An Exploration of Mood Classification in the Million Songs Dataset

Humberto Corona, Michael P. O'Mahony
Insight Centre for Data Analytics
School of Computer Science and Informatics
University College Dublin, Ireland
firstname.lastname@insight-centre.org

ABSTRACT

As the music consumption paradigm moves towards streaming services, users have access to increasingly large catalogs of music. In this scenario, music classification plays an important role in music discovery. It enables, for example, search by genres or automatic playlist creation based on mood. In this work we study the classification of song mood, using features extracted from lyrics alone, based on a vector space model representation. Previous work in this area reached contradictory conclusions based on experiments carried out using different datasets and evaluation methodologies. In contrast, we use a large freely-available dataset to compare the performance of different term-weighting approaches from a classification perspective. The experiments we present show that lyrics can successfully be used to classify music mood, achieving accuracies of up to 70% in some cases. Moreover, contrary to other work, we show that the performance of the different term weighting approaches evaluated is not statistically different using the dataset considered. Finally, we discuss the limitations of the dataset used in this work, and the need for a new benchmark dataset to progress work in this area.

1. INTRODUCTION

Most of the research on music classification is based on features obtained by audio analysis [1–3]. However, previous work by Besson et al. [4] concluded that semantic (lyrics) and harmonic (tunes) information are processed independently by the brain, even when these information sources are closely related to each other. This indicates the relevance of lyrics in music classification, as it can be complementary to the study of harmonic information.

It is noteworthy that the results obtained by different previous work are not consistent in their methodology or outcomes. For example, [5, 6] found that lyrical features can outperform audio features in music mood classification in certain categories. However, [7] suggests that lyrics perform worse than cultural or audio features. Moreover, comparing results from previous work is difficult, as the class labels selected for classification are not consistent, and the datasets used are different and/or are not publicly available.

In this paper we focus on the analysis of lyrics for mood classification. We follow state-of-the-art approaches to infer music mood using social tags, and study three different levels of granularity for mood classification. We study a vector space model (VSM) representation of songs using different term-weighting approaches, with the aim of establishing a comprehensive benchmark for music mood classification using lyrical features. Moreover, we present a feature analysis to further explain the main findings of this work.

As lyrics are copyrighted material, it is difficult to legally obtain a large dataset. Contrary to previous work, which are based on different small-scale datasets [8,9], we use the *Million Song Dataset (MSD)* [10], a large freely-available dataset which facilitates the reproducibility and comparison of the findings presented in this work. With this goal in mind, we also make our source code publicly available [1] .

The remainder of this paper is organised as follows. First, Section 2 describes the related work. Section 3 describes the datasets used and Section 4 presents the term-weighting metrics studied in this work. Section 5 presents a feature analysis of song lyrics. Section 6 introduces the classification approach studied in this paper and the results obtained. Finally, in Section 7, conclusions are presented.

2. RELATED WORK

Most of the existing music classification approaches rely on audio analysis to infer mood or genres [11, 12], while other approaches combine audio analysis with other features, such as cultural features [7] or lyrics [13]. However, the exploration of lyrics alone as a source of information for music classification is an interesting problem and it has not been widely explored. In this section we present an overview of the main approaches for mood representation and music mood classification using lyrics.

2.1 Mood Representation

Music moods are difficult to infer: people perceive them differently [14] and they are culturally dependent [15]. Moreover, some songs (e.g., *Bohemian Rhapsody* by *Queen* [2]) express a wide range of moods over the course of the song.

[1] http://github.com/hcorona/SMC2015
[2] http://open.spotify.com/track/1fNo4jzUtg9EC0yyHcZY5j

Most of the proposed mood ontologies rely on models developed in the psychology field — *Russell's model of affect* [16] being one of the most widely used. This model is based on the evidence that the affective dimensions are built in a *highly systematic fashion*, instead of being independent dimensions. Each mood $m \in M$ is mapped onto a two-dimensional space defined by valence v (which measures the good–bad dimension of sentiment) and arousal a (which measures the active–passive dimension of sentiment). Therefore, in this model each mood can be represented by a vector in the two-dimensional valence-arousal space $m \in M = (v, a)$.

Russell's theoretical model has been adapted to the music classification problem [17] using social tags to infer the categories. The authors also expose the lack of consensus on the names for concepts to be learned; some authors refer to *mood* while others refer to *sentiment* to define the same concept. There is also little consensus in defining the mood categories to classify, which makes comparing research output in this area problematic.

In this work, we use the term *mood* to refer to the categories to be learned in the classification problem. Moreover, we infer mood from social tags as proposed in [17, 18]. Then, we group those moods into four categories, following Russell's model of affect. This allows us to build a large dataset in which the mood groups are clearly defined. Moreover, the approach facilitates different levels of granularity that can be used in the classification task.

2.2 Music Classification using Lyrics

Hu et al. [8] propose a method for detecting mood for 500 manually labeled Chinese songs using lyrics. The approach maps mood into a two-dimensional space of valence and arousal and uses a translated and expanded version of the ANEW (Affective Norms for English Words) dataset [19]. Then a fuzzy clustering method is used to group the lyrics' sentences according to their mood and to extract one prominent mood from each song. The results show that lyric mood are more correlated to valence than to the arousal dimension.

Downie et al. [18] propose a lyric-based approach to mood classification using a binary SVM classifier. The article proposes a vector space model feature set, combined with other statistical textual features such as part-of-speech tags. The results are evaluated using a private dataset with 5,585 songs and using the 18 mood categories presented in [20]. The results show that the combination of both audio and lyrical features can improve classification performance. In later work [21], the authors explore different lyrical features and modifiers, such as stylistic features or features obtained from the ANEW dataset. The results show that a lyrics-based classifier can outperform an audio based classifier for some mood categories.

Kim et al. [22] also explore lyric-based mood classification. The approach uses partial syntactic analysis to extract emotions or mood from songs, achieving an accuracy of around 60% when evaluated in a manually labeled dataset of 500 Korean songs. This paper proposes an approach which includes novel features such as negation detection, time of emotion and change of emotion. A different approach is adopted by Kumar et al. [23], who use SentiWordNet [3] to extract mood features from 185 lyrics labeled with one of four mood categories: *happy, angry, love* and *sad*. This work compared three classifiers: KNN, SVM, and Naïve Bayes; the latter classifier performed best, achieving a classification accuracy of up to 81%.

Dodds et al. [24] use features extracted from lyrics and the ANEW dataset to measure the sentiment of songs, (also from blogs and State of the Union presidential speeches). The aim of the work is to quantify the evolution of the overall happiness in the different contexts. The approach calculates the average valence of each instance (song, blog post or speech) as a measure of happiness. The results show that, for example, valence can help distinguish between genres, when a large number of songs are considered.

From the related work it is clear that mood classification of music using lyrics is an emerging and interesting problem. However, it is difficult to compare previous findings since different works have reached contradictory conclusions based on experiments carried out on different private datasets using different evaluation methodologies. Thus, in this paper we present a comparison of different approaches for classifying music mood using lyrical features, and evaluate them using a large freely-available dataset using categories derived from Russell's model of affect.

3. DATASET

We perform our experimental studies using the *Million Song Dataset (MSD)* [10]. It is a large, freely-available dataset, which contains rich metadata and audio features for one million contemporary popular music tracks. We also use the *LastFm* [4] and *MusixMatch* [5] datasets, which expand the original *Million Song Dataset* providing metadata and lyrics for a subset of tracks.

The *LastFM dataset* contains song-level tags for more than 500,000 songs. The mood categories are derived using the social tags found in this dataset, following the approach proposed in [17, 18].

The *MusixMatch dataset* contains lyrics for 237,662 songs. Each song is described by word-counts of the top 5,000 stemmed terms across the set [6]. Specifically, we use the songs from the MSD for which social tags and lyrics are available. Furthermore, we only consider English language lyrics in this study. Thus, the resulting dataset used in this work contains 32,302 songs.

The use of this dataset is key regarding the reproducibility of the work presented here. However, given the format of the dataset (only word-counts for the top 5,000 terms are

[3] SentiWordNet: a database of sentiment information for english words, designed for opinion mining. `http://sentiwordnet.isti.cnr.it/`

[4] Last.fm dataset, the official song tags and song similarity collection for the Million Song Dataset, available at: `http://labrosa.ee.columbia.edu/millionsong/lastfm`.

[5] musiXmatch dataset, the official lyrics collection for the Million Song Dataset, available at: `http://labrosa.ee.columbia.edu/millionsong/musixmatch`.

[6] The terms were selected by its document frequency, normalised by the term frequencies in each song. We do not perform any post-processing on this set of terms (i.e. stop-words are not removed).

provided), our analysis is limited to a vector space model representation of songs, and more sophisticated natural language processing techniques [25] cannot be considered. This is a significant limitation of this particular dataset from a classification perspective as we will discuss further below.

3.1 Building the Mood Dataset

Three levels of granularity are considered for mood classification. To build the dataset we select a subset of songs for which the mood-related tags described in [18] are available. We select songs using the same criteria as used for the *MIREX 2009 mood multi-tag dataset* [7]; a song has to be tagged at least twice with one term in a tag group, or with at least two terms in a tag group, each at least once. Moreover, we remove repeated songs, (i.e. songs which have the same title and lyrics, but different ids in the dataset).

The mood groups are inferred as described in [18], where different *LastFM* tags are grouped to form a subset of predefined groups. For example, the group *G29* contains songs tagged as *aggression* and *aggressive*.

Finally, the mood quadrants as described in [8] are considered, where each quadrant represents a positive or negative value for valence and arousal. Table 1 shows the mood tags, groups and quadrants used in this work [8].

Tag	Group	Quadrant
aggression, aggressive.	G29	v^-a^+
angst, anxiety, anxious, etc.	G25	
anger, angry, choleric, fury, etc.	G28	
excitement, exciting, thrill, etc.	G1	v^+a^+
upbeat, gleeful, enthusiastic, etc.	G2	
cheerful, festive, jolly, etc.	G6	
happy, happiness, happy music, etc.	G5	
depressed, blue, dark, gloom, etc.	G16	v^-a^-
sad, sadness, unhappy, etc.	G15	
grief, heartbreak, sorrow, etc.	G17	
brooding, contemplative, etc.	G8	v^+a^-
alm, comfort, quiet, etc.	G12	

Table 1. Mood tags, groups and quadrants.

To illustrate the above, consider the song *Orchestra of Wolves* [9], by the British hardcore-punk band *Gallows*. This song is tagged as *aggressive* in the *LastFM* dataset, and therefore it is included in mood group $G29$ and quadrant v^-a^+ (given its negative valence and positive arousal values).

4. TERM-WEIGHTING SCHEMES

In this work, the vector space model is used to represent documents (songs), where each document $d = (t_1, t_2, \cdots, t_y)$ is represented by a vector in the y-dimensional term space.

[7] http://www.music-ir.org/mirex/wiki/2013:Audio_Tag_Classification

[8] Only tags which are associated with at least 100 songs are considered. Moreover, we discard groups G7, G9, G11, G14, G31 and G32 because they do not have enough tags or they can not be easily described in the valence-arousal space.

[9] http://open.spotify.com/track/5BorbORef4VQUlNOjAjoDT

The basic term weighting scheme we consider is the binary approach, in which each element of the vector is set to 1 or 0 to indicate the presence or absence of the corresponding term. A number of other term weighting schemes have been proposed in the literature [26,27]; in what follows, we describe some well known term-weighing schemes which are used in this work.

Term Frequency (tf) [27] accounts for the number of times a term t occurs in document d (denoted by $tf_{t,d}$). The rationale for this scheme is to assign higher weights to frequently occurring terms, since such terms are likely to be more characteristic of document content. Several normalisation approaches have been proposed for the original term frequency metric. Here, we use a standard logarithm normalisation, as shown in Equation 1.

$$ntf_{t,d} = \log\left(1 + tf_{t,d}\right). \tag{1}$$

Term frequency – inverse document frequency (tf-idf) combines the *tf* metric described above, with inverse document frequency (*idf*), which gives higher weights to terms which are rare in the collection. The *tf-idf* metric [28, 29] for a term t in document d is calculated as the product of $tf_{t,d}$ and $idf_{t,D}$, as shown in Equation 2.

$$tf\text{-}idf_{t,d,D} = tf_{t,d} \cdot \log \frac{|D|}{df_{t,D}}, \tag{2}$$

where the document frequency ($df_{t,D}$) is the number of documents in the collection D that contain the term t.

BM25 [30], is a sophisticated term-weighting scheme which has been widely used in text classification and retrieval. It is computed as per Equation 3:

$$BM25_{t,d,D} = \log \frac{|D| - df_{t,D} + 0.5}{df_{t,D} + 0.5} \frac{(k_1+1)tf_{t,d}}{k_1((1-b)+b\frac{L}{\bar{L}})+tf_{t,d}}, \tag{3}$$

where L is the document length and $\bar{L}$ is the average document length in the collection D. In this work, the parameters k_1 and b are set to typical values of 1.20 and 0.75, respectively [27].

what is doc length?

Delta tf-idf [31] is a scheme specifically proposed for sentiment classification. As shown in Equation 4, the term frequency of a term is multiplied by the δ function (Equation 5), which measures the relative document frequencies of a term in positive and negative instances. Thus, higher weights are assigned to terms which appear primarily in one class [10].

$$delta\ tf\text{-}idf_{t,d,D} = tf_{t,d} \cdot \delta_{t,D}. \tag{4}$$

$$\delta_{t,D} = \log_2 \left(\frac{df_{t,D^+} + 1}{df_{t,D^-} + 1} \right). \tag{5}$$

In the above, df_{t,D^+} and df_{t,D^-} are the document frequencies for term t in documents labeled as positive and negative, respectively.

[10] The original weight results in an infinite or undefined value if a particular term does not appear at least once in both classes. Thus, we modify the original equation by adding 1 to the document frequencies, as shown in Equation 5.

5. FEATURE ANALYSIS

In this section, we perform a preliminary feature analysis of song lyrics, examining how does the term distributions and different term weighting schemes affect the classification performance. For the sake of clarity, we perform the analysis on the mood quadrants dataset; however, similar trends are found in the mood groups and mood tags datasets.

We first study the term distribution across documents (songs) and classes (mood quadrants). To achieve high classification performance using lyrics alone, the vocabulary should be very different across moods. If many of the same terms occur in all classes, it will be difficult to classify those songs that contain these terms.

In total, there are 4,481 distinct terms in the dataset (i.e., vocabulary size). From Table 2, it can be seen that the majority of these terms (between 3,903 and 4,248 terms) occur in all classes. The overall distribution of terms across classes is as follows: 365 terms appear in a single class, 328 terms appear in two classes, 300 terms appear in 3 classes and 3,488 terms are common to all four classes. Thus, only a very small fraction of terms are unique to a single class, between 25 (class v^-a^+) and 180 (class v^+a^-) terms, indicating that the vocabulary of lyrics is, to a high degree, common across the four moods considered.

	v^+a^+	v^+a^-	v^-a^+	v^-a^-
Number of instances	6,973	14,685	1,958	8,686
Number of distinct terms per class	3,903	4,248	3,616	4,106
Number of unique terms per class	57	180	25	103
Mean (std. dev.) number of distinct terms per song	65 (48)	46 (50)	78 (40)	61 (49)
Mean (std. dev.) number of terms per song	189 (147)	134 (160)	224 (139)	180 (160)

Table 2. Term statistics for the mood quadrant dataset.

Table 3 shows the top ten terms for each mood quadrant, where the rank is produced by measuring the correlation between the term and the class, using Pearson correlation [32]. Nine of the top terms of the v^-a^+ quadrant shown in the table are intuitively related with moods from this quadrant (*aggressive, angry, etc.*). However, while some terms are correlated to one particular class, the same terms (*got, get, yeah*) are most highly correlated to both the v^+a^+ and v^-a^- mood quadrants. These are *connector* terms that are not related to mood. Moreover, a number of the top terms shown in Table 3 are stop-words (or at least would be considered as such in traditional information retrieval contexts), indicating the relative lack of discriminating terms in the dataset.

Table 2 also presents statistics on the total number (song length) and the number of distinct terms per song per class. While differences in these statistics are apparent — for example, on average, songs in class v^+a^- tend to be short while those in class v^-a^+ are the longest — there is significant variance evident in these statistics, thereby limiting their value from a classification perspective.

Rank	v^+a^+	v^+a^-	v^-a^+	v^-a^-
1	got (0.14)	dead (0.075)	love (0.231)	got (0.113)
2	get (0.14)	f**k (0.068)	f**k (0.217)	get (0.094)
3	yeah (0.136)	death (0.067)	hate (0.157)	yeah (0.09)
4	it (0.119)	love (0.062)	dead (0.155)	die (0.082)
5	oh (0.112)	die (0.061)	kill (0.149)	pain (0.08)
6	gonna (0.11)	scream (0.057)	blood (0.143)	it (0.077)
7	up (0.108)	blood (0.055)	s**t (0.13)	babi (0.074)
8	a (0.104)	hate (0.05)	burn (0.124)	tear (0.074)
9	do (0.101)	the (0.048)	death (0.122)	you (0.074)
10	you (0.1)	hell (0.047)	die (0.119)	up (0.073)

Table 3. Top terms ranked by Pearson correlation.

Figure 1 presents a histogram of term frequency (*tf*) values per song in the dataset, calculated over all songs. The graph shows the term frequency values on the horizontal axis and the count for each value on the vertical axis (both axis are presented in logarithmic scale). As can be seen, the vast majority of terms (98%) occur just once in songs, while only 1.1% of terms occur twice. Given these findings, little or no difference in classification results can be expected when the binary or term frequency weighting schemes are applied.

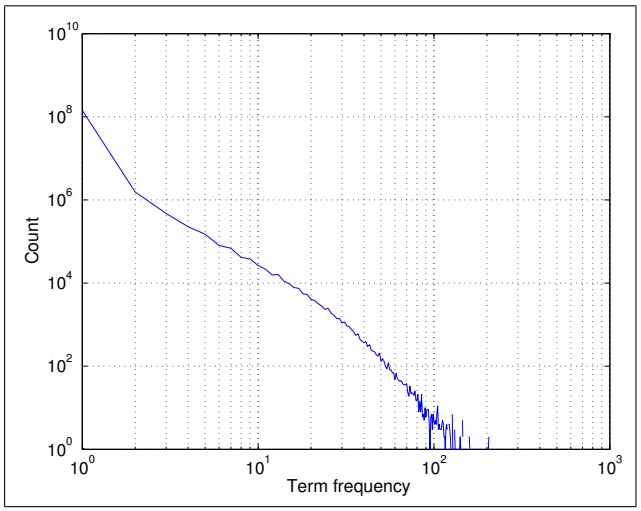

Figure 1. Term frequency distribution.

Figure 2 shows a histogram of the document frequency values (*df*) for all terms and documents. The horizontal axis shows the document frequency value for each term, with corresponding counts shown on the vertical axis (both axis are presented in logarithmic scale). The figure shows that the document frequency histogram follows a long-tail distribution, where most terms appear in a small subset of documents; for example, 1,044 terms (23%) appear in 20 documents (songs) or less, while 272 terms (6%) appear in more than 2000 (out of a total of 32,302) documents. Thus, this distribution of terms across documents is likely to limit the effect of *idf* term weighting. Moreover, the *idf* scheme does not consider term distribution with respect to class, and hence we also consider the *delta tf-idf* term weighting scheme, which does take class distribution into account.

The distribution of values for the δ function (Equation 5) is shown in Figure 3, where the horizontal axis represents δ values, while the vertical axis shows the count for each

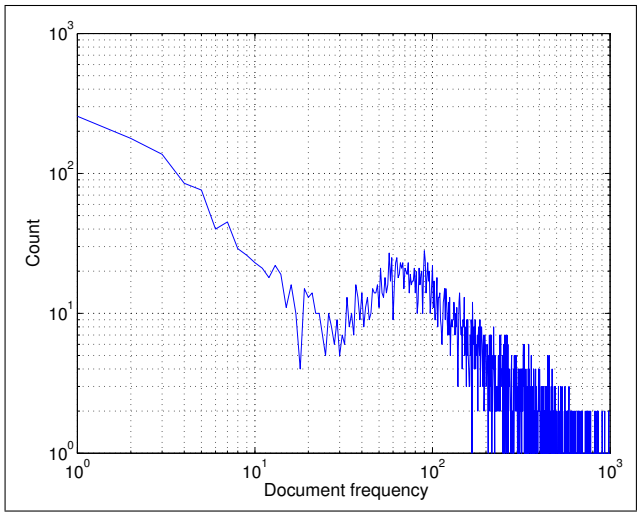

Figure 2. Document frequency distribution.

value, presented in a logarithmic scale. Each of the lines in the graph corresponds to a mood quadrant [11] . From the figure it is clear that a large number of terms are evenly distributed among the classes (i.e. at $\delta \approx 0$), while, on average across the mood quadrants, only 13% and 2% of term *delta* values are beyond ± 1 (i.e. corresponding to a ratio of 2:1 or above of term distribution across classes) and ± 2 (i.e. a ratio or 4:1 or above), respectively. Thus, given the distribution of terms across classes, the *delta tf-idf* weighting scheme may not appreciably affect classification performance.

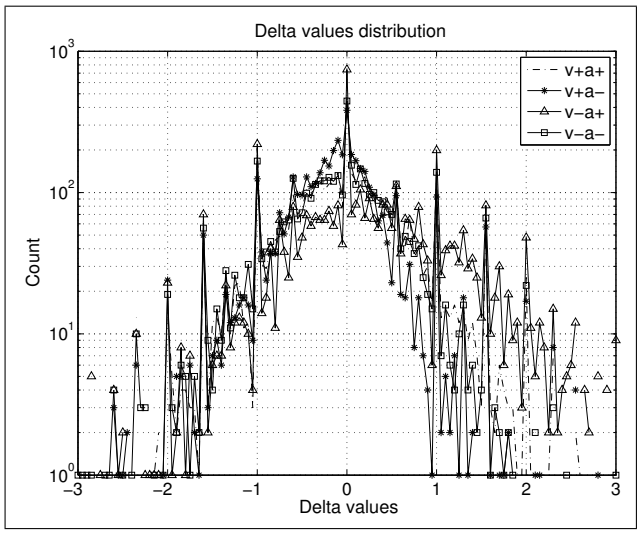

Figure 3. Distribution for δ values.

The analysis presented in this section, in particular the relative lack of discriminating terms in the dataset, may be an artefact of how the 5,000 terms were selected for inclusion in the *MusixMatch* dataset. As such, the analysis indicates a limitation in the use of this dataset for lyrics-based mood classification, a point to which we will return later in the paper.

[11] For each mood quadrant, δ values for terms are computed based on a random selection of 1,000 positive songs (i.e. from the mood quadrant in question) and 1,000 negative songs (i.e. from other mood quadrants).

6. MOOD CLASSIFICATION

We adopt a supervised classification approach where songs are represented using the vector space model. We experiment with the term weightings approaches described in Section 4 (*binary*, *tf*, *tf-idf*, *BM25* and *delta tf-idf*), comparing their performance using the three different mood granularities (i.e. class labels) described in Section 3.1. With this experiment, we aim to present a comprehensive and reproducible evaluation of music mood classification based on lyrics using the large, publicly available *Million Song Dataset*.

6.1 Experimental Methodology

We experiment with three different datasets in this evaluation, where each song is labelled according to one of the three mood granularities (i.e. mood quadrants, groups or tags) as shown in Table 1. In particular, balanced binary classifiers are created for each mood granularity by randomly selecting 1,000 positive training instances from each class; 1000 negative training instances are also randomly selected from other classes [12] .

Classification was performed using the Weka machine learning framework [33] with the LIBLinear L2-SVM classification algorithm [34], which is known to perform efficiently on large sparse datasets. Moreover, SVM classifiers have been used in the past in many binary classification scenarios with success [21, 31, 35].

Classification performance is evaluated using a standard 5-fold cross validation approach using the accuracy metric:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \qquad (6)$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives.

A Kruskal-Wallis test [36] at the 0.05 level is performed to determine whether statistically significant differences in results exist between the various term weighting schemes. Finally, the *delta tf-idf* weights are computed over training set instances only and these weights are then applied to test set instances.

6.2 Results

Tables 4, 5 and 6 show the results for the three different mood granularities and term-weighting schemes considered. Overall, it can be seen that the performance of the different term weighting approaches is very similar in terms of classification accuracy; no statistically significant differences in results were found. These results are expected given the analysis presented in the previous section. For example, most (98%) of the term frequency values seen in the dataset are equal to one. Thus, classification accuracy using the *tf* weighting scheme is close to the binary representation. Further, the distribution of terms

[12] When less than 1,000 positive instances are available, the maximum number of positive instances are selected, together with an equal number of negative instances. In the group and tags dataset, the mean number of positive (and negative) training instances is 744 and 426 per class, respectively.

across documents and classes in the dataset also limited the effect of the *idf*, *BM25* and *delta tf-idf* schemes.

Table 4 shows the results for mood quadrant granularity. Here, we can see that the classifier performs best for the v^-a^+ mood, which is interesting given that this class contains the least number (25) of unique terms (see Table 2). This result may be due to songs in this class containing the greatest numbers of total and distinct terms, although further analysis is required to test this hypothesis.

mood	size	accuracy				
		binary	tf	tf-idf	BM25	δ tf-idf
v^+a^+	1000	0.561	0.572	0.586	0.569	0.581
v^+a^-	1000	0.525	0.537	0.536	0.532	0.520
v^-a^+	1000	0.638	0.656	0.626	0.653	0.636
v^-a^-	1000	0.554	0.562	0.549	0.560	0.541

Table 4. Classification results for each mood quadrant.

Table 5 shows the classification results for mood groups. The best results are obtained for groups G29 (*aggression*, *aggressive*) and G28 (*anger*, *angry* etc.), where classification accuracies of 0.695 and 0.671 using binary term weighting are achieved, respectively. The remaining mood groups all have accuracies less than 0.6 (binary term weighting). Since both G29 and G28 belong to the v^-a^+ quadrant, these results confirm that the lyrics-based classification approach works well for songs in this quadrant.

mood	size	accuracy				
		binary	tf	tf-idf	bm25	δ-tf
G5	1000	0.566	0.573	0.559	0.570	0.542
G12	1000	0.549	0.548	0.525	0.562	0.517
G2	1000	0.555	0.571	0.571	0.570	0.558
G29	619	0.695	0.720	0.690	0.718	0.670
G28	1000	0.671	0.651	0.648	0.672	0.628
G1	196	0.574	0.571	0.543	0.567	0.543
G8	561	0.536	0.516	0.515	0.516	0.521
G15	1000	0.552	0.534	0.522	0.538	0.515
G6	530	0.555	0.558	0.555	0.558	0.531
G25	267	0.586	0.545	0.526	0.586	0.520
G17	749	0.592	0.599	0.571	0.592	0.570
G16	1000	0.599	0.600	0.575	0.598	0.574

Table 5. Classification results for each mood group.

Finally, the results obtained for the individual mood tags (Table 6) also align with the above findings, where high classification accuracies are seen for songs with tags belonging to the G29 group. Although. the best performance (0.767) is achieved for the tag "cool down" (which belongs to group G12 and quadrant v^+a^-), this particular tag appears infrequently in the dataset, thereby limiting its effectiveness.

mood	size	accuracy				
		binary	tf	tf-idf	bm25	δ-tf
mellow	1000	0.519	0.516	0.518	0.523	0.514
chillout	1000	0.562	0.558	0.551	0.553	0.542
happy	1000	0.562	0.568	0.554	0.561	0.558
aggressive	589	0.699	0.698	0.666	0.705	0.654
angry	821	0.649	0.668	0.633	0.665	0.624
soothing	271	0.505	0.494	0.494	0.514	0.522
melancholic	1000	0.557	0.577	0.569	0.570	0.554
calm	535	0.480	0.479	0.498	0.485	0.491
sad	1000	0.557	0.559	0.563	0.553	0.543
reflective	216	0.518	0.502	0.486	0.530	0.500
cheer up	112	0.544	0.563	0.535	0.536	0.544
depressing	267	0.524	0.597	0.585	0.584	0.554
depressive	126	0.703	0.644	0.620	0.651	0.616
dark	1000	0.582	0.615	0.597	0.603	0.577
depression	127	0.511	0.566	0.578	0.554	0.598
happiness	169	0.524	0.497	0.535	0.529	0.517
heartache	125	0.544	0.536	0.544	0.516	0.508
calming	131	0.505	0.550	0.588	0.531	0.554
wistful	209	0.510	0.493	0.505	0.493	0.486
sunny	156	0.519	0.536	0.516	0.513	0.510
cheerful	150	0.557	0.557	0.560	0.590	0.527
heartbreaking	173	0.552	0.532	0.518	0.523	0.468
rage	115	0.683	0.696	0.635	0.696	0.626
angst	179	0.547	0.565	0.556	0.579	0.541
cool down	172	0.767	0.796	0.770	0.779	0.776

Table 6. Classification results for each mood tag.

7. DISCUSSION

In this paper, we have presented a comprehensive evaluation of music mood classification, relying solely on lyrics as a source of information. We have studied three different granularities for mood representation (quadrants, groups and mood tags) and evaluated four term-weighting schemes (*tf, tf-idf, BM25* and *delta tf-idf*). In particular, we have used a large publicly available dataset in our analysis to enable reproducibility of experiments. This approach contrasts with previous work in this area, where much of the work has relied on small-scale, private datasets, and where contradictory results were reported in some instances.

The results obtained show that lyrics alone can be used for the mood classification task, performing particularly well for some moods (e.g. the v^-a^+ mood quadrant, where classification accuracies up to 70% were reached). However, in contrast to findings reported in [37], in this work no statistically significant differences in classification performance were found when using the various term-weighting schemes considered. These results align with [18], where the use of a smaller subset of term-weighting approaches (*binary*, *tf* and *tf-idf*) evaluated on a different dataset led to similar performance. The *delta tf-idf* term-weighting scheme also did not outperform other approaches in our analysis, which is somewhat surprising given this scheme takes term distribution across classes into account. Moreover, the results presented in [37], where a term-weighting scheme which also considers class distribution of terms is proposed, do not align with our findings, as they show a substantial improvement in classification performance over the *tf* approach.

Given the discrepancies in findings between the various works discussed above, clearly there is a need for a benchmark dataset to assess the performance of lyrics-based classification approaches. While the The *MusixMatch Dataset* and *Million Songs Dataset* represent significant steps in this direction, the analysis presented in Section 5 of this pa-

per highlights some important limitations in them. For example, only counts for the top 5,000 terms per song across the collection are made available, which precludes the application of more sophisticated natural language processing techniques to the classification task. Moreover, the approach used to selected the top 5,000 terms leads to a high degree of common terms across moods; as shown in Table 2, only 365 of the 4,481 terms are unique to one mood quadrant, which severely limits the discriminating power of these terms. In this regard, we conducted a small scale study involving 800 songs (200 for each mood quadrant) for which full lyrics are available. The term statistics in this dataset are very different: in total, there are 9,276 distinct terms (across all quadrants), with between 4,000 and 4,600 distinct terms per class, of which between 1,200 and 1,600 (on average, 32% per class) of these terms are *unique* to each class — which is clearly very different to the very low percentage of unique terms (on average, 2% per class) in the publicly available *MusixMatch* dataset used in this work.

In conclusion, while acknowledging that lyrics are copyrighted material and the legal considerations involved in making (full) song lyrics publicly available, the analysis presented in this paper highlights the need for a new benchmark dataset to progress work in this area. The provision of such a dataset would facilitate a true comparison of the different approaches to music classification, the reproducibility of experiments, and allow the true potential for lyrics-based classification approaches to be established.

8. ACKNOWLEDGEMENTS

This work is supported by Science Foundation Ireland through the Insight Centre for Data Analytics under grant number SFI/12/RC/2289.

9. REFERENCES

[1] A. Schindler and A. Rauber, *Capturing the Temporal Domain in Echonest Features for Improved Classification Effectiveness*. Springer International Publishing, 2012.

[2] M. F. McKinney and J. Breebaart, "Features for Audio and Music Classification." *ISMIR*, vol. 3, pp. 151–158, 2003.

[3] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *Speech and Audio Processing, IEEE transactions*, vol. 10, no. 5, pp. 293–302, 2002.

[4] M. Besson, F. Faita, and I. Peretz, "Singing in the Brain: Independence of Lyrics and Tunes," *Psychological Science*, vol. 9, no. 6, pp. 494–498, 1998.

[5] J. S. Downie, "When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis," *In Proceedings of the 11th Internation Society for Music Information Retrieval Conference (ISMIR 2010)*, pp. 619–624, 2010.

[6] R. Mayer, R. Neumayer, and A. Rauber, "Combination of Audio and Lyrics Features for Genre Classification in Digital Audio Collections Categories and Subject Descriptors," *In Proceedings of the 16th ACM International Conference on Multimedia*, pp. 159–168, 2008.

[7] C. McKay and I. Fujinaga, "Improving Automatic Music Classification Performance by Extracting Features from Different Types of Data," *In Proceedings of the international Conference on Multimedia Information Retrieval - MIR'10*, pp. 257–266, 2010.

[8] Y. Hu, X. Chen, and D. Yang, "Lyric-based Song Emotion Detection with Affective Lexicon and Fuzzy Clustering Method," *In Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pp. 123–128, 2009.

[9] R. Mihalcea and C. Strapparava, "Lyrics, Music, and Emotions," *In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 590–599, 2012.

[10] T. Bertin-mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The Million Song Dataset," *In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, pp. 591–596, 2011.

[11] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multi-Label Classification of Music into Emotions." *ISMIR*, vol. 8, pp. 325–330, 2008.

[12] R. Foucard and S. Essid, "Exploring New Features for Music Classification," *In Proccedings of the 14th International IEEE Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pp. 1–4, 2013.

[13] R. Mayer and A. Rauber, "Musical Genre Classification by Ensembles of Audio and Lyrics Features," *In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, pp. 675–680, 2011.

[14] Y. Song, S. Dixon, M. Pearce, and A. Halpern, "Do Online Social Tags Predict Percieved or Induced Emotial Responses to Music?" *In Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, pp. 89–94, 2013.

[15] K. Kosta, Y. Song, G. Fazekas, and M. B. Sandler, "A Study of Cultural Dependence of Perceived Mood in Greek Music," *In Proceedings of the 14th International Society for Music Information Retrieval (ISMIR 2013)*, pp. 317–322, 2013.

[16] J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[17] X. Hu, "Music and mood: Where Theory and Reality Meet," *In Proceedings of iConference*, pp. 1–8, 2010.

[18] H. Xiao, J. S. Downie, and A. F. Ehmann, "Lyric Text Mining in Music Mood Classification," *American Music*, vol. 183, no. 5040, pp. 411–416, 2009.

[19] M. Bradley and P. Lang, "Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings," The Center for Research in Psychophysiology, University of Florida, Tech. Rep., 1999.

[20] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann, "The 2007 MIREX Audio Mood Classification Task: Lessons Learned." *In Proceedings of the 9th International Conference of Music Information Retrieval (ISMIR 2008)*, pp. 462 – 467, 2008.

[21] X. Hu and J. S. Downie, "Improving Mood Classification in Music Digital Libraries by Combining Lyrics and Audio," *In Proceedings of the 10th Annual Joint Conference on Digital Libraries*, pp. 159–168, 2010.

[22] M. Kim and H.-C. Kwon, "Lyrics-Based Emotion Classification Using Feature Selection by Partial Syntactic Analysis," *23rd IEEE International Conference on Tools with Artificial Intelligence*, pp. 960–964, Nov. 2011.

[23] V. Kumar and S. Minz, "Mood Classifiaction of Lyrics using SentiWordNet," *2013 International Conference on Computer Communication and Informatics (ICCCI-2013)*, pp. 1–5, Jan. 2013.

[24] P. S. Dodds and C. M. Danforth, "Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents," *Journal of Happiness Studies*, vol. 11, no. 4, pp. 441–456, Jul. 2009.

[25] J. P. G. Mahedero, A. Martinez, P. Cano, M. Koppenberger, and F. Gouyon, "Natural Language Processing of Lyrics," *In Proceedings of the 13th Annual ACM International Conference on Multimedia - MULTIMEDIA '05*, pp. 475–478, 2005.

[26] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. ACM Press, New York, 1999, vol. 9.

[27] S. Ceri, A. Bozzon, and M. Brambilla, *An Introduction to Information Retrieval*. Springer Berlin Heidelberg, 2013.

[28] K. S. Jones, "A Statistical Interpretation of Term Specificity and its Application in Retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[29] A. Aizawa, "An Information-theoretic Perspective of tf-idf Measures," *Information Processing & Management*, vol. 39, no. 1, pp. 45–65, Jan. 2003.

[30] K. Sparck Jones, S. Walker, and S. Robertson, "A Probabilistic Model of Onformation Retrieval: development and comparative experiments," *Information Processing & Management*, vol. 36, pp. 809–840, 2000.

[31] J. Martineau and T. Finin, "Delta TFIDF: An Improved Feature Space for Sentiment Analysis," *In Proceedings of the Third International ICWSM Conference*, pp. 258–261, 2009.

[32] M. a. Hall, "Correlation-based Feature Selection for Machine Learning," Ph.D. dissertation, 1999.

[33] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: an Update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[34] R. Fan, K. Chang, and C. Hsieh, "LIBLINEAR: A Library for Large Linear Classification," *The Journal of Machine Learning*, vol. 9, pp. 1871–1874, 2008.

[35] Y. Song, S. Dixon, and M. Pearce, "A Survey of Music Recommendation Systems and Future Perspectives," *9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)*, pp. 19–22, 2012.

[36] E. Theodorsson-Norheim, "Kruskal-Wallis test: BASIC computer program to perform nonparametric one-way analysis of variance and multiple comparisons on ranks of several independent samples." *Computer Methods and Programs in Biomedicine*, vol. 23, no. 1, pp. 57–62, 1986.

[37] M. V. Zaanen and P. Kanters, "Automatic Mood Classification Using TF*IDF Based on Lyrics." *In Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pp. 75–80, 2010.