Regressing Controversy of Music Artists from Microblogs

Mhd Mousa HAMAD
Johannes Kepler University
Linz, Austria
mhd.mousa.hamad@gmail.com

Marcin Skowron
OFAI
Vienna, Austria
marcin.skowron@ofai.at

Markus Schedl
Johannes Kepler University
Linz, Austria
markus.schedl@jku.at

***Abstract*—Social media represents a valuable data source for researchers to analyze how people feel about a variety of topics, from politics to products to entertainment. This paper addresses the *detection of controversies involving music artists*, based on microblogs. In particular, we develop a *new controversy detection dataset* consisting of 53,441 tweets related to 95 music artists, and we devise and evaluate a *comprehensive set of user- and content-based feature candidates to regress controversy*. The evaluation results show a strong performance of the presented approach in the controversy detection task: F1 score of 0.811 in a classification task and RMSE of 0.688 in a regression task, using controversy scores in the range [1, 4].**

In addition, the results obtained in applying the presented approach on a dataset from a different domain (CNN news controversy) demonstrate transferability of the developed feature set, with a significant improvement over prior approaches. A combination of the adopted Gradient Boosting based classifier and the developed feature set results in an F1 score of 0.775, which represents an improvement of 9.8% **compared to the best prior result on this dataset.**

***Index Terms*—controversy detection, sentiment analysis, Twitter, dataset, music**

I. Introduction

Social media users express their feelings about various entities (e.g., persons, products, movies, etc.) in the form of user-generated content. This data provides a real-time view of opinions, activities, and trends around the world. As a result, companies and organizations are analyzing social media to better understand their customers and the public at large, to gain business values or provide better services.

Users with different backgrounds, preferences, values, and beliefs are participating in the massive open collaboration ongoing in social media. The exchange of opinions with opposing of multifaceted views between the users often leads to prolonged public disagreements, an extended discussion marked by the expression of opposing views - a controversy.

Different institutions, companies and research groups are interested in identifying issues that focus public attention and increase users' engagement, and in an early prediction of events and topics that generate such controversies. In the music domain, detecting controversies involving an artist, an album, a song, or a music event in the early stages is of particular importance for music producers, artists, public relations, and marketing departments, providing them the ability for a suitable and timely response. Similarly, controversy detection is also a useful feature in music-related systems, such as automatic recommendation systems [1], which could tailor their recommendations of controversial artists or songs to the listener's preferences (e.g., do not recommend a controversial song to specific listeners even if it would match their music taste). Furthermore, controversy detection from social media extends the scope of the large scale studies on artists' popularity, e.g., recognition, sales of albums, performance of tracks on charts, and their relation to controversies involving the artists in social media.

In this paper, we build and thoroughly evaluate multiple prediction models to detect controversies involving music artists in Twitter, using a set of 41 features, partly adopted from previous work, partly newly devised. We consider an artist controversial if we find tweets with highly differing points of view (admiration, dislike, criticism, etc.) related to this artist. Our experiments were conducted on a newly created dataset for controversy detection, which constitutes another original contribution.

II. Related Work

A controversy is defined as an "argument that involves many people who strongly disagree about something".[1] Controversies can occur in one text written by multiple authors concerning an entity or in many texts from different authors. Controversy detection is the process of automatically analyzing the text(s) about the entity under investigation to detect whether it is/they are controversial. Controversy detection approaches may be categorized into content-based and feature-based approaches.

Content-based controversy detection approaches analyze the content of a single text and/or its edit history to predict whether it is related to a controversial entity. For instance, in [2], the authors propose a basic and two controversy rank (CR) models to identify controversial articles in Wikipedia. These models draw clues from the collaboration between the contributors and the edit history of an article. In [3]–[5], the authors map a webpage to multiple Wikipedia articles by searching for a set of representative terms of that webpage in a large set of Wikipedia articles using the Blekko[2] search

[1] https://www.merriam-webster.com/dictionary/controversy

[2] https://www.blekko.com

engine, then selecting the top k search results which are also considered as neighboring articles for that webpage. They propose approaches to detect if a webpage is controversial or not using its neighboring articles whose controversy scores are aggregated to calculate the controversy score of the webpage. The representative terms used to search for articles is either the top ten most frequent non-stop terms [4] or the topics and subtopics in the webpage [5].

Feature-based controversy detection approaches use a set of features and metadata related to one or multiple texts involving an entity to build a model capable to predict if that entity is controversial. In [6], the authors propose an approach to detect controversial issues in news articles based on the magnitude and the difference of the scores of the sentiments expressed within the terms of these issues. In [7], the authors compute the contradiction score involving topics using a set of texts based on three values: the mean and the variance of the scores of sentiments expressed within the texts and the number of these texts. These approaches, among others [8], [9] rely solely on sentiment analysis, which makes them closer to *polarity detection* approaches than the broader *controversy detection* approaches. While the former set of approaches rely solely on sentiment analysis to look for exactly two points of view (negative and positive), controversy detection approaches use sentiment analysis as one important feature among other features to detect and extract multiple points of view [10]–[13]. Other important features/techniques for detecting controversies include topic analysis and detection since topic changes may reflect a new point of view. For instance, in [11], the authors formulate the task of detecting controversies involving celebrities in Twitter. They define a *snapshot* denoting a triple $s = (e, \Delta t, tweets)$, where e represents an entity (celebrity), Δt represents a period and $tweets$ represents the set of tweets published during the target period involving the target entity. The controversy score for a snapshot is then computed based on the disagreement in sentiments expressed in the tweets and the occurrences of controversial terms, extracted from Wikipedia's *list of controversial issues*.[3] The proposed approach used the *Subjectivity Lexicon* (cf. Table I) to analyze the sentiment of the tweets. In [10], the authors extend the work proposed in [11] and reformulate the task of detecting controversies to differentiate between *event snapshots* and *non-event snapshots*. Three regression models using Gradient Boosted Decision Trees (GBDT) [14] are used with a rich set of features, including but not limited to sentiment-based features, representing each snapshot. In [12], [13], the authors propose approaches to detect controversies in news articles and pages submitted to *Reddit*, a social navigation site, respectively. They use a rich set of features extracted from the comments on the articles and pages. The feature set is derived from the one used in [10]. The proposed approach in [12] takes the time of comments into consideration to study how early it is possible to detect controversial news articles after publishing them. In [15] propose a graph-based approach to detect and quantify controversies involving any topic using a *conversation graph* built based on the activities (e.g., comments, mentions) occurring among a set of people. This graph is split into two partitions, using a partitioning algorithm based on serial programming, from the METIS[4] toolbox [16]. The controversy level involving the topic is finally measured by how separated the two partitions were. In [17], the authors propose an approach to detect controversies in Twitter by exploiting motifs from two interaction-based graphs: *user* and *reply*. The user graph models the relationships between users whereas the reply graph models the activity among them (e.g., discussions and comments). Multiple prediction models, including AdaBoost and SVM, are applied using features extracted from both graphs. In addition to the motifs-based features, the proposed approach uses structure-, propagation- and temporal-based features.

As we mentioned above sentiment-based features play an important role in detecting controversies. These features are extracted using *sentiment analysis* techniques. Sentiment analysis is the process of analyzing a text (or multimedia item) to identify or quantify the emotional state expressed in it. Its approaches may be categorized into lexicon-based and machine learning approaches. *Lexicon-based approaches* rely on an underlying sentiment lexicon which is a list of lexical features (e.g. words) attached with values declaring how negative or positive they are based on their semantic meaning. The lexicons can be categorized based on the labels assigned to their words into polarity-based (with binary label values) or valence-based (with numeric label values representing sentiment intensities) lexicons. Table I shows a summary of widely-used lexicons, most of which were surveyed in [18]. Achieving state-of-the-art performance, the VADER lexicon [18] is used in our proposed method for the analysis of sentiment expressed in tweets, compared to different lexicons used in all the aforementioned controversy detection approaches. *Machine learning approaches* for sentiment analysis include the use of Naïve Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM), which are among the most common classifiers used in text analysis and mining. Variants of Neural Networks (NN) and Deep Learning models have also gained a lot of attention recently in the domain of text analysis, as these models have outperformed other models in multiple tasks [19]–[22].

To best of our knowledge, feature-based controversy detection approaches relied solely on features extracted or related to the texts under investigation but not the authors of these texts who, as suggested in [2], may influence the controversy level involving an entity. In our work, we formulate the task of controversy detection in way similar to [11] but without considering time. Given an entity (music artist) and a set of tweets involving this entity, we regress controversy using a fine-grained controversy taxonomy and focus on the design and evaluation of a comprehensive set of user- and content-based feature candidates. Unlike most other works, we also

[3] https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

[4] http://glaros.dtc.umn.edu/gkhome/metis/metis

TABLE I
SENTIMENT ANALYSIS LEXICONS SUMMARY WITH FEATURES RELATED TO THE DOMAIN OF SOCIAL MEDIA.

Lexicon	**Reference**	**No. Tokens**	**Annotation**	**Acronyms**	**Slangs**	**Emoticons**
Harvard General Inquirer (GI)	[23]	2406	Binary			
Linguistic Inquiry and Word Counts (LIWC)	Commercial http://liwc.wpengine.com/	905	Binary			
Opinion Lexicon	[24]	6800	Binary		TRUE	
Subjectivity Lexicon (MPQA)	[8]	8200+	Binary + Neutral			
Affective Norms for English Words (ANEW)	[25]	1040	Numeric [1, 9]			
SentiWordNet	[26]	147306 (Synsets)	Numeric [0.0, 1.0]			
SentiStrength	[27]	2310	Numeric [-5, -1], [1, 5]			TRUE
AFINN	[28]	2477	Numeric [-5, +5]	TRUE	TRUE	
SenticNet	[29]	50000 (Concepts)	Numeric [-1.0, +1.0]			
Valence Aware Dictionary for sEntiment Reasoning (VADER)	[18]	7500	Numeric [-4, +4]	TRUE	TRUE	TRUE
Emoji Sentiment Ranking	[30]	751 (Emojis)	Binary + Neutral			TRUE

make the created ground truth dataset available for research purposes.

III. DATA COLLECTION AND PROCESSING

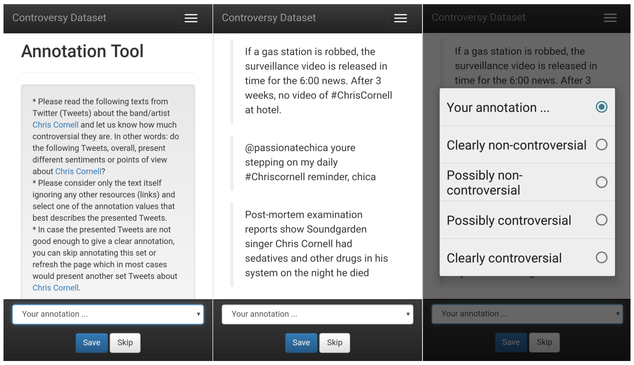
To the best of our knowledge, the only controversy detection dataset available in the domain of microblogs is the one used in [10], but this dataset is not publicly available, neither for research nor for commercial use. We therefore created a new dataset to evaluate controversy detection approaches for music artists. The dataset is available for download from http://www.cp.jku.at/datasets/artist_controversy_dataset.zip.

A. Data Streaming and Processing

We used the *Filter realtime Tweets*[5] (known previously as *public streams*) of Twitter Developers APIs[6] to collect data about music artists. The APIs give developers a low latency access to the public tweet data flowing through Twitter. *Filter realtime Tweets* can be filtered using a list of *track terms*. We compiled a list of about 300 artist names retrieved from different sources (*Last.fm*, *Spotify*, *Billboard*, and *The Top Tens*) to filter English tweets streamed for 13 days (Dec 31, 2016, Feb 25, 2017, May 20, 2017 and Jun 03, 2017 to Jun 13, 2017), to cover different times. We used the names of the most popular artists to be able to get a sufficient amount of tweets during the streaming days.

We applied carefully selected filters on the streamed tweets. The filters were developed after examining the tweets involving the artists by 8 annotators who provided feedback about possible problems discovered in the initially acquired test dataset. In the acquisition of the final dataset, we applied the following processes to address the most common issues identified at the preliminary stage:

- *Filter out ambiguous artist names:* many names referred to other, non-artist, entities (e.g., Air, Berlin, FloRida, Pitbull, etc.). We manually reviewed all the artist names that were skipped during the annotation process (explained below) and removed all the ambiguous ones and their tweets.



Annotation Tool

* Please read the following texts from Twitter (Tweets) about the band/artist Chris Cornell and let us know how much controversial they are. In other words: do the following Tweets, overall, present different sentiments or points of view about Chris Cornell?
* Please consider only the text itself ignoring any other resources (links) and select one of the annotation values that best describes the presented Tweets.
* In case the presented Tweets are not good enough to give a clear annotation, you can skip annotating this set or refresh the page which in most cases would present another set Tweets about Chris Cornell.

If a gas station is robbed, the surveillance video is released in time for the 6:00 news. After 3 weeks, no video of #ChrisCornell at hotel.

@passionatechica youre stepping on my daily #Chriscornell reminder, chica

Post-mortem examination reports show Soundgarden singer Chris Cornell had sedatives and other drugs in his system on the night he died

If a gas station is robbed, the surveillance video is released in time for the 6:00 news. After 3

Your annotation ... ●
Clearly non-controversial ○
Possibly non-controversial ○
Possibly controversial ○
Clearly controversial ○

Fig. 1. Multiple screen shots of the annotation tool.

- *Filter out duplicate tweets:* during the streaming process, we used the original tweet when the streamed tweet was a retweet and the streamed and the original tweets combined when the streamed tweet was quoting another one.
- *Filter out advertising tweets:* we calculated how many times a word (not a stop-word or a special symbol) occurred in all the tweets streamed during the first three streaming days (about 2 million tweets). Then, we manually compiled a list of 44 unwanted words used for advertising purposes (e.g., vote, ticket, tix, etc.). Finally, we removed all tweets containing any of these words or a hyperlink.

B. Data Annotation

We built a simple web-based annotation tool (cf. Figure 1), where only authorized annotators could help estimating how controversial sets of tweets involving music artists are. The tool was developed using Python and Django, a free and open source Python Web framework, and hosted on Amazon Web Services (AWS).

We defined an *annotation set* as an artist with a set of only 50 tweets selected randomly from all the tweets related to the artist. Each annotation set can be annotated by selecting one of four presented options: *"clearly non-controversial"*, *"pos-*

[5] https://developer.twitter.com/en/docs/tweets/filter-realtime/overview
[6] https://developer.twitter.com/en/docs

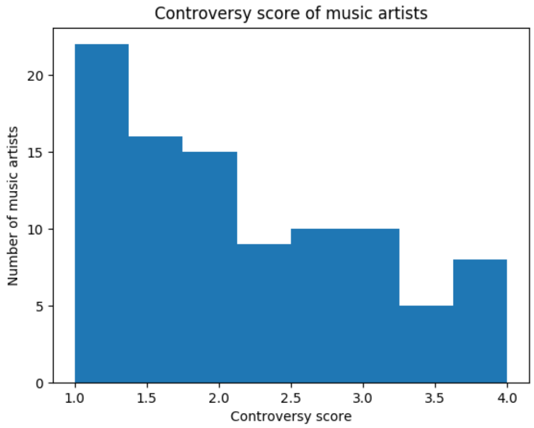


Fig. 2. Controversy scores of music artists.

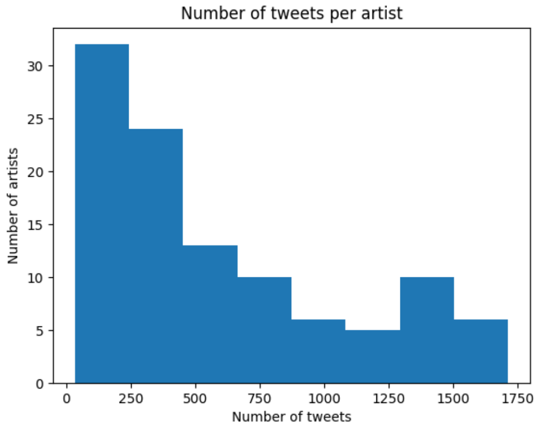


Fig. 3. Number of tweets involving music artists.

sibly non-controversial, *”possibly controversial”*, and *”clearly controversial”*. Using these options is based on [3] where the authors manually annotated Wikipedia articles to detect controversial webpages.

We explained the annotation options using examples to 8 annotators, 2 females and 6 males, with an excellent English command and mean age of 27 with a standard deviation of 2.7 years. The annotators were asked to consider an annotation set as *”clearly non-controversial”* if a single point of view dominated over 90% of the tweets, *”possibly non-controversial* if at least two points of view were present with one of them being expressed in 10%-15% of the tweets, *”possibly controversial”* if at least two points of view were present with one of them being expressed in 15%-20% of the tweets, and *”clearly controversial”* if at least two points of view were present within a comparable percentage of the tweets. They could skip annotating an annotation set or refresh the page presenting the annotation set which changed the presented set of tweets. We selected the artist to be annotated as the one with the least number of received annotations. This kept a balance between the number of annotations received for each artist who, as a result, received exactly 3 annotations (only 7 artists received 4 annotations). We averaged *weighted kappa* agreement values between each two annotators to get a general inter-rater agreement value of 0.335, which is considered a fair agreement [31].

To compute the controversy score of each artist, we mapped the annotation options into numeric values using a simple convention: ”clearly non-controversial” → 1, ”possibly non-controversial” → 2, ”possibly controversial” → 3, and ”clearly controversial” → 4. The final controversy score associated with each artist was computed by averaging all annotation scores received for this artist. Figure 2 shows a histogram of the controversy scores of all artists in the created dataset.

The created dataset contains features related to 95 artists annotated with a continuous controversy score ranging from 1 for *”clearly non-controversial”* artists to 4 for *”clearly controversial”* artists. The scores were manually computed based on annotating randomly selected (as described above) 53,441 tweets, involving these artists and published by 43,141 Twitter users during 13 days. Figure 3 shows a histogram of the number of tweets involving these artists.

The dataset, including user, tweet and artist features (cf. Section IV-A), is freely available for research purposes.[7] Artist features, together with human annotations of their controversy, are available in separate file representing the actual dataset used in our experiments. We also distributed the features related to users and tweets, along with their Twitter IDs, in two different files for researchers interested in this data. Our experiments can be reproducible starting from these two files.

IV. Experiments and Results

We used the created dataset to evaluate the machine learning approaches in detecting controversies using a large set of features. This section describes the adopted as well as proposed features and the prediction models built using them.

A. Feature Extraction

The evaluated feature set contains 41 features representing each artist, in addition to a numeric target feature representing how controversial this artist is (as determined in the human annotation process described above). These features were determined based on a multilevel feature extraction process representing Twitter users, tweets and finally, based on both feature sets, features representing artists.

Most of the **tweet-based features** are derived from the approaches presented in [10], [12], [13]. We extended this set with features related to users and others extracted based on multiple lists manually compiled from different external resources:

- *Profession List* which contains 112 occupations (e.g., ”Engineer”).[8][9]

[7] http://www.cp.jku.at/datasets/artist_controversy_dataset.zip
[8] https://en.wikipedia.org/wiki/Lists_of_occupations
[9] https://www.123test.com/professions

- *Controversial Word List* which contains 1474 controversial terms (e.g., ”feminism”). The list was compiled based on Wikipedia’s list of controversial issues.[10]
- *Controversial Abbreviation List* which contains 42 controversial abbreviations compiled during processing the controversial word list (e.g., ”World War II” was also compiled as ”WWII”).
- *Positive Word List* which contains 266 positive words (e.g., ”wow”).[11]
- *Negative Word List* which contains 939 negative words (e.g., ”boring”).[12,13,14]
- *Slang Word List* which contains 5379 slang words (e.g., ”afc” for ”away from computer”).[15]
- *Positive and Negative Emoticon List* which contains 47 positive emoticons (e.g., ”☺ ”) and 21 negative ones (e.g., ”☹ ”).[16]

User features were extracted from the information provided by Twitter (verified, description and tweets, likes, followings, followers and lists count). These features were used to compute one new feature representing a Twitter user and referred to as *participation score*. The *participation score* is calculated by averaging the normalized scores of all numeric features. We used the logarithm (log base 10) of the number of tweets, likes, followings, followers and lists to normalize their associated feature values as these values span over a large scale.

User and tweet features were combined and processed to define **artist features**. The following list summarizes these features. In this list, features marked with (**) are new, features marked with (*) are adopted from previous work, and the rest are used as they were originally presented in [10], [12], [13]. For the adopted features, we either changed the extraction model (i.e., we used VADER [18] to extract sentiment-related features as it outperformed other sentiment analysis approaches applied in [10], [12], [13]) or the resources used in the extraction process (i.e., we used the new manually compiled lists of controversial, positive, negative and Slang terms. We had to manually process these lists to clean them up to better fit an automatic text processing system).

- Controversy Features
  - Controversial terms (words or abbreviations) mean and SD*
  - Controversial terms count*
- Sentiment Features
  - Positive and negative tweets percentage*
  - Neutral, positive, negative and compound sentiment mean and SD*
  - Positive and negative words mean*
  - Positive and negative emoticons** mean
- User-based Features
  - User participation mean and SD**
  - Verified users percentage**
  - Username includes artist name percentage**
  - Tweets by verified users percentage**
  - Tweets by active users percentage**
- Syntactic Features
  - Tokens mean
  - Nouns, verbs, adjectives** and adverbs** mean
  - Capitalized words mean**
  - Slang words mean**
  - Expressive punctuation marks mean**
- Twitter-based Features
  - Likes mean and SD
  - Retweets mean and SD
  - User mentions and hashtags mean
  - Reply and quoting tweets percentage
- Other Features
  - Tweets count, involving the artist
  - Unique users count, publishing those tweets
  - Tweets per user mean

TABLE II
SELECTED HIGHLY CORRELATED FEATURES.

1st Feature	2nd Feature	Correlation
Tweets count	Users count	0.970
Controversial terms mean	Controversial terms count	0.922
% Positive tweets	Positive sentiment mean	0.977
% Positive tweets	Compound sentiment mean	0.971
% Negative tweets	Negative sentiment mean	0.979
Positive sentiment mean	Compound sentiment mean	0.801
Likes SD	Retweets SD	0.890

B. Feature Analysis

We analyzed the extracted features for artists using the caret (Classification and Regression Training) package available in R [32]. The main goal of conducting this analysis was to select the most relevant and non-redundant features.

1) Feature Correlation Analysis: We analyzed the correlation between each possible pair of features to detect redundant features. Two features were considered highly correlated, and thus redundant, if they had a correlation score smaller than -0.75 or greater than $+0.75$. Table II shows a selection of these highly correlated features.

2) Feature Importance Analysis: We analyzed the importance of each feature to detect irrelevant features which are usually ranked as the least important ones. Feature importance is usually measured by building many prediction models using different subsets of the feature set and evaluating the accuracy of each model. Features that continuously generate the worst accuracy are considered the least important. We used a linear regression model evaluated using R^2 (also known as coefficient of determination) and a Learning Vector Quantization (LVQ) model, a special Artificial Neural Network (ANN) that applies a winner-takes-it-all Hebbian-learning-based approach [33], evaluated using area under ROC to evaluate feature importance. Both models were evaluated in

[10] https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues
[11] http://www.enchantedlearning.com/wordlist/positivewords.shtml
[12] http://www.enchantedlearning.com/wordlist/negativewords.shtml
[13] https://gist.github.com/jamiew/1112488
[14] http://www.noswearing.com/dictionary
[15] https://www.noslang.com/dictionary
[16] https://github.com/bear/python-twitter

TABLE III
FEATURE IMPORTANCE SCORES FOR THE MOST IMPORTANT FEATURES. FEATURES HIGHLY CORRELATED WITH A MORE IMPORTANT FEATURE ARE NOT GIVEN A NUMBER.

No.	Feature Name	LR Score	LVQ Score	Product Score
#1	Negative words mean	0.704	1.000	0.704
#2	Slang words mean	0.738	0.830	0.612
#3	Positive emoticons mean	0.904	0.673	0.609
#4	Compound sentiment mean	0.501	0.628	0.315
#5	Negative sentiment mean	0.405	0.674	0.273
	Negative sentiment SD *	0.452	0.574	0.260
#6	Compound sentiment SD	1.000	0.252	0.252
#7	User participation mean	0.589	0.421	0.248
	Negative tweets percentage *	0.267	0.733	0.196
#8	Reply tweets percentage	0.472	0.326	0.154
#9	Tweets by active users percentage	0.719	0.201	0.145
#10	Verified users percentage	0.834	0.171	0.143
#11	Controversial terms mean	0.349	0.396	0.138
#12	Verbs mean	0.712	0.175	0.125
	Positive tweets percentage *	0.474	0.259	0.123
	Tweets by verified users percentage *	0.864	0.129	0.112
	Controversial terms count *	0.297	0.372	0.110
#13	Neutral sentiment mean	0.406	0.270	0.110
	Positive sentiment mean *	0.408	0.187	0.076
#14	Tokens mean	0.325	0.198	0.064
#15	Retweets mean	0.509	0.115	0.058

three independent runs using 10-fold cross validation. LVQ is a classification model that can be applied on data with binary or categorical class attribute. We adapted the created dataset to fit to classification models by converting the *controversy score* from a numeric attribute into a binary one indicating whether an artist is controversial or not using a threshold $\alpha = 2.4$ identified experimentally. Table III shows the most important features along with their normalized importance scores using both prediction models. The features are ranked using the multiplication of the two normalized importance scores for each of them. We can notice that the new user-related features extracted in this work, such as *User participation mean* (#7) and *Tweets by active users percentage* (#9), are ranked high which reflects their importance in detecting controversies.

C. *Evaluation of Machine Learning Models*

We evaluated two categories of prediction models based on the features analyzed before and using 10-fold cross validation. These models were evaluated using the most important features starting by the most important one (single feature) and adding one feature at a time until we included all of them. In this process, the features that are highly correlated with one of the already added ones (marked in Table III with *) were not considered. All evaluations were conducted using WEKA (Version 3.8.1) [34] with the default parameters.

Decision Table, *Sequential Minimal Optimization (SMO)* which builds a support vector machine (SVM), *Multilayer Perceptron*, and *Random Forest* support numeric and nominal class attributes, i.e., they can be built as both regression and classification prediction models. We evaluated these models for both categories in addition to *Linear Regression* for the regression category. *ZeroR*, i.e., a majority voter, was used as a baseline.

To evaluate the classification prediction models using the created dataset, we changed the controversy score from numeric into a binary attribute indicating whether an artist is controversial or not. We used a threshold α as a cutoff so that each artist with a controversy score greater than α is considered controversial. We evaluated multiple threshold values within the range $[1.5, 2.5]$ with a step of 0.1. These evaluations are based on the complete feature set and using the default parameters defined in WEKA. A cutoff at $\alpha = 2.4$ was used as this value resulted in the best performance and kept the distribution of the class attribute values in balance. This value is also very close to the mean of the controversy scores of all artists $mean = 2.17$ which also indicates how good the selection is (cf. figure 2).

Root Mean Squared Error (RMSE) was used as our evaluation metric for the regression models. *F1 score* was used as our evaluation metric for the classification models. Figure 4 shows these metrics for all evaluated models using the included number of the most important features as the horizontal axis.

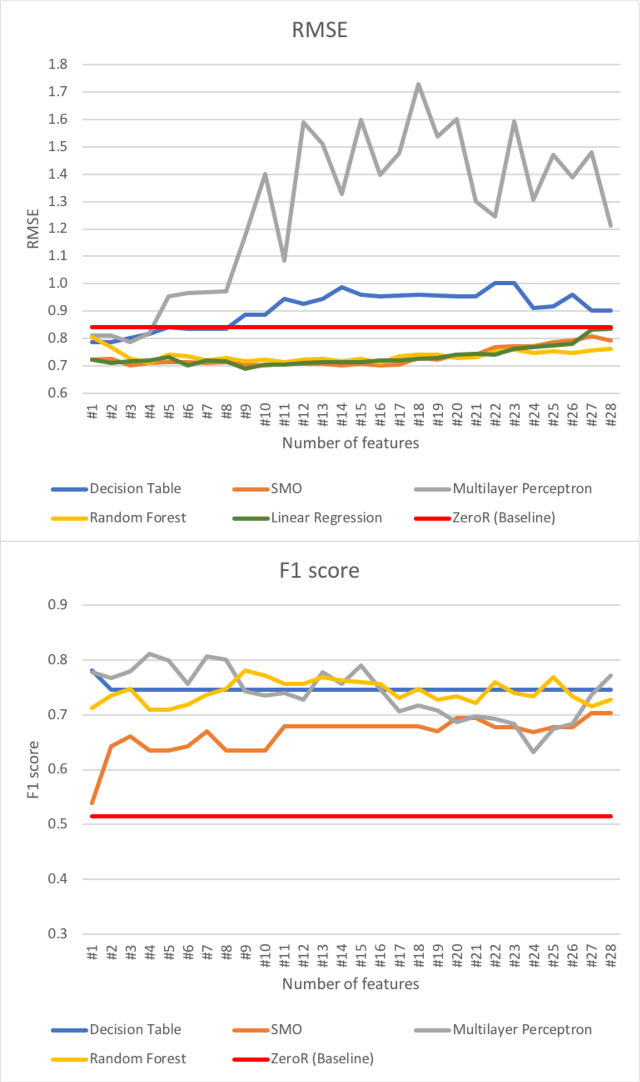


Fig. 4. RMSE and F1 score for the prediction models using the most important feature.

SMO and linear regression performed the best as regression models. Linear regression had a slightly better performance using the 9 most important features $(RMSE = 0.688)$.

Multilayer Perceptron performed the best as a classification model using the 4 most important features ($F1 = 0.811$).

D. Discussion

We evaluated a comprehensive set of features using multiple prediction models. These evaluations showed that using only a subset of the most relevant non-redundant features (between 4 and 9) leads to the best controversy detection system involving music artists in Twitter on the presented dataset. This subset of features does not only improve the performance of the detection model, but also optimizes the resources required to build this model. It also contains some of the new features extracted in this work (cf. Table III).

Since this is, to the best of our knowledge, the first work to address controversy detection of music artists, we cannot directly compare results to existing work. Nevertheless, to assess the ability of generalizing our approach to other data sources and domains, we adopted and applied the complete feature extraction and evaluation processes on a controversy detection dataset of CNN news articles presented in [12]. The dataset contains 728 articles (376 controversial and 352 non-controversial) published by CNN, and 522,595 comments written by 40,826 CNN users. The articles were annotated by multiple annotators as controversial or non-controversial. We extracted two feature sets related to CNN users and comments which correspond to the sets related to Twitter users and tweets, respectively. Some user features, however, are different as they are platform-specific. CNN provides an automatic reputation score for each of its users. We used this score in generating a user participation score combining this feature with other available user features. Some of the Twitter-based features for tweets had direct replacements (i.e., likes, replies) in CNN comments. Features such as retweets, hashtags and user mentions were excluded as no similar or related feature was available in the news dataset. The final set of features representing news articles corresponds thus to the set representing the artists in Twitter, with the exclusion of features related to verified user accounts, retweets, and user mentions. The best performance reported in [12] considered the feature set extracted using all comments on news articles and was obtained with a Decision Table classifier, yielding an F1 score of 0.706. Using the same classifier with its reported parameters (the default parameters in WEKA), the best performance obtained with the 14 most important features representing news articles was an F1 score of 0.722. Using SMO (SVM) as classifier, we achieved an even better performance ($F1 = 0.748$) when using the 18 most important features, compared to the SVM results reported in the original work ($F1 = 0.507$). Further, we evaluated the performance of two additional classifiers which were not applied in [12]: Random Forest and Gradient Boosting. The F1 scores achieved with our approach that combines the new classifiers and the developed set of features are $F1 = 0.769$ and $F1 = 0.775$, for Random Forest and Gradient Boosting, respectively. The overall best performing approach provided 9.8% improvement in F1 score, compared to the best prior result on this dataset.

In general, comparing the most relevant features in the presented dataset with the relevant features reported in [12] shows the specificity of Twitter compared to other media sources. Specifically, negative words, slang words, and emoticons have higher relevance for the controversy score in Twitter compared to other features. Conversely, the controversial terms and sentiments expressed in the comments on CNN news articles have the highest relevance to the controversy score. The relatively short length of tweets may be the cause of this difference as people tend to use more emoticons and slang expressions to shorten the messages as well as a difference in the average age and education level of users in both domains. This corroborates the findings from [35], [36].

V. Conclusions

Controversy detection is important in multiple domains ranging from social studies to marketing. In the particular domain of music, it is highly relevant and useful for music companies, producers, and listeners. It can also improve the performance of music-related systems such as recommender and classification systems. However, controversy detection of music artists has to the best of our knowledge not been researched so far in a social media context. Therefore, we addressed the problem of controversy detection of music artists using data streamed from Twitter. We developed a new ground truth dataset and evaluated a comprehensive set of features (parts of which we newly proposed) using multiple prediction models to analyze their performance for this task.

The evaluation results are promising, achieving a root mean squared error (RMSE) of 0.688 with a linear regression model built from only the 9 most important features, and an F1 score of 0.811 using a Multilayer Perceptron and only the 4 most important features. Our results show that using the right feature set is more important than using a comprehensive one for the task at hand. In addition, the results obtained in experiments conducted on a dataset from a different domain (CNN news controversy) demonstrate transferability of the developed feature set, with a significant improvement over the prior approach using a significantly smaller set of features.

The created dataset involves 95 music artists annotated manually by 8 annotators with a continuous controversy score in the range $[1, 4]$. The admittedly small number of annotated artists and the inter-rater agreement between the annotators (0.335) are two limiting factors of this dataset. In the future, we plan to considerably extend the dataset, involving more artists along with a more fine-tuned processing pipeline to filter out erroneous tweets. This will also allow to devise and apply deep learning models for detecting controversies about artists.

VI. Acknowledgments

This research is partially funded by the Austrian Science Fund (FWF) under grant no. P 27530. We thank the annotators for their work.

References

[1] M. Schedl, P. Knees, B. McFee, D. Bogdanov, and M. Kaminskas, *Recommender Systems Handbook*, 2nd ed. Springer, 2015, ch. Music Recommender Systems.

[2] B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, H. W. Lauw, and K. Chang, "On ranking controversies in wikipedia: Models and evaluation," in *The 2008 International Conference on Web Search and Data Mining (WSDM)*. Palo Alto, CA, USA: ACM, 2008, pp. 171–182.

[3] S. Dori-Hacohen and J. Allan, "Detecting controversy on the web," in *The 22nd ACM International Conference on Information and Knowledge Management (CIKM)*. San Francisco, CA, USA: ACM, 2013, pp. 1845–1848.

[4] ——, "Automated controversy detection on the web," in *Advances in Information Retrieval - The 37th European Conference on IR Research (ECIR)*, vol. 9022. Vienna, Austria: Springer, 2015, pp. 423–434.

[5] M. Jang and J. Allan, "Improving automated controversy detection on the web," in *The 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. Pisa, Italy: ACM, 2016, pp. 865–868.

[6] Y. Choi, Y. Jung, and S.-H. Myaeng, "Identifying controversial issues and their sub-topics in news articles," in *Pacific-Asia Workshop on Intelligence and Security Informatics (PAISI)*. Hyderabad, India: Springer Berlin Heidelberg, 2010, pp. 140–153.

[7] M. Tsytsarau, T. Palpanas, and K. Denecke, "Scalable detection of sentiment-based contradictions," in *The 1st International Workshop on Knowledge Diversity on the Web (DiversiWeb)*, Hyderabad, India, 2011, pp. 9–16.

[8] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *The Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT)*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, 2005, pp. 347–354.

[9] M. Tsytsarau, T. Palpanas, and K. Denecke, "Scalable discovery of contradictions on the web," in *The 19th International Conference on World Wide Web (WWW)*. Raleigh, NC, USA: ACM, 2010, pp. 1195–1196.

[10] A.-M. Popescu and M. Pennacchiotti, "Detecting controversial events from twitter," in *The 19th ACM International Conference on Information and Knowledge Management (CIKM)*. Toronto, ON, Canada: ACM, 2010, pp. 1873–1876.

[11] M. Pennacchiotti and A.-M. Popescu, "Detecting controversies in twitter: A first study," in *The NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media (WSA)*. Los Angeles, CA, USA: Association for Computational Linguistics, 2010, pp. 31–32.

[12] R. V. Chimmalgi, "Controversy trend detection in social media," Engineering Science (Interdepartmental Program), Louisiana State University and Agricultural and Mechanical College, Baton Rouge, LA, USA, Master's Thesis, 2013.

[13] O. P. Anifowose, "Identifying controversial topics in large-scale social media data," Faculty of Media, Bauhaus-Universität Weimar, Weimar, Germany, Master's Thesis, 2016.

[14] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, pp. 1189–1232, 2001.

[15] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, "Quantifying controversy in social media," in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, ser. WSDM '16. New York, NY, USA: ACM, 2016, pp. 33–42. [Online]. Available: http://doi.acm.org/10.1145/2835776.2835792

[16] G. Karypis and V. Kumar, "Metis – unstructured graph partitioning and sparse matrix ordering system, version 2.0," Tech. Rep., 1995.

[17] M. Coletto, K. Garimella, A. Gionis, and C. Lucchese, "A motif-based approach for identifying controversy," *CoRR*, vol. abs/1703.05053, 2017. [Online]. Available: http://arxiv.org/abs/1703.05053

[18] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *The Eighth International AAAI Conference on Weblogs and Social Media*. Ann Arbor, MI, USA: AAAI Publications, 2014.

[19] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Seattle, WA, USA: The Association for Computational Linguistics, 2013, pp. 1631–1642.

[20] L. Dong, F. Wei, C. Tan, D. Tang—, M. Zhou, and K. Xu, "Adaptive recursive neural network for target-dependent twitter sentiment classification," in *The 52nd Annual Meeting of the Association for Computational Linguistics: Short Papers*. Baltimore, MD, USA: Association for Computational Linguistics, 2014, pp. 49–54.

[21] C. N. dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *COLING*. ACL, 2014, pp. 69–78.

[22] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 task 4: Sentiment analysis in Twitter," in *Proceedings of the 11th International Workshop on Semantic Evaluation*, ser. SemEval '17. Vancouver, Canada: Association for Computational Linguistics, August 2017.

[23] P. J. Stone, D. C. Dunphry, M. S. Smith, and D. M. Ogilvie, *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, England: MIT Press, 1966.

[24] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *The ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*. Seattle, WA, USA: ACM, 2004, pp. 168–177.

[25] M. M. Bradley and P. J. Lang, "Affective norms for english words (anew): Instruction manual and affective ratings," The Center for Research in Psychophysiology, University of Florida, Gainesville, FL, USA, Tech. Rep., 1999.

[26] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: MIT Press, 1998.

[27] M. Thelwall, *The Heart and Soul of the Web? Sentiment Strength Detection in the Social Web with SentiStrength*. Springer, 2017, pp. 119–134.

[28] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," in *The ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages*, Heraklion, Greece, 2011, pp. 93–98.

[29] E. Cambria, S. Poria, R. Bajpai, and B. Schuller, "Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives," in *Commonsense Knowledge*. Osaka, Japan: AAAI Publications, 2016, pp. 2666–2677.

[30] P. Kralj Novak, J. Smailović, B. Sluban, and I. Mozetič, "Sentiment of emojis," *PLOS ONE*, vol. 10, no. 12, pp. 1–22, 12 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0144296

[31] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, pp. 213–220, 1968.

[32] M. Kuhn, "Building predictive models in r using the caret package," *Journal of Statistical Software*, vol. 28, pp. 1–26, 2008.

[33] T. Kohonen, *Learning Vector Quantization*. Cambridge, MA, USA: MIT Press, 1998, pp. 537–540.

[34] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix*, 4th ed. Morgan Kaufmann, 2016.

[35] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!" in *ICWSM*, L. A. Adamic, R. A. Baeza-Yates, and S. Counts, Eds. The AAAI Press, 2011.

[36] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PLOS ONE*, vol. 8, no. 9, pp. 1–16, 09 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0073791