0.968.

By analyzing the results of detecting music-event relevance by initializing colony information with MI, the following conclusions were drawn:

a) The classification result can be clearly improved by reducing the feature dimension: the average number of tokens that occur in the documents that are correctly classified after feature selection changes from 73 to 46 and the average classification accuracy is improved from 0.938 to 0.954.
b) The selected features are informative for the classification as shown by the results in Table IX.
c) The false negative classified documents are still classified incorrectly after the selection of features.

It seems that the last mentioned conclusion depends on the distribution of the features in the documents. It was noticed that some music related features lead to unexpected results during the classification. For example, considering the following document:

> "Marienmünster am Freitag. Abtei Marienmünster, 15 Uhr weihnachtliche Orgelmusik, Arien und Instrumentalstücke zum Fest.",

words like *Orgelmusik* (organ music) and *Instrumentalstücke* (instrumentals) give a strong feeling that this document is somehow related to a music event in public. However, as the features only occur once in the corpus, they fail to be considered as strong features by the model to make a decision.

TABLE VII. RESULTS OF THE 3-FOLD CROSS VALIDATION FOR THE MUSIC DATASET USING STEMMING AND ACO FEATURE SELECTION WITH MI INITIALIZED PHEROMONE AND HEURISTIC.

| Model | Feature | Micro P. | Micro R. | $F_1$ |
|---|---|---|---|---|
| Naïve Bayes | BoW | 0.811 | 0.983 | 0.889 |
| | TF-IDF(BoW) | **0.990** | 0.817 | 0.895 |
| | BoW+NE | 0.830 | **0.994** | 0.905 |
| | TF-IDF(BoW+NE) | 0.981 | 0.885 | 0.931 |
| MLP | BoW | 0.948 | 0.894 | 0.920 |
| | TF-IDF(BoW) | 0.915 | 0.926 | 0.920 |
| | BoW+NE | 0.972 | 0.911 | 0.941 |
| | TF-IDF(BoW+NE) | 0.935 | 0.951 | 0.943 |
| SVM | TF-IDF(BoW) | 0.969 | 0.897 | 0.932 |
| | TF-IDF(BoW+NE) | 0.967 | 0.934 | **0.950** |
| AdaBoost | *svm*_TF-IDF(BoW) | 0.969 | 0.897 | 0.932 |
| | *svm*_TF-IDF(BoW+NE) | 0.967 | 0.934 | **0.950** |
| | *tree*_TF-IDF(BoW) | 0.987 | 0.645 | 0.780 |
| | *svm*_TF-IDF(BoW+NE) | **0.996** | 0.722 | 0.837 |
| RandomForest | *gini*_TF-IDF(BoW) | 0.993 | 0.777 | 0.871 |
| | *gini*_TF-IDF(BoW+NE) | 0.991 | 0.860 | 0.923 |
| | *ent*_TF-IDF(BoW) | 0.983 | 0.762 | 0.859 |
| | *ent*_TF-IDF(BoW+NE) | 0.990 | 0.862 | 0.922 |

## VI. CONCLUSION AND FUTURE WORK

In this paper, two gold standard datasets for music event detection were presented and made publicly available here

TABLE VIII. RESULTS OF THE 3-FOLD CROSS VALIDATION FOR THE MUSIC DATASET USING STEMMING AND ACO FEATURE SELECTION, INITIALIZED WITH A CONSTANT VALUE FOR PHEROMONE AND HEURISTIC.

| Model | Feature | Micro P. | Micro R. | $F_1$ |
|---|---|---|---|---|
| Naïve Bayes | BoW | 0.667 | 0.974 | 0.792 |
| | TF-IDF(BoW) | 0.979 | 0.673 | 0.798 |
| | BoW+NE | 0.684 | **0.986** | 0.808 |
| | TF-IDF(Bow+NE) | **0.989** | 0.765 | 0.863 |
| MLP | BoW | 0.915 | 0.862 | 0.888 |
| | TF-IDF(BoW) | 0.904 | 0.891 | 0.899 |
| | Bow+NE | 0.946 | 0.903 | 0.924 |
| | TF-IDF(Bow+NE) | 0.921 | 0.940 | **0.930** |
| SVM | TF-IDF(BoW) | 0.964 | 0.840 | 0.897 |
| | TF-IDF(Bow+NE) | 0.981 | 0.885 | 0.931 |

TABLE IX. RESULTS OF THE 3-FOLD CROSS VALIDATION FOR THE MUSIC EVENT DATASET USING STEMMING AND ACO FEATURE SELECTION WITH MI INITIALIZED PHEROMONE AND HEURISTIC.

| Model | Feature | Micro P. | Micro R. | $F_1$ |
|---|---|---|---|---|
| Naïve Bayes | BoW | 0.945 | 0.985 | 0.965 |
| | TF-IDF(BoW) | 0.938 | 0.981 | 0.959 |
| | BoW+NE | 0.949 | 0.985 | 0.966 |
| | TF-IDF(Bow+NE) | 0.932 | **0.989** | 0.959 |
| MLP | BoW | 0.956 | 0.909 | 0.932 |
| | TF-IDF(BoW) | 0.957 | 0.939 | 0.948 |
| | Bow+NE | 0.965 | 0.943 | 0.954 |
| | TF-IDF(Bow+NE) | 0.954 | 0.951 | 0.952 |
| SVM | TF-IDF(BoW) | 0.966 | 0.970 | **0.968** |
| | TF-IDF(Bow+NE) | 0.966 | 0.958 | 0.962 |
| AdaBoost | *svm*_TF-IDF(BoW) | 0.966 | 0.970 | **0.968** |
| | *svm*_TF-IDF(Bow+NE) | 0.966 | 0.958 | 0.962 |
| | *tree*_TF-IDF(BoW) | 0.973 | 0.696 | 0.812 |
| | *svm*_TF-IDF(Bow+NE) | **0.986** | 0.795 | 0.880 |
| RandomForest | *gini*_TF-IDF(BoW) | 0.965 | 0.844 | 0.901 |
| | *gini*_TF-IDF(Bow+NE) | 0.975 | 0.882 | 0.926 |
| | *ent*_TF-IDF(BoW) | 0.978 | 0.837 | 0.902 |
| | *ent*_TF-IDF(Bow+NE) | 0.975 | 0.894 | 0.933 |

TABLE X. RESULTS OF THE 3-FOLD CROSS VALIDATION FOR THE MUSIC EVENT DATASET USING STEMMING AND ACO FEATURE SELECTION, INITIALIZED WITH A CONSTANT VALUE FOR PHEROMONE AND HEURISTIC.

| Model | Feature | Micro P. | Micro R. | $F_1$ |
|---|---|---|---|---|
| Naïve Bayes | BoW | 0.882 | 0.935 | 0.908 |
| | TF-IDF(BoW) | 0.865 | 0.947 | 0.904 |
| | BoW+NE | 0.904 | 0.966 | 0.934 |
| | TF-IDF(Bow+NE) | 0.874 | **0.973** | 0.921 |
| MLP | BoW | 0.912 | 0.905 | 0.908 |
| | TF-IDF(BoW) | 0.882 | 0.935 | 0.908 |
| | Bow+NE | 0.961 | 0.932 | 0.946 |
| | TF-IDF(Bow+NE) | 0.940 | 0.947 | 0.943 |
| SVM | TF-IDF(BoW) | 0.912 | 0.901 | 0.906 |
| | TF-IDF(Bow+NE) | **0.976** | 0.935 | **0.955** |

[38]. Furthermore, a process chain for the categorization of music event related texts was proposed and a first baseline evaluation conducted. For finding much more representative features for event detection, a couple of seed words are given for learning extended features correspondingly. The results change slightly in comparison to results by using the same working process. Further experiments with ACO-based feature selection show that a frequency-based approach considering music specific named entities performs best together with an SVM model for the classification of music relevant texts and an SVM-based ensemble model for the classification of music event relevant texts. As a strategy for initializing the relevant parameters for ACO (e.g., heuristic and pheromone), using MI gives promising results. As discussed in Section V-B, the documents, in which the occurrence of non-domain related features is more dominant than the one of domain related features, are intended to be classified as non-music and event relevant, although these documents are truly music and event relevant. For this purpose, the weights of the features should be considered in classification.

Obviously, the proposed event detection approach can also be applied to similar domains such as movie showings, however, it should be analyzed how it may be applicable to other more general types of events such as social events spread via a social network, or economic events in market and event copyright violation issues. In future work, the following aspects need to be given more attention:

*Gold standard datasets:* The datasets used in this study are relatively small, especially the one including music event related texts and need to be extended in the future. Alternatively, classification results may be improved without extended datasets by considering further strategies. For example, transfer learning enables the model to use pre-trained knowledge to transfer it to the original problem domain, where there is no sufficient training data available [39]. Based on neural probabilistic language models [40], the texts can be represented by using pre-trained vectors of words that enable the models to observe the semantic in the sentences. Furthermore, active learning shows the ability to reduce the training samples as reported in [41].

*Named Entities as Features:* In [20], the authors used named entities to represent the documents in a boolean model and a vector space model. Similar work was conducted in [19], where the entity power coefficient is used to create occurrences of all terms that related to a named entity. For the study at hand it would be important to distinguish the different named entities into hierarchical classes, since some music events are related to music that is not protected by any copyright laws. In such cases, there is a need to distinguish these events from those that might include copyright infringements. For this purpose, a more fine-grained categorization to separate different types of events can be realized by applying hierarchical classification methods, such as discussed in [20] [42].

*Hyper-parameter of ACO:* As discussed in Section IV-D, the ACO approach is accompanied with hyper-parameters and different strategies for initializing the pheromone and heuristic values. In this study, it was shown that the best results on

the music relevant dataset is clearly improved by the correct initialization of the pheromone and heuristic values in an ant colony. Furthermore, it would be interesting to analyze, how the other parameter influence the results, e.g. the trade-off between exploitation and exploration under the control of the hyper-parameter, in this study $q_0 = 0.6$; or how the hyper-parameter $\mu$ changes the final results, if each ant has a bigger capacity for the feature subset, in this study $\mu = 0.35$. Some conclusions about how to select hyper-paremeters like $\beta, \rho$ for other applications, e.g., for the well-known traveling salesman problem are drawn by [27]. Additionally, the optimal population size of ants in a colony in the feature selection process was discussed in [30]. How the other parameters are initialized in selecting the features, should also be investigated in the future.

*Knowledge Base:* There are music-related named entities that occur in German texts but there is no corresponding entry in databases like dbpedia or Musicbrainz for those entities. Hence, the availability of music-related entities in such databases plays a crucial role in expanding the knowledge base in order to assign correct music roles. Here, suitable fall-back strategies should be considered in future work.

*Polysemy & Synonymy:* During the experiments it has been noticed that no text representation captures phenomena like polysemy or synonymy of words. In German, the word *spielen* (play) has the following syntactic and semantic environment: "Orgel spielen" (playing organ music) and "gegen eine Mannschaft spielen" (playing against a team). For the former one, it is certain that it is music relevant, the latter implies a sport event. However, without further contextual information, even the first one might not be recognized as music relevant. Similarly, the problem of synonymy deserves much more attention in future work, e.g., "... in der Gosecker Schloss-Krypta klangen wieder kristallene Engel..." (in the Gosecker castle-crypt crystal angels sounded again). In this context, the word *klangen* means to play music. Those words are quite often used in public announcement in Germany. Thus, it could be a meaningful feature to decide whether a text is music relevant or not.

## ACKNOWLEDGMENT

The authors would like to thank the deecoob GmbH for acting as experts during the creation of the gold standard and providing data. The project was funded by the German federal ministry for economics and energy.

## REFERENCES

[1] J. Xi, M. Spranger, H. Siewerts, and D. Labudde, "Towards an automated system for music event detection," in SEMAPRO 2018, The Twelfth International Conference on Advances in Semantic Processing. ThinkMind, 2018, pp. 22–27.

[2] GEMA, "Geschäftsbericht mit Transparenzbericht 2017," https://www.gema.de/fileadmin/user_upload/Gema/geschaeftsberichte/GEMA_Geschaeftsbericht2017.pdf, 2018 [retrieved: September, 2018].

[3] V. Gupta and G. S. Lehal, "A Survey of Text Mining Techniques and Applications," Journal of Emerging Technologies in Web Intelligence, vol. 1, no. 1, 2009, pp. 60–76. [Online]. Available: www.alerts.yahoo.com

[4] M. Spranger, H. Siewerts, J. Hampl, F. Heinke, and D. Labudde, "SoNA: A Knowledge-based Social Network Analysis Framework for Predictive Policing," International Journal On Advances in Intelligent Systems, vol. 10, no. 3 & 4, 2017, pp. 147–156.

[5] K. Giridhar and A. James, "Text Classification and Named Entities for New Event Detection," in Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2004, pp. 297–304.

[6] A. Edouard, "Event detection and analysis on short text messages," Ph.D. dissertation, Université Côte D'Azur, 2017, https://hal.inria.fr/tel-01680769/document [retrieved: September, 2018].

[7] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text Classification Using Machine Learning Techniques," WSEAS Transactions on Computers, vol. 4, 2006, pp. 966–974.

[8] R. Bellman, "The Theory of Dynamic Programming," Bulletin of the American Mathematical Society, vol. 60, no. 6, 1954, pp. 503–515.

[9] Y. Yang, T. Pierce, and J. Carbonell, "A Study of Retrospective and Online Event Detection," in Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1998, pp. 28–36.

[10] Y. Raimond and S. Abdallah, "The Event Ontology," http://motools.sourceforge.net/event/event.html, 2007 [retrieved: July, 2019].

[11] J. Allan, Topic detection and tracking: event-based information organization, J. Allan, Ed., 2002.

[12] M. Cordeiro and J. Gama, "Online social networks event detection: A survey," in Solving Large Scale Learning Tasks. Challenges and Algorithms, S. Michaelis, N. Piatkowski, and M. Stolpe, Eds., vol. 9580. Springer Verlag, 2016, pp. 1–41.

[13] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in Proceedings of the 19th International Conference on World Wide Web. ACM, 2010, pp. 851–860.

[14] E. Lefever and V. Hoste, "A classification-based approach to economic event detection in dutch news text," in Proceedings of the Tenth International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA), 2016, pp. 330–335.

[15] G. Jacobs, E. Lefever, and V. Hoste, "Economic event detection in company-specific news text," in Proceedings of the First Workshop on Economics and Natural Language Processing. Association for Computational Linguistics, 2018, pp. 1–10.

[16] F. Hogenboom, F. Frasincar, U. Kaymak, and F. de Jong, "An overview of event extraction from text," in CEUR Workshop Proceedings, vol. 779, 2011, pp. 48–57.

[17] E. Benson, A. Haghighi, and R. Barzilay, "Event discovery in social media feeds," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 389–398.

[18] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," Linguisticae Investigationes, vol. 30, 2007, pp. 3–26.

[19] M. K. Stefan Andelic, "Text classification based on named entities," 2017, pp. 23–28.

[20] Y. Gui, Z. Gao, R. Li, and X. Yang, "Hierarchical text classification for news articles based-on named entities," Advanced Data Mining and Applications, vol. 7713, 2012, pp. 318–329.

[21] M. Spranger and D. Labudde, "Towards Establishing an Expert System for Forensic Text Analysis," International Journal on Advances in Intelligent Systems, vol. 7, no. 1/2, 2014, pp. 247–256.

[22] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in Proceedings of the Fourteenth International Conference on Machine Learning, 1997, pp. 412–420.

[23] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," Data classification: Algorithms and applications, 2014, p. 37.

[24] H. Liu and H. Motoda. Boca Raton, Florida: Chapman and Hall/CRC, Oktober 2007, ch. 2,13.

[25] M. Dorigo, "Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem," IEEE Transactions on Evolutionary Computation, vol. 1, April 1997, pp. 53–66.

[26] D. J. T. Sumpter, Collective Animal Behavior. Princeton (New Jersey): Princeton University Press, 2010, ch. 1,5.

[27] M. Dorigo and T. Stützle, Ant Colony Optimization. Cambridge, Massachusetts. London, England: The MIT Press, 2004, ch. 1,3.

[28] R. K. Sivagaminathan and S. Ramakrishnan, "A hybrid approach for feature subset selection using neural networks and ant colony optimization," Expert Syst. Appl., vol. 33, 2007, pp. 49–60.

[29] D. Martens, M. De Backer, R. Haesen, J. Vanthienen, M. Snoeck, and B. Baesens, "Classification With Ant Colony Optimization," IEEE Transactions on Evolutionary Computation, vol. 11, no. 5, 2007, pp. 651–665.

[30] M. H. Aghdam, N. Ghasem-Aghaee, and M. E. Basiri, "Text feature selection using ant colony optimization," Expert Systems with Applications, vol. 36, no. 3, April 2009, pp. 6843–6853.

[31] E. Saraç and S. A. Özel, "An Ant Colony Optimization Based Feature Selection for Web Page Classification," The Scientific World Journal, vol. 2014, 2014.

[32] M. M. Kabir, M. Shahjahan, and K. Murase, "An Efficient Feature Selection Using Ant Colony Optimization Algorithm," Neural Information Processing, vol. 36, 2009, pp. 242–252.

[33] M. M. Kabir and K. Murase, "A Wew Wrapper Feature Selection Approach Ssing Neural Network," Neurocomputing, vol. 36, 2010, pp. 3273–3283.

[34] F. Pedregosa et al., "Scikit-learn: Machine learning in python," J. Mach. Learn. Res., vol. 12, Nov. 2011, pp. 2825–2830. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2078195

[35] M. F. Porter, "An algorithm for suffix stripping," Program, vol. 14, no. 3, 1980, pp. 130–137.

[36] H. Schmid and H. Schmid, "Improvements in part-of-speech tagging with an application to german," Proceedings of the ACL SIGDAT-Workshop, 1995, pp. 47–50.

[37] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. ACM, 2005, pp. 363–370.

[38] J. Xi, "Music Classification Gold Standard Datasets," https://github.com/fosil-mw/music_classification_data, 2018 [retrieved: July, 2019].

[39] Q. Yang and S. J. Pan, "A Survey on Transfer Learning," IEEE Transactions on Knowledge Data Engineering, vol. 22, 2009, pp. 1345–1359.

[40] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," Journal of Machine Learning Research, vol. 3, 2003, pp. 1137–1155.

[41] D. K. Xiao Li and C. X. Ling, "Active learning for hierarchical text classification," Tan PN., Chawla S., Ho C.K., Bailey J. (eds) Advances in Knowledge Discovery and Data Mining, vol. 7301, 2012, pp. 14–25.

[42] A. McCallum, R. Rosenfeld, T. M. Mitchell, and A. Y. Ng, "Improving text classification by shrinkage in a hierarchy of classes," in Proceedings of the Fifteenth International Conference on Machine Learning, 1997, pp. 359–367.