Music Social Tags Representation in Dimensional Emotion Models

Na He
School of Computer Science, Faculty of Engineering & IT
University of Technology Sydney
Sydney, Australia
winterhn@gmail.com

Sam Ferguson
School of Computer Science, Faculty of Engineering & IT
University of Technology Sydney
Sydney, Australia
samuel.ferguson@uts.edu.au

***Abstract*—Much research has sought to recognize and retrieve music on the basis of emotion labels. These labels are usually obtained from either subjective test or social tags. Researchers use social tags usually either by grouping tags to emotion categories or clusters directly, or by mapping tags to dimensional quadrants simply. Few research work have undertaken semantic analysis on social tags for projecting them into a dimensional emotion space, especially based on recent neural word embedding techniques using large-scale datasets. In this paper, we propose an effective solution to analyse music tag information and represent them in a 2-dimension emotion plane without limiting the corpus to contain only emotion terms. In our solution, we apply neural word embedding methods for tag representation, including Skip-gram, Continuous Bag-Of-Words (CBOW) and Global Vectors (GloVe). In our experiment, we compare these methods with traditional Latent Semantic Analysis (LSA) model based on Procrustes Analysis evaluation metrics. The results shows that neural tag embedding methods outperform LSA and represent tags with high approximations with classic circumplex emotion definitions.**

***Keywords*-dimensional emotion model, semantic analysis, social tags, tag embedding**

I. Introduction

During recent years, music emotion recognition in Music Information Retrieval (MIR) systems has become a very active research area, driven by the need to automatically detect emotion for music. A great deal of research has been contributed for this area such as music emotion classification [1], dynamic emotion detection [2], multimodal recognition [3].

Generally, there are two taxonomic methods for emotion representation: the categorical method and the dimensional method. The categorical method maps emotion descriptions into some typical discrete terms [4], categories [5] or clusters [6], while the dimensional method considers emotion as continuous values within a 2 or 3-dimensional space. Increasingly, research tends to use a dimensional model [7] or quadrants in dimensional space [8] to narrow the semantic gap between music experience and human perception, as opposed to using limited categorical emotion terms.

Corresponding to dimensional emotion representation, quantified emotion annotations are required. Such annotations could be conducted by subjective test, but it usually result in a heavy load on time consumption and labor cost [9], which is not tractable with large-scale datasets such as those seen in MIR. As an alternative source of annotation, increasing interest has been shown in crowdsourcing resources [10].

With the fast growth of web social media, social tags from community users is considered as a good way to providing annotation for music related tasks such as music automatic tagging [11], music emotion recognition [12] or sentiment analysis [13]. Compared with subjective annotation, social tags save more effort and serve training models better for large-scale dataset. As for dimensional emotion representation, previous research usually simply map social tags to the quadrants of classic dimensional emotion model such as Russell's circumplex plane [14], then define annotation schemes to project songs associated with tags to dimensional space [8], rather than analysing tags relationships in the context of music tag dataset. Only a few research projects focus on tags analysis to reflect themselves on dimensional emotion model [15], [16]), but these approaches use conventional latent semantic analysis (LSA) which purely depends on geometrical transformations rather than neural text analysis to automatically learn latent concepts.

In this research project, we propose an neural tag analysis solution for dimensional emotion representation with large-scale music datasets. Fig. 1 shows the work flow for this solution. In tags analysis, we consider tags as terms rather than single words because tags may be phrases or even sentences. The social tags dataset is preprocessed to generate a structured input such as a text corpus or a factorized matrix for the subsequent tag analysis models. Then the neural tag embedding model is trained and output vector-based terms. After extracting emotion-related term vectors, non-metric Multidimensional Scaling (nMDS) method is applied to this vector-based data to reduce the dimensionality into 2 dimensions (2D) or 3 dimensions (3D). Finally, Procrustes Analysis approach is used to make terms conform to dimensional emotion space and measure tag representation performance based on Warriner's emotional ratings [17].

The crucial part of this solution is tags embedding, where neural word embedding methods are implemented to replace LSA, including Skip-gram, Continuous Bag-Of-Words(CBOW) [18] and Global Vectors (GloVe) [19]. We compare their performance by using Procrustes root mean squred error (RMSE). The results show that neural word

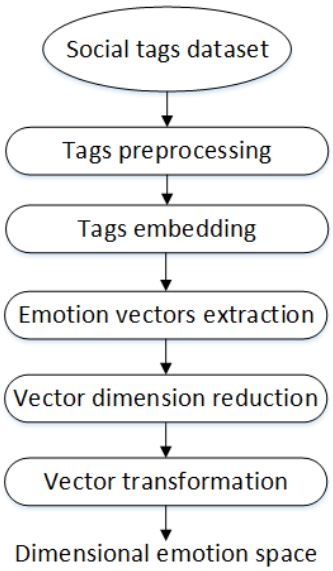

Fig. 1. Tags analysis solution overview

embedding methods outperform traditional semantic analysis methods. Using this solution, we are able to model joint representations of tags rather than limited to single type of tag corpus (such as emotion or genre only) and utilize social tags as a source of emotion annotation more reliably to quantify music in dimensional emotion space.

II. Backgrounds and Related Work

In this section, we introduce some backgrounds about emotion representation, emotion annotation and related work about word embedding models.

A. Emotion Representation

Music emotion model were originally developed in psychological studies. Early research by Hevner used experiments with approximately 450 subjects, resulting in eight adjective clusters (or categories) of affective terms, which were laid out in a circle [20]. While it was ostensibly a categorical model, it clearly laid the groundwork for a dimensional model of emotion. The dimensional approach was clearly articulated by Russell [14], who proposed that 28 affect words located in a 2 dimensional circumplex model, with a horizontal axis of *valence* (positive-negative) and a vertical axis of *arousal* (active-inactive). Following this work, Thayer [21] proposed another dimensional model based on tension and energy. These two models were integrated into a classic 2-dimension emotion model by Scherer [22] by positioning the two sets of two axes within a single space. Paltoglou and Thelwall [23] continued to refine emotion term positions within these spaces with improvements to the emotion coordinate accuracy. These dimension representation could be adopted for emotion prediction either in continuous values as regression problems or in categories based on quadrants as classification problems.

B. Music Emotion Annotation

Annotation is the process of obtaining emotion labels for songs so that emotion recognition training models can use them as the target labels. One method for this purpose is subjective testing, which can be conducted by either experts or candidates, annotating songs in categories [24], or rating songs in predefined range in a dimensional space [25], [26]. Another common method is to utilize crowdsourcing resources such as MTurk workers [2], collaborative games [27], social tags [13] or web service [28]. Among them, social tags is relatively mature resource to be explored and ready to use. However, social tags also contain some problems such as polysemy, misspellings, junk words and popularity bias [29], which need to be preprocessed before using as emotion labels.

Social tags is utilized for music emotion recognition (MER) in several ways. In a categorical model, researchers usually use tags as ground truth data directly [30]. While in a dimensional model, most research mainly uses subjective experiments [31] or maps tags to other existing models [8]. Then how to use tags to construct dimensional annotation is a real need for large-scale music dataset. Because such dataset is hard to be annotated manually considering labor and time cost. Saari and Eerola [15] have done some research in this area using conventional text analysis methods to generate Affective Circumplex Transformation (ACT), then calculate songs emotion based on associated tag weights and tag coordinates. In another research [12], songs are labelled with continuous arousal/valence values based on tags and crowdsourcing tag-related emotional ratings. However, no research mentions the most recent word embedding methods to mine the latent relationship of tags during tags analysis.

C. Word Representation

To facilitate text analysis, researchers usually do word representation using a vector-based model. In such model, a document or word is represented as a vector, where each dimension corresponds to one feature. Then these vectors could be used in language modelling and feature learning in a variety of applications, such as information retrieval [32], opinion mining [33], question answering [34], named entity recognition [35] and syntactico-semantic parsing [36].

Usually, the approaches of vector-based representation could be divided into supervised and unsupervised methods [37]. In supervised methods, one-hot representation in Natural Language Processing (NLP) means a simple word-based vector. It encodes binary vectors which have the same length as the size of the vocabulary. And only one element is '1' in each vector. Therefore it is easy to represent but fails to capture syntactic (structure) and semantic (meaning) relationships in text context. Another classic representation is Vector Space Model (VSM) [38] which is a document-based vector and use term-specific weights rather than binary data as element value. However, these two methods expose the drawback of data sparsity in large-scale text analysis.

In contrast, unsupervised methods show better effectiveness in handling large-size vocabulary and documents, which gen-

erate compact vectors with real value in low dimensions. They reduce the vector sparsity effectively and can better measure semantic similarity with other words. These models are commonly known as word embedding. One conventional word embedding method is distributional representation, the essence of which is dimensionality reduction and utilizing matrix factorization strategies. Among a variety of methods, a popular one is Latent Semantic Analysis (LSA) [39], [40] which actually performs Singular Value Decomposition (SVD). This method has been used widely in tag representation [16], and in music emotion modelling [15], [41], [42].

In emerging word embedding research, neural networks are leveraged to directly learn low-dimensional word representations rather than first reducing dimensionality. One notable technique in this branch of research is `word2vec` [18]. There are two typical models for this technique: Skip-gram and Continuous Bag-Of-Words (CBOW). Skip-gram aims to predicting context words from a given target word. While the CBOW architecture tries to predict the target word through its surrounding context. Word2vec focus on context information but poorly utilize global statistical data. Thus, it captures more syntactic regularities but few semantic regularities. Another popular technique is GloVe [19], which is a new global log-bilinear regression model combining global matrix factorization like LSA with local context window like word2vec. Due to this, it could cover both semantic and syntactic information better and outperform other models on word analogy, word similarity, and named entity recognition tasks.

III. Methodology

In this section, we describe how tags information is processed and represented in our solution. Tag preprocessing and tag embedding play an significant roles in reducing the sparsity of social tags.

A. Tag Preprocessing

We collect large-scale social tags from `Last.fm`[1] which is combined with the Million Song Dataset (MSD) [43] and have been used in many music classification research projects [16], [44], [45]. Based on previous research [29], it is necessary to preprocess tags to reduce the impact of noisy information and irrelevant information in tags dataset.

First, we need to construct a text corpus to facilitate tag preprocessing. In order to describe our solution comprehensively, we consider each track in songs dataset as a document and tags for one track as text in one document. Besides that, we call each tag as "term", not "word" since not all tags are single words. Like term frequency in documents, the `Last.fm` dataset contains tag popularity for tracks. Some researchers used these normalised counts to calculate Term Frequency-Inverse Document Frequency (TF-IDF) [15], [46]. In our solution, we combine these normalised counts and corresponding tags to build up tag content for each track so as to construct a text corpus for all tracks.

[1]http://www.last.fm

Once the text corpus is ready, we categorize tags to determine what strategies should be applied to this corpus to process different types of tags. Based on previous research work [5], [13], [15] and our cognition, we summarize music social tag categories with examples as below:

- **meaningless terms:**
  stop words: a, the, this, no, not
  junk tags or misspellings: zzzzzzz, Grrl
- **non-emotion terms:**
  opinion words: good, bad, poor
  genre, instrument, epoch, locale: jazz, guitar, 60s, usa
  ambiguous tags: love
  emotion-irrelevance tags : song, beat
- **emotion terms:**
  lemmatization: depression, depressive, depressed
  synonym: melancholy and sadness

1) Meaningless Terms: For the stop words, most of them are meaningless for semantic analysis but take the high proportion in the corpus. We remove them by referring to the `snowball` list of stopwords[2]. At the same time, we remain negative words from this list thereby keeping the original meaning for some terms such as 'not happy'. For the junk tags and misspellings, they should be removed to improve the validity of tags. Considering the variety of tag content, it is impossible to find out all tags mentioned above and filter them manually. Supposing that these terms are either very common or low-frequency, we set a series of statistical thresholds to filter them:

- **term_count_min:** minimum number of occurrences over all documents
- **doc_proportion_max:** maximum proportion of documents which should contain term
- **doc_proportion_min:** minimum proportion of documents which should contain term

Through this way, we filter out most of meaningless, noisy, high-frequency terms. To some extent, the thresholds determines the quality of term analysis to balance between removing irrelevant information and avoiding information loss.

2) Non-emotion Terms: Previous research work [29] explored `Last.fm` tags dataset and found that tags mainly include genre, emotion, instruments,locales, opionions and so on, among of which, genre accounts for high proportion (68%) followed by locale (12%) while mood only accounts for 5% followed by opinion (4%) and instruments (4%). Focusing on emotion tags analysis, most research usually remove all non-emotional terms and only keep tracks labeled by emotion tags. Such approach results in a great deal of information loss and limits generalization since a large number of tracks without emotion tags are excluded. In our research, we keep these tracks involved in subsequent tags embedding analysis so that we can explore more tags relationship in vector space. Using our method, even a track is not labelled by emotion tags, it could be linked emotion terms through term similarity and analogy. For ambiguous tags and other emotion-irrelevance

[2]http://snowball.tartarus.org/algorithms/english/stop.txt

tags, we exclude most of them through statistical filtering mentioned in Section III-A1

3) Emotion Terms: Considering the inflection of words and synonyms, some researchers [5] tried to build up synsets for clustering emotion terms while others [13] extended term inflected forms derived from a lemmatization process to construct emotion corpus. In our research, we make no change for all emotion terms and explore that whether these terms have distinct dimensional values from each other.

After the above preprocessing, the final tag corpus is established and then a corpus of textual data representing all tracks is vectorized for further use.

B. Tag Embedding

Corresponding to different word embedding models, different types of input are required and constructed from the vectorized text corpus mentioned above.

Conventional Latent Semantic Analysis (LSA) technique requires Document-term matrix (DTM) as input. DTM describes the frequency of terms that occur in a collection of tag documents. To reduce the impact of high-frequency terms, the term-weighting scheme TF-IDF [47] is usually applied to adjust term weights. In our research, we apply LSA to our tag analysis as a baseline.

Neural word-embedding models CBOW and Skip-gram take the vectorized text corpus as input directly rather than using global matrix factorization. These two models are able to learn local context for each term automatically.

GloVe trains its model based on term-co-occurrence matrix (TCM). TCM is the statistics of terms in the vectorized corpus in a form of matrix X. Each element X_{ij} in such matrix represents how often term i appears in context of term j. The algorithm utilizes a new weighted least squares regression model. It defines a cost function like this:

$$J = \sum_{i=1}^{V}\sum_{j=1}^{V} f(X_{ij})(w_i^T w_j + b_i + b_j - log(X_{ij}))^2 \qquad (1)$$

Here w_i means the vector for the main term i and w_j means the vector for the context term j. b_i, b_j are scalar biases for the main and context terms. f is a weighting function that avoids frequent co-occurrences being overweighted, see definition as in:

$$f(X_{ij}) = \begin{cases} (\mathrm{X}_{ij}/X_{max})^{\alpha} & \text{if } \mathrm{X}_{ij} < X_{max} \\ 1 & \text{otherwise} \end{cases} \qquad (2)$$

Here X_{max} defines the threshold of term co-occurrences value. Only X_{ij} less than X_{max} take effect to regression model through f. α is a factor in weighting function, set to 0.75 by default.

Through tag embedding, all of these models output a vector-based matrix, where

- each row corresponds to each term in tag corpus
- all elements in each row is a feature vector for that term
- the vector size means dimensions of tag embedding

C. Emotion Vectors Extraction

Once vector-based matrix is ready, we could extract vectors for specified terms to analyse tags relationship and check the tag-embedding model performance. A set of terms can belong to one specified type such as emotion, genre or themes, which reflects the generalization of our solution. In this paper, we focus on emotion tags analysis, hence the emotion vocabulary is defined for vectors extraction.

D. Vector-based Data Scaling and Transformation

In this step, we apply non-metric multidimensional scaling (nMDS) method and Procrustes Analysis (PA) to emotion-term vectors for vector-based dimensionality reduction and transformation. Dimensional emotion models proposed in the previous research are usually two or three dimensions with the reason that more than 3 dimensions could not reflect emotion variation intuitively and one dimension could not distinguish emotion sufficiently. In this research, we utilize nMDS [48] to generate 2D and 3D models separately for performance comparison, then Procrustes transformation [49] make our tags approximate to a classic Valence-Arousal (VA) model.

To assess the quality of nMDS, we compare nMDS with other typical dimensionality reduction solutions including Principal component analysis (PCA), Locally Linear Embedding (LLE), kernel PCA (kPCA) and AutoEncoder, through R_{NX} measurement defined in [50]. Taking a set of our vector-based data as input, the R_{NX} values with log-scaled rank K are shown in Fig. 2. The results show that nMDS is the best way to representing the pairwise distance and dissimilarity for terms in a low dimensions meanwhile keeping the pairwise relationship changing as few as possible.

In the procedure of Procrustes Analysis, we transform our 2D and 3D vectors through translation, rotation and scaling to find a optimal approximation to the VA reference. Given the VA reference as target X, our vectors as Y is transformed to conform to X. Y_t is the final 2D emotion tags representation. Equation (3) shows how PA works.

$$Y_t = f(Y) = b * Y * T + c \qquad (3)$$

where b is scale component, T is orthogonal rotation and reflection component and c is translation component.

To obtain the better goodness-of-fit, b, T and c are adjusted to minimize the root mean squared error (RMSE) defined as in:

$$f(Y) \rightarrow \sqrt{\frac{\sum_{min}(X - Y_t)^2}{N}} \qquad (4)$$

where N means the number of terms in our vector-based data.

If Y is 3D vectors, then 2D X is filled with one zero column to match dimensions. Alternatively, the dominance ratings [17] could be used as the third dimension but we don't discuss this situation here.

IV. Experiments

Following the workflow of our solution, we describe our experiment step by step.

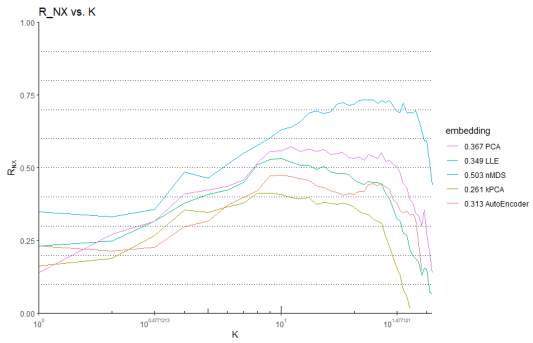

Fig. 2. Performance comparison of different dimensionality reduction methods, where a value of 0 corresponds to a random embedding and a value of 1 to a perfect embedding into the k neighborhood. The legend contains AUC_{lnk} meansurement defined in [50].

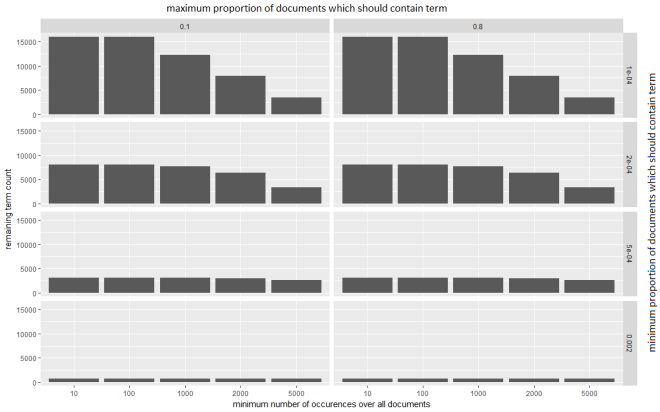

Fig. 3. Terms count statistics after filtering

A. Dataset Preprocessing

We collect social tag dataset from `Last.fm` associated with 504555 tracks in MSD and create a vocabulary containing a total of 463487 unique tag terms after removing stop words. Then we filter very common terms and low-frequency terms based on statistical thresholds. Fig. 3 shows how many terms are left with different combinations of thresholds and Table I is our final setting to balance information noise with information loss. This results in a total of 7685 terms and a corpus of 470280 tracks. Finally, we vectorize this corpus for subsequent process.

B. Tag Embedding

In the process of tag embedding, embedding dimensions K need to be determined. Given that the final emotion model is 2 or 3 dimensions, the higher K values would increase the dimension gap and may give rise to underfitting by leaving out important dimensions of the dissimilarity data when implementing dimensionality reduction. Due to this, K is selected from the set of {4,8,16,32,64,128}. Table II lists the detail of input and some key hyper-parameters used in embedding models. For other parameters, we just reuse the default values.

TABLE I
THRESHOLDS OF TERM FILTERING

Parameter	Value
term_count_min	1000
doc_proportion_max	0.8
doc_proportion_min	0.0002

TABLE II
TAG-EMBEDDING MODELS SUMMARY

Model	Input	Hyper-parameters for embedding
LSA	DTM with TF-IDF	vector size = K
CBOW	Corpus	vector size = K context window = 5 training epoch = 10
Skip-gram	Corpus	vector size = K context window = 5 training epoch = 10
GloVe	TCM context window = 5 "right" context terms are used for statistics	vector size = K X_{max} = 10 training epoch = 25 learning rate = 0.15

In this experiment, we use the natural language processing package `text2vec`[3] in R software environment to generate pruned corpus, DTM with TF-IDF, TCM and run LSA and GloVe models. CBOW and Skip-gram models are implemented by `word2vec` function in python library `gensim`[4].

C. Emotion Terms Selection

The selecting of emotion vocabulary is based on dimensional emotion models [14], [15], [22], emotion clustering [5], [16], and MIREX Mood Categories [51]. In our experiment, we select 44 common emotion terms rather than all of emotion terms in vector-based matrix, corresponding to emotion quadrants with considering term balance in each quadrant. See Table III.

TABLE III
THE EMOTION TAGS IN DIMENSIONAL QUADRANTS

Q1	Q2	Q3	Q4
happy	angry	sad	relax
joyful	brutal	melancholy	calm
party	aggressive	sadness	peaceful
fun	scary	depressive	mellow
sexy	frustration	bittersweet	sweet
upbeat	bitter	gloomy	soothing
uplifting	sarcastic	sorrow	hopeful
exciting	cynical	desperate	dreamy
triumphant	black	dark	chill
intense	quirky	lonely	serious
romantic	heartbroken	sleepy	quiet

D. Data Scaling and Transformation

In this step, we applied non-metric MDS method to reduce K-dimension vectors to 2D and 3D vectors respectively, followed by Procrustes transformation. In our experiment, we

[3]https://cran.r-project.org/web/packages/text2vec/
[4]https://pypi.org/project/gensim/

choose the target Arousal-Valence reference from Warriner's list [17] which provides continuous ratings of valence, arousal and dominance for almost 14000 English words. The PA performance comparison between four tag embedding models is shown in Fig. 4 based on Procrustes RMSE as the evaluation metrics.

E. Tags Visualization

Finally, we use the transformed results from PA to visualize the final 2D emotion space based on social tags. Here we select two models with best performance: *GloVe_64D+MDS_2D* and *Skip-gram_64D+MDS_3D*. The results are shown in Fig. 5.

V. Discussion

In this section, we compare the performance of neural tag embedding models with LSA baseline. Further, tag topology structure based on dimensional emotion space is visualized for analysing the influence factors of tag representation.

A. Tag Embedding Models Performance

As seen in Fig. 4, better performance are located in higher K range $\{32, 64, 128\}$ for all models because that values narrow the gap between the sparse high-dimension corpus and K-dimension embedding vectors. For each K-dimension embedding, the best solution is one of neural tag embedding models rather than LSA. It demonstrates that neural word embedding techniques outperform conventional text analysis methods. Further, the performance of GloVe and Skip-gram models vary dramatically with K value changing while the performance of CBOW model is relatively stable. It can be seen that GloVe and Skip-gram models are more sensitive to the embedding size, and hence selecting appropriate size could achieve better performance. With regard to the selection of 2D or 3D vector space, there is no a certain regularity to impact performance. In our experiment, the best results are shown in 64D tag embedding, where GloVe-based vectors are reduced to 2D while Skip-gram-based vectors are reduced to 3D.

B. Tags Visualization

In the 2D emotion models as shown in Fig. 5, we can see that tags could be divided into four parts conforming to classic emotion models. (b) representation shows the typical terms better than (a) such as happy, angry, sad, relax. The deviation of 'sad' in (a) is caused by GloVe combining global matrix factorization. In such latent semantic analysis, a lot of co-occurrence terms with 'sad' are non-emotion terms, but these terms co-occur with other high-frequency emotion terms which are not located in Q3. While Skip-gram model in (b) only utilize local context information which reflects latent relationship with 'sad' better and reduce the impact of noise information. It illustrates that terms in our tag corpus are strongly correlated with the terms nearby with similar popularity, and cleaning irrelevant information is very important for GloVe models as it covers global context. Skip-gram is the better choice without cleaning the corpus on a large scale.

Another reason for the position deviation is that PA minimize the sum of residuals for all tags between our term vectors and AV reference. To serve the whole performance, some terms sacrifice their correct positions to bridge the gap of inappropriate positions of terms influenced by irrelevant information, such as angry and aggressive. Besides that, the Warriner's AV ratings bring about some deviations. Because those ratings are based on word stimuli rather than music, the difference exist for some terms. For example, 'heartbroken' in our tag analysis models are located in low arousal, but its reference is labeled in high arousal. Similarly, 'black' and 'dark' are usually linked with low valence and middle or high arousal but their ratings are opposite in AV reference. Therefore, a better AV reference could enhance the transformation performance as well.

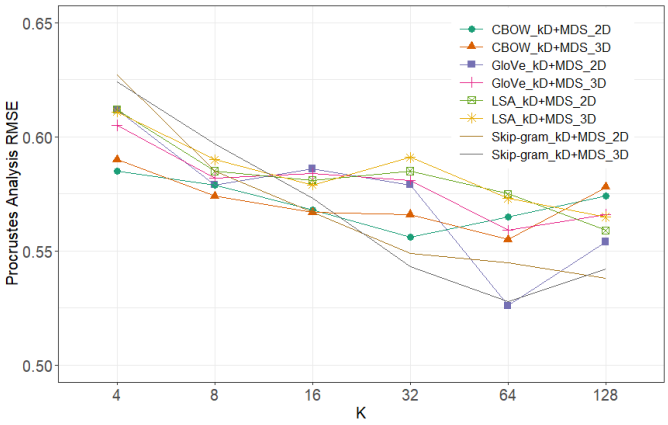

Fig. 4. Procrustes analysis performance comparison between different models

C. Music Emotion Annotation based on Tags

In Music Emotion Recognition (MER) tasks, social tags is used for music annotation in several ways [8], [13], [16]. But most prior work applied these tags to solve music classification problems rather than regression problems. Because there is no good way to quantify tags and then quantify emotion for songs. While our solution provides a more flexible and effective way to represent tags as embedding vectors. On one hand, we could calculate the quantified emotion for music based on tags as mentioned in [15]. On the other hand, for songs without labelled emotion tags, we still can represent them in dimensional emotion model based on tag semantic analysis and songs similarities. It is more reasonable to get emotion labels for more songs.

VI. Conclusion

In this research, we propose an effective solution to analyse music social tags relationships where the neural word embedding techniques are applied to obtain vector-based terms for large-scale tags dataset. Apart from that, non-metric MDS and Procrustes transformation are utilized to visualize terms in 2D emotion model. The experimental results show that neural tag embedding models outperform conventional LSA model. The whole solution achieves the objective of transforming the sparse, discrete and messy tags information to dense,

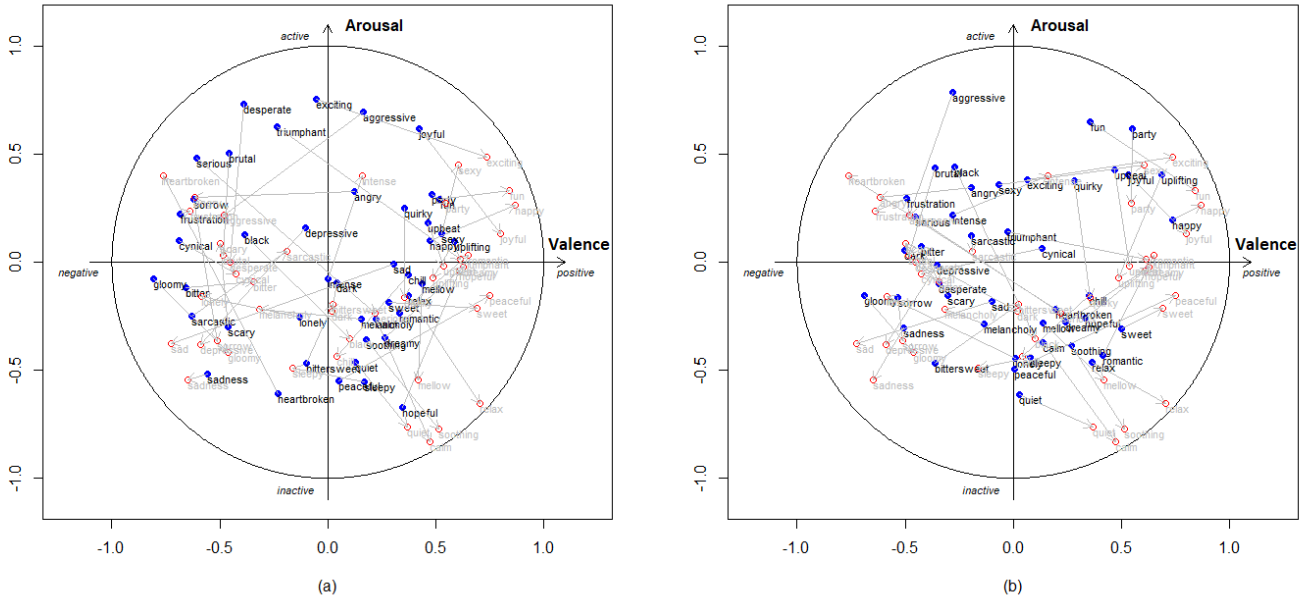

Fig. 5. 2D emotion model computed from Procrustes Analysis. (a) denotes *GloVe_64D+MDS_2D* and (b) denotes *Skip-gram_64D+MDS_3D*. In each model, blue dots represent our tag positions, red circles represent Warriner's AV reference of tags.

quantified and correlated data. In this way, it allows more than one type of terms (e.g. genre, emotion) in the corpus for tag analysis and is capable of reflecting tags relationship covering multiple vocabulary sets. Moreover, songs without labelling emotion tags could be linked with emotion terms through semantic analysis. This provides a good resource for emotion annotation in music emotion recognition work.

REFERENCES

[1] X. Hu, K. Choi, and J. S. Downie, "A framework for evaluating multimodal music mood classification," *Journal of the Association for Information Science and Technology*, vol. 68, no. 2, pp. 273–285, 2017.

[2] A. Aljanaki, Y. H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLoS ONE*, vol. 12, no. 3, 2017.

[3] B. Jeon, C. Kim, A. Kim, D. Kim, J. Park, and J. W. Ha, "Music emotion recognition via end-To-end multimodal neural networks," in *CEUR Workshop Proceedings*, vol. 1905, 2017.

[4] C. Laurier, O. Meyers, J. Serrà, M. Blech, P. Herrera, and X. Serra, "Indexing music by mood: Design and integration of an automatic content-based annotator," in *Multimedia Tools and Applications*, vol. 48, no. 1, 2010, pp. 161–184.

[5] X. Hu, J. S. Downie, and A. F. Ehmann, "Lyric text mining in music mood classification," *American Music*, vol. 183, no. 5,049, pp. 2–209, 2009.

[6] A. Bhattacharya and K. V. Kadambari, "A Multimodal Approach towards Emotion Recognition of Music using Audio and Lyrical Content," *arXiv preprint arXiv:1811.05760*, 2018. [Online]. Available: http://arxiv.org/abs/1811.05760

[7] J. Grekow, "Music emotion maps in the arousal-valence space," in *Studies in Computational Intelligence*, 2018, vol. 747, pp. 95–106.

[8] R. Panda, R. M. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, 3 2018.

[9] Y.-H. Yang and H. H. Chen, "Machine Recognition of Music Emotion: A Review," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 3, pp. 1–30, 2012.

[10] E. Çano and M. Morisio, "Crowdsourcing Emotions in Music Domain," *International Journal of Artificial Intelligence & Applications*, vol. 8, no. 4, pp. 25–40, 2017.

[11] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," in *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016*, 2016, pp. 805–811.

[12] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, "Music mood detection based on audio and lyrics with deep neural net," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018*, 2018, pp. 370–375.

[13] E. Cano and M. Morisio, "Music Mood Dataset Creation Based on Last FM Tags," in *2017 International Conference on Artificial Intelligence and Applications, Vienna, Austria*, 2017, pp. 15–26.

[14] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[15] P. Saari and T. Eerola, "Semantic Computing of Moods Based on Tags in Social Media of Music," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 10, pp. 2548–2560, 2014.

[16] C. Laurier, M. Sordo, and J. Serrà, "Music mood representations from social tags," in *Proc. International Society for Music Information Retrieval Conference*, 2009, pp. 381–386.

[17] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behavior Research Methods*, 2013.

[18] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119. [Online]. Available: http://arxiv.org/abs/1310.4546

[19] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[20] K. Hevner, "Experimental Studies of the Elements of Expression in Music," *The American Journal of Psychology*, vol. 48, no. 2, pp. 246–268, 1936.

[21] R. Thayer, *The biopsychology of mood and arousal*, 1989.

[22] K. R. Scherer, "What are emotions? and how can they be measured?" *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005.

[23] G. Paltoglou and M. Thelwall, "Seeing stars of valence and arousal in blog posts," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 116–123, 2013.

[24] C. Lin, M. Liu, W. Hsiung, and J. Jhang, "Music emotion recognition based on two-level support vector classification," in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2016, pp. 375–389.

[25] J. Grekow, "Audio features dedicated to the detection of arousal and valence in music recordings," in *Proceedings - 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications, INISTA 2017*, 2017, pp. 40–44.

[26] Y. H. Yang and H. H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 762–774, 2011.

[27] E. L. Law, L. V. Ahn, R. B. Dannenberg, and M. Crawford, "Tagatune: A game for music and sound annotation," in *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007*, 2007, pp. 361–364.

[28] K. Knautz, D. R. Neal, S. Schmidt, T. Siebenlist, and W. G. Stock, "Finding Emotional-Laden Resources on the World Wide Web," pp. 217–246, 2011.

[29] P. Lamere, "Social tagging and music information retrieval," *Journal of New Music Research*, vol. 37, no. 2, pp. 101–114, 2008.

[30] Y.-C. Lin, Y.-H. Yang, and H. H. Chen, "Exploiting online music tags for music emotion classification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 7S, no. 1, pp. 1–16, 2011.

[31] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 448–457, 2008.

[32] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones, "Word Embedding based Generalized Language Model for Information Retrieval," 2015, pp. 795–798.

[33] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, "Sentiment analysis leveraging emotions and word embeddings," *Expert Systems with Applications*, vol. 69, pp. 214–224, 3 2017.

[34] A. Bordes, S. Chopra, and J. Weston, "Question Answering with Subgraph Embeddings," in *Empirical Methods in Natural Language Processing(EMNLP)*, 2014. [Online]. Available: http://arxiv.org/abs/1406.3676

[35] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *arXiv preprint arXiv:1603.01360*, 2016.

[36] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, "Parsing With Compositional Vector Grammars," *Advances in Neural Information Processing Systems*, pp. 455–465, 2013.

[37] J. Turian, L. Ratinov, and Y. Bengio, "Word representations : A simple and general method for semi-supervised learning," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, 2010, pp. 384–394.

[38] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975. [Online]. Available: http://portal.acm.org/citation.cfm?doid=361219.361220

[39] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[40] N. E. Evangelopoulos, "Latent semantic analysis," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 4, no. 6, pp. 683–692, 2013.

[41] M. Levy and M. B. Sandler, "A semantic space for music derived from social tags," *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, 2007.

[42] A. Schindler and P. Knees, "Multi-Task Music Representation Learning from Multi-Label Embeddings," in *Proceedings - International Workshop on Content-Based Multimedia Indexing*, vol. 2019-Septe. IEEE, 2019, pp. 1–6.

[43] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The Million Song Dataset." in *Ismir*, vol. 2, no. 9, 2011, p. 10.

[44] X. Hu and J. S. Downie, "Exploring Mood Metadata: Relationships with Genre, Artist and Usage Metadata," in *8th International Society for Music Information Retrieval Conference – ISMIR 2007*, 2007, pp. 67–72.

[45] Y. Song, S. Dixon, M. T. Pearce, and A. R. Halpern, "Perceived and Induced Emotion Responses to Popular Music: Categorical and Dimensional Models," *Music Perception: An Interdisciplinary Journal*, vol. 33, no. 4, pp. 472–492, 2016.

[46] M. Levy and M. Sandler, "Learning latent semantic models for music from social tags," *Journal of New Music Research*, vol. 37, no. 2, pp. 137–150, 2008.

[47] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting TF-IDF term weights as making relevance decisions," *ACM Transactions on Information Systems*, vol. 26, no. 3, pp. 1–37, 2008.

[48] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[49] J. C. Gower, *Procrustes Analysis*, 2015.

[50] G. Kraemer, M. Reichstein, and M. D. Mahecha, "dimRed and coRanking-unifying dimensionality reduction in R," *R Journal*, vol. 10, no. 1, pp. 342–358, 2018.

[51] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann, "The 2007 mirex audio mood classification task: Lessons learned," in *ISMIR 2008 - 9th International Conference on Music Information Retrieval*, 2008, pp. 462–467.