Multi-Emotion Classification for Song Lyrics

Darren Edmonds
Donald Bren School of ICS
University of California, Irvine
dedmond1@uci.edu

João Sedoc
Stern School of Business
New York University
jsedoc@stern.nyu.edu

Abstract

Song lyrics convey a multitude of emotions to the listener and powerfully portray the emotional state of the writer or singer. This paper examines a variety of modeling approaches to the multi-emotion classification problem for songs. We introduce the Edmonds Dance dataset, a novel emotion-annotated lyrics dataset from the reader's perspective, and annotate the dataset of Mihalcea and Strapparava (2012) at the song level. We find that models trained on relatively small song datasets achieve marginally better performance than BERT (Devlin et al., 2019) fine-tuned on large social media or dialog datasets.

1 Introduction

Text-based sentiment analysis has become increasingly popular in recent years, in part due to its numerous applications in fields such as marketing, politics, and psychology (Rambocas and Pacheco, 2018; Haselmayer and Jenny, 2017; Provoost et al., 2019). However, the vast majority of sentiment analysis models are built to identify net positive or negative sentiment rather than more complex, ambiguous emotions such as anticipation, surprise, or nostalgia (Jongeling et al., 2017). As a result, current models usually fail to portray the coexistence of multiple emotions within a text sample, resulting in limited characterization of a human's true emotions. Songs are often created to elicit complex emotional responses from listeners, and thus are an interesting area of study to understand nuanced emotions (Mihalcea and Strapparava, 2012).

This paper examines a variety of approaches to address the multi-emotion classification problem. We aim to build an emotion classification model that can detect the presence of multiple emotions in song lyrics with comparable accuracy to the typical inter-annotator agreement for text-based sentiment analysis (70-90%) (Diakopoulos and Shamma, 2010; Bobicev and Sokolova, 2017; Takala et al., 2014). Building such a model is especially challenging in practice as there often exists considerable disagreement regarding the perception and interpretation of the emotions of a song or ambiguity within the song itself (Kim et al., 2010).

There exist a variety of high-quality text datasets for emotion classification, from social media datasets such as CBET (Shahraki, 2015) and TEC (Mohammad, 2012) to large dialog corpora such as the DailyDialog dataset (Li et al., 2017). However, there remains a lack of comparable emotion-annotated song lyric datasets, and existing lyrical datasets are often annotated for valence-arousal affect rather than distinct emotions (Çano and Morisio, 2017). Consequently, we introduce the Edmonds Dance Dataset[1], a novel lyrical dataset that was crowdsourced through Amazon Mechanical Turk. Our dataset consists of scalar annotations for the 8 core emotions presented by Plutchik (2001), with annotations collected at the song level and from the reader's perspective.

We find that BERT models trained on out-of-domain data do not generalize well to song lyrics and have lower F1 scores than Naive Bayes classifiers for emotions such as disgust and fear. However, BERT models trained on small lyrical datasets achieve marginally better performance, despite in-domain datasets being orders of magnitude smaller than their counterparts. We also find that surprise has significantly lower inter-annotator agreement and test accuracy than other core emotions.

2 Related Work

A multitude of models and techniques have been explored for song emotion classification. Both He et al. (2008) and Wang et al. (2011) found that fea-

[1]The Edmonds Dance dataset is available by request from the authors of this paper.

Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 221–235

ture extraction from lyrics improves emotion classification performance. Researchers have trained Naive Bayes, HMM, SVM, clustering, and Random Forest models on lyrical and sometimes audio features to predict emotion in songs (Hu et al., 2009; Kim and Kwon, 2011; Jamdar et al., 2015; An et al., 2017; Rachman et al., 2018). Deep learning frameworks have also been widely utilized for song emotion classification, ranging from CNNs and LSTMs (Delbouys et al., 2018; Abdillah et al., 2020) to transformer-based models such as BERT and ELMo (Parisi et al., 2019; Liu and Tan, 2020).

Multiple researchers have taken a multi-modal approach to emotion prediction. Strapparava et al. (2012), introduced a novel corpus of both music and lyrics, and achieved promising results when using both musical and lyrical representations of songs in emotion classification. Similarly, Yang et al. (2008) found an increase in 4-class emotion prediction accuracy from 46.6 to 57.1 percent when incorporating lyrics into models trained on audio.

However, audio data can lead to problematic bias in emotion classification. Susino and Schubert (2019b) explored the presence of emotion stereotyping in certain genres, and found that heavy metal and hip-hop music were perceived to have more negative emotions than pop music with matched lyrics. Susino and Schubert (2019a) also found that emotional responses to an audio sample of a song could be predicted by stereotypes of the culture with which the song's genre was associated. Additionally, Fried (1999) found that violent lyrical passages were seen to be significantly more negative when represented as rap songs rather than country songs. Dunbar et al. (2016) validated Fried's findings through multiple studies in which participants believed that identical lyrics were more offensive when portrayed as rap rather than country music.

Lyrics are paramount for the accurate prediction of emotion in music. Yang and Lee (2009) transformed song lyrics into psychological feature vectors using a content analysis package and concluded that song lyrics alone can be used to generate promising, human-comprehensible classification models. Hu et al. (2009) found that audio features did not always outperform lyric features for mood prediction, and that combining lyric and audio features does not necessarily improve mood prediction over simply training on lyrics features. In later research, Hu and Downie (2010) found that lyrics features significantly outperformed audio features in 7 of 18 mood categories, while audio features outperformed lyrical features in only one.

Research is split regarding crowdsourced emotion annotation quality; while Mohammad and Bravo-Marquez (2017) achieved strong results through crowdsourcing labels, Hasan et al. (2014) found crowd labels to sometimes not even be in agreement with themselves. Surprise is an emotion that is especially difficult to model (Buechel and Hahn, 2017; Schuff et al., 2017), less frequent (Oberländer and Klinger, 2018), and is sometimes divided into positive and negative surprise (Alm et al., 2005).

3 Datasets

3.1 In-domain Datasets

Lyrics are valuable for song emotion prediction and decent classification models can be generated solely on song lyrics. However, many lyrical datasets for song emotion classification are based on valence-arousal and lack emotions such as surprise or fear, which are important components of mood (Ekman and Friesen, 2003). In addition, there is a lack of large, high quality datasets capturing complex emotion in music.

A Novel Lyrics Dataset Annotated for Emotion Consequently, we created the Edmonds Dance dataset, a novel corpus of English song lyrics annotated for emotion from the reader's perspective. By searching a Spotify playlist consisting of 800 songs, both lyrical and instrumental, and collecting available lyrics from LyricFind, Genius, and MusixMatch (Lyr; Gen; Mus), we retrieved lyrics for 524 songs. We then labeled our dataset based on Plutchik's 8 core emotions of Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, and Trust (Plutchik, 2001). Table 1 depicts a subsection of the Edmonds Dance dataset, while the Appendix has more information on our labeling methods.

Mihalcea/Strapparava Dataset Reannotation In addition to the Edmonds Dance dataset, we also reannotated the dataset introduced in Mihalcea and Strapparava (2012), a multimodal corpus of songs that includes scalar annotations of both audio and lyrics for Ekman's six core emotions: Anger, Disgust, Fear, Joy, Sadness, and Surprise (Ekman, 1993). The original dataset was annotated from the songwriter's perspective and at a line level. We averaged these line-level lyrical annotations to achieve classifications at higher levels, thus gen-

Song	Artist	Lyrics	Emotion
I'm Done	MYNGA, Hechmann	You ruined my life What you said in your message that night Left me broken and bruised...	Anger, Disgust, Sadness
I Lived	OneRepublic	I'd like to teach the world to sing Hope when you take that jump You don't fear the fall...	Anticipation, Joy, Trust
Jubel	Klingande	Save me, save me, save me You think I don't laugh, oh Do things I can like, so Why are we losing time...	Fear, Sadness, Surprise, Trust

Table 1: Examples from the Edmonds Dance Dataset

	Mihalcea/Strapparava	Edmonds Dance
Songs	100	524
Lines	4976	22924
Words	109332	708985
Vocabulary	2233	6563

Table 2: Lyrical Dataset Basic Statistics

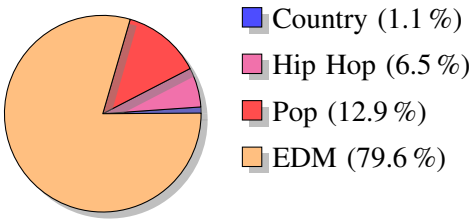

Figure 1: Genres within Edmonds Dance HITs

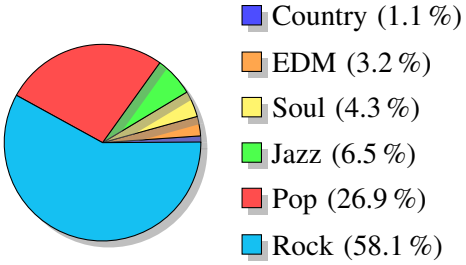

Figure 2: Genres within Mihalcea/Strapparava HITs

erating 452 verse-based and 100 song-based annotations. Table 2 provides some basic statistics for the lyrical datasets used in our research.

Mechanical Turk We submitted HITs on Mechanical Turk to validate lyric annotations. Each HIT contained three songs to be annotated from the reader's perspective for 8 emotions on a 6-point Likert scale. We also queried whether the annotator had heard of each song (yes/no), and whether they liked it (yes/no/unsure). Of the 186 songs annotated in total, 93 were from the Edmonds Dance dataset and 93 were from the Mihalcea/Strapparava dataset. HITs encompassed multiple genres, with the Edmonds Dance dataset mostly consisting of electronic music and the Mihalcea/Strapparava dataset mostly consisting of rock music. Figures 1 and 2 summarize HIT breakdowns by genre.

Annotation Guidelines To generate reliable annotations, our HIT included detailed annotation instructions. We organized these guidelines into four sections: initial instructions, important notes, definitions, and examples. The initial instructions section provided the annotator with basic task information, stating that he or she will be given a set of song lyrics, and is expected to record the degree to which the lyrics contain eight specific emotions. We also stated that emotions would be rated on a 6-point scale ranging from the complete absence of an emotion to the extreme presence of an emotion.

The important notes section emphasized that English speakers were required for the task, and that completion of all fields was required. The definitions section provided dictionary-level definitions for each of the eight emotions, while the examples section provided two annotated examples, along with general annotation guidelines (see Appendix A.2 for HIT images).

Each HIT contained the same two example songs. Each of the eight emotions was present in at least one of the songs, and emotions evoked by each song were apparent from the lyrics. Our HITs are available upon request.

Error Analysis We evaluated annotator reliability by calculating the average Cohen's Kappa of each annotator against others assigned to the same HIT, and discarding those below the threshold of 0.25. We then analyzed agreement across emotions by calculating Krippendorf's Alpha on the

remaining annotators, and examined the agreement between original and Turker annotations using Pearson's correlation coefficient. Table 3 depicts our results, with more details available in the Appendix.

Surprise had significantly lower inter-annotator agreement than other emotions. Krippendorf's Alpha and Pearson's Correlation values were lowest for Surprise, with significant correlation differences compared to all other emotions except Anticipation. Meanwhile, Joy and Sadness had relatively higher alpha and correlation values, suggesting a hierarchy of difficulty in emotion classification.

Emotion	Krippendorf's Alpha	Correlation±90 % CI
All	0.625	0.276±0.029
Anger	0.571	0.249±0.083
Anticipation*	0.572	0.294±0.149
Disgust	0.563	0.302±0.081
Fear	0.564	0.421±0.072
Joy	0.641	0.407±0.074
Sadness	0.63	0.411±0.072
Surprise	0.532	0.078±0.087
Trust*	0.617	0.384±0.138

Table 3: Inter-annotator Agreement at Cohen's Kappa Threshold of 0.25, and Pearson's Correlation between Original and Turker labels. Starred emotions were not present in the original Mihalcea/Strapparava dataset.

Analysis of Crowd Workers To confirm the quality of our dataset, we analyzed differences in annotation patterns between included and discarded Turkers. Discarded annotators had lower median completion time across the Edmonds Dance and Mihalcea/Strapparava datasets ($p<.005$), were more likely to say that they disliked a song ($p<.005$), and were less likely to say that they were unfamiliar with a song ($p<.001$). We also found that discarded annotators spent less time than included annotators on labeling songs that they disliked ($p<.001$). Further details are in the Appendix.

3.2 Out of Domain Datasets

To explore the efficacy of out-of-domain model training, we used the CBET (Shahraki, 2015), TEC (Mohammad, 2012), and DailyDialog (Li et al., 2017) datasets, three large collections of text annotated for multiple emotions including 6 core emotions present in both the Edmonds Dance and Mihalcea/Strapparava datasets. The CBET and TEC datasets respectively consist of 81,163 and 21,048 tweets, while the DailyDialog dataset consists of 102,979 statements collected from 13,118 transcripts of two-person conversations. Emotion distributions of the CBET, TEC, Daily Dialog, Edmonds Dance, and Mihalcea/Strapparava datasets are depicted in Table 4.

Emotion	CBET	TEC	DD	Dance	M/S
Anger	11.2%	7.4%	1.0%	13.7%	9.1%
Disgust	10.7%	3.6%	0.3%	21.9%	2.9%
Fear	11.2%	13.3%	0.2%	19.7%	1.8%
Joy	13.4%	39.1%	12.5%	43.9%	50.4%
Sadness	11.4%	18.2%	1.1%	35.3%	33.0%
Surprise	11.4%	18.3%	1.8%	13.0%	0.9%

Table 4: Presence of Emotion by Dataset

To train more robust baseline models, we also created augmented and transformed versions of the datasets; details on this process are available in the Appendix. While no versions of the CBET, TEC, and DailyDialog datasets include music lyrics, they are large enough to train deep models which we hypothesized could accurately predict emotions in smaller, gold-standard test datasets of song lyrics.

4 Model Implementation

We chose Naive Bayes as our first baseline emotion classification model due to its widespread applications in text classification and sentiment analysis (Raschka, 2014). Given its robustness to outliers and ability to deal with imbalanced data (Chen et al., 2004), a Random Forest baseline model was also implemented. Lastly, we utilized a Most Frequent Sense (MFS) baseline model, given its strong performance in word sense disambiguation tasks and its applications to emotion classification (Preiss et al., 2009). We trained our Naive Bayes model on bag-of-words features and our Random Forest model on transformed feature vectors which were generated from our textual datasets using the NRC Hashtag Emotion Lexicon (Mohammad and Turney, 2013); see Appendix for further details.

To improve upon emotion classification quality, we also explored more complex models. Due to its ability to generate powerful contextualized word embeddings and its state-of-the-art results in numerous language understanding tasks (Devlin et al., 2019), the $\text{BERT}_{\textbf{BASE}}$ uncased architecture was fine-tuned for multi-emotion classification from the text of song lyrics. $\text{BERT}_{\textbf{BASE}}$ consists of 12 Transformer blocks, a hidden size of 768, 12 self-attention heads, and an additional output layer

which we used for fine-tuning.[2]

5 Evaluation

We trained separate BERT models for each emotion on the original and augmented CBET datasets, and tested their performance on the Edmonds Dance and Mihalcea/Strapparava datasets. We then compared these results with those of our baseline Naive Bayes, Random Forest, and Most Frequent Sense models. To compare emotion prediction accuracy across multiple text corpora, we also trained BERT models on the TEC and DailyDialog datasets, and tested them on our lyrical datasets.

We found that BERT models trained on the CBET, TEC, and DailyDialog datasets did not generalize well to lyrical data. While models for joy and sadness improved upon the performance of baseline classifiers, models for disgust and fear performed worse than our Naive Bayes baseline. Furthermore, data augmentation techniques improved the performance of our baseline Naive Bayes model, but did not significantly increase BERT model accuracy.

To compare in-domain model accuracy with our out-of-domain results, we trained and tested BERT models on the Edmonds Dance and Mihalcea/Strapparava datasets, and vice versa. Models trained and tested on lyrical datasets had marginally better accuracy and F1 scores than out-of-domain models for anger, joy, and sadness. Given the much smaller sizes of lyrical datasets compared to their counterparts, as well as the differences in song genre and annotation perspective across lyrical datasets, our findings suggest a significant advantage in using in-domain data to train models for complex emotion classification of songs.

Finally, all models performed poorly when classifying surprise, and F1 scores for anger, disgust, and fear remained consistently low across models, suggesting a steep hierarchy of difficulty regarding emotion classification. Inter-annotator agreement was much lower for surprise than other emotions, and none of our models were able to accurately predict the presence of surprise in song lyrics. Our work implies that surprise is unique from the perspective of emotion classification.

Tables 5 and 6 highlight our model results. A complete version of our evaluation results is available in the Appendix.

[2]$\text{BERT}_{\text{BASE}}$ is available at https://tfhub.dev/google/bert_uncased_L-12_H-768_A-12/1.

Emotion	MFS	Naive Bayes	CBET BERT	Lyrics BERT
Anger	0.88	0.67	0.85	**0.88**
Disgust	0.88	0.79	0.88	0.88
Fear	0.89	0.77	0.89	0.89
Joy	0.47	0.53	0.58	**0.7**
Sadness	0.66	0.58	0.7	**0.73**
Surprise	0.92	0.92	0.92	0.92

Table 5: Model Accuracy on Lyrics By Emotion

Emotion	Naive Bayes	CBET BERT	Lyrics BERT
Anger	0.17	0.04	**0.2**
Disgust	**0.21**	0	0
Fear	**0.18**	0.14	0
Joy	0.03	0.24	**0.69**
Sadness	**0.55**	0.48	0.54
Surprise	0	0	0

Table 6: Model F1 Score on Lyrics By Emotion

6 Conclusion

In this paper we explore a variety of approaches to the multi-emotion classification problem for songs. We introduce the Edmonds Dance dataset, a novel lyrical dataset annotated for emotion at the song level and from the reader's perspective. We find that emotion classification of song lyrics using state-of-the-art methods is difficult to accomplish using out-of-domain data; BERT models trained on large corpora of tweets and dialogue do not generalize to lyrical data for emotions other than joy and sadness, and are outperformed by Naive Bayes classifiers on disgust and fear. On the other hand, models trained on song lyrics achieve comparable accuracy to models trained on out-of-domain data, even when lyrical datasets are orders of magnitude smaller than their counterparts, have been aggregated from line to song level, have been annotated from different perspectives, and are composed of different genres of music. Our findings underscore the importance of using in-domain data for song emotion classification.

7 Ethical Consideration

Our dataset was annotated by 184 Amazon Mechanical Turk crowdworkers. Annotators were paid \$0.15 per task or $\sim$ \$6.75 per hour, and reliable annotators (see Appendix A.2) were awarded a bonus of \$0.10 per task or $\sim$ \$11.25 per hour.

References

Genius website. https://genius.com/.

Lyricfind website. https://www.lyricfind.com/.

Musixmatch website. https://www.musixmatch.com/.

Jiddy Abdillah, Ibnu Asror, Yanuar Firdaus Arie Wibowo, et al. 2020. Emotion classification of song lyrics using bidirectional lstm method with glove word representation weighting. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 4(4):723–729.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 579–586.

Yunjing An, Shutao Sun, and Shujuan Wang. 2017. Naive bayes classifiers for music emotion classification based on lyrics. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pages 635–638. IEEE.

Victoria Bobicev and Marina Sokolova. 2017. Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *RANLP*, pages 97–102.

Sven Buechel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585.

Erion Çano and Maurizio Morisio. 2017. Moodylyrics: A sentiment annotated lyrics dataset. In *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, pages 118–124.

Chao Chen, Andy Liaw, Leo Breiman, et al. 2004. Using random forest to learn imbalanced data. *University of California, Berkeley*, 110(1-12):24.

Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. 2018. 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, Paris, France, September 23-27, 2018*, pages 370–375.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas A Diakopoulos and David A Shamma. 2010. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1195–1198.

Adam Dunbar, Charis E Kubrin, and Nicholas Scurich. 2016. The threatening nature of "rap" music. *Psychology, Public Policy, and Law*, 22(3):280.

Paul Ekman. 1993. Facial expression and emotion. *American psychologist*, 48(4):384.

Paul Ekman and Wallace V Friesen. 2003. *Unmasking the face: A guide to recognizing emotions from facial clues*. ISHK.

Carrie B Fried. 1999. Who's afraid of rap: Differential reactions to music lyrics 1. *Journal of Applied Social Psychology*, 29(4):705–721.

Maryam Hasan, Emmanuel Agu, and Elke Rundensteiner. 2014. Using hashtags as labels for supervised learning of emotions in twitter messages. In *ACM SIGKDD workshop on health informatics, New York, USA*.

Martin Haselmayer and Marcelo Jenny. 2017. Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. *Quality & quantity*, 51(6):2623–2646.

Hui He, Jianming Jin, Yuhong Xiong, Bo Chen, Wu Sun, and Ling Zhao. 2008. Language feature mining for music emotion classification via supervised learning from lyrics. In *International Symposium on Intelligence Computation and Applications*, pages 426–435. Springer.

Xiao Hu and J Stephen Downie. 2010. When lyrics outperform audio for music mood classification: A feature analysis. In *ISMIR*, pages 619–624.

Xiao Hu, J Stephen Downie, and Andreas F Ehmann. 2009. Lyric text mining in music mood classification. *American music*, 183(5,049):2–209.

Adit Jamdar, Jessica Abraham, Karishma Khanna, and Rahul Dubey. 2015. Emotion analysis of songs based on lyrical and audio features. *arXiv preprint arXiv:1506.05012*.

Robbert Jongeling, Proshanta Sarkar, Subhajit Datta, and Alexander Serebrenik. 2017. On negative results when using sentiment analysis tools for software engineering research. *Empirical Software Engineering*, 22(5):2543–2584.

Minho Kim and Hyuk-Chul Kwon. 2011. Lyrics-based emotion classification using feature selection by partial syntactic analysis. In *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, pages 960–964. IEEE.

Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. 2010. Music emotion recognition: A state of the art review. In *Proc. ismir*, volume 86, pages 937–952.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Gaojun Liu and Zhiyuan Tan. 2020. Research on multimodal music emotion classification based on audio and lyirc. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 1, pages 2331–2335. IEEE.

Rada Mihalcea and Carlo Strapparava. 2012. Lyrics, music, and emotions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 590–599.

Saif Mohammad. 2012. # emotional tweets. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada. Association for Computational Linguistics.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Laura Ana Maria Oberländer and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119.

Loreto Parisi, Simone Francia, Silvio Olivastri, and Maria Stella Tavella. 2019. Exploiting synchronized lyrics and vocal features for music emotion detection. *arXiv preprint arXiv:1901.04831*.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Judita Preiss, Jon Dehdari, Josh King, and Dennis Mehay. 2009. Refining the most frequent sense baseline. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 10–18.

Simon Provoost, Jeroen Ruwaard, Ward van Breda, Heleen Riper, and Tibor Bosse. 2019. Validating automated sentiment analysis of online cognitive behavioral therapy patient texts: An exploratory study. *Frontiers in psychology*, 10:1065.

Fika Hastarita Rachman, Riyanarto Sarno, and Chastine Fatichah. 2018. Music emotion classification based on lyrics-audio using corpus based emotion. *International Journal of Electrical & Computer Engineering (2088-8708)*, 8(3).

Meena Rambocas and Barney G Pacheco. 2018. Online sentiment analysis in marketing research: a review. *Journal of Research in Interactive Marketing*.

Sebastian Raschka. 2014. Naive bayes and text classification i-introduction and theory. *arXiv preprint arXiv:1410.5329*.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23.

Ameneh Gholipour Shahraki. 2015. Emotion detection from text. Master's thesis, University of Alberta.

Carlo Strapparava, Rada Mihalcea, and Alberto Battocchi. 2012. A parallel corpus of music and lyrics annotated with emotions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2343–2346, Istanbul, Turkey. European Language Resources Association (ELRA).

Marco Susino and Emery Schubert. 2019a. Cultural stereotyping of emotional responses to music genre. *Psychology of Music*, 47(3):342–357.

Marco Susino and Emery Schubert. 2019b. Negative emotion responses to heavy-metal and hip-hop music with positive lyrics. *Empirical Musicology Review*, 14(1-2):2–15.

Pyry Takala, Pekka Malo, Ankur Sinha, and Oskar Ahlgren. 2014. Gold-standard for topic-specific sentiment analysis of economic texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2152–2157, Reykjavik, Iceland. European Language Resources Association (ELRA).

Xing Wang, Xiaoou Chen, Deshun Yang, and Yuqian Wu. 2011. Music emotion classification of chinese songs based on lyrics using tf* idf and rhyme. In *ISMIR*, pages 765–770. Citeseer.

Dan Yang and Won-Sook Lee. 2009. Music emotion identification from lyrics. In *2009 11th IEEE International Symposium on Multimedia*, pages 624–629. IEEE.

Yi-Hsuan Yang, Yu-Ching Lin, Heng-Tze Cheng, I-Bin Liao, Yeh-Chin Ho, and Homer H Chen. 2008. Toward multi-modal music emotion classification. In *Pacific-Rim Conference on Multimedia*, pages 70–79. Springer.

A Appendix

A.1 Lyrical Datasets

The Mihalcea/Strapparava dataset initially consisted of 4976 lines across 100 songs which were annotated using a scale from 0 to 10, with 0 as the absence of emotion and 10 as the highest intensity of emotion. Annotations were based on Ekman's six core emotions: Anger, Disgust, Fear, Joy, Sadness, and Surprise (Ekman, 1993). As the dataset was annotated at a line level, we averaged emotion annotations on each line to achieve classifications at higher levels. Through averaging, we generated 452 verse-based and 100 song-based annotations.

With regards to the Edmonds Dance Dataset, the basis for label selection was provided by Plutchik's Theory of Emotion, which postulates that all emotions are combinations of the 8 core emotions present in our label (Plutchik, 2001). As a result, the label can lead to additional classification models for emotions which are theorized to be dyads of the core emotions (e.g, PLove = PJoy $*$ PTrust, or PAggressiveness = PAnger $*$ PAnticipation). Our dataset was initially labeled using an array of size 8; each array index contained a binary value to indicate an emotion's presence.

Dataset	Emotion	Krippendorf's Alpha
Dance	All	0.717
Dance	Anger	0.704
Dance	Anticipation	0.706
Dance	Disgust	0.691
Dance	Fear	0.69
Dance	Joy	0.718
Dance	Sadness	0.73
Dance	Surprise	0.653
Dance	Trust	0.705
M/S	All	0.532
M/S	Anger	0.438
M/S	Anticipation	0.437
M/S	Disgust	0.435
M/S	Fear	0.437
M/S	Joy	0.564
M/S	Sadness	0.53
M/S	Surprise	0.411
M/S	Trust	0.528

Table A1: Interannotator Agreement at Cohen's Kappa Threshold of 0.25

Dataset	Emotion	Pearson's Correlation	90% CI
Dance	All	0.396	(0.343, 0.445)
Dance	Anger	0.204	(0.033, 0.363)
Dance	Anticipation	0.294	(0.129, 0.443)
Dance	Disgust	0.429	(0.278, 0.559)
Dance	Fear	0.31	(0.146, 0.457)
Dance	Joy	0.362	(0.203, 0.502)
Dance	Sadness	0.316	(0.154, 0.462)
Dance	Surprise	0.175	(0.003, 0.336)
Dance	Trust	0.384	(0.228, 0.522)
M/S	All	0.183	(0.124, 0.241)
M/S	Anger	0.28	(0.114, 0.431)
M/S	Disgust	0.214	(0.045, 0.371)
M/S	Fear	0.499	(0.358, 0.618)
M/S	Joy	0.439	(0.289, 0.568)
M/S	Sadness	0.477	(0.333, 0.6)
M/S	Surprise	0.01	(-0.161, 0.18)

Table A2: Pearson's Correlation between Original and Turker annotations

A.2 Annotator Error Analysis

To evaluate the reliability of our Mechanical Turk annotations, we first used Cohen's Kappa to calculate the average inter-annotator agreement of each Turker against others assigned to the same HIT. We then discarded all annotators who failed to meet a threshold of 0.25, and calculated average agreement for each emotion using Krippendorf's Alpha on the remaining annotators. Krippendorf's Alpha values were highest for the emotions of joy, sadness, and trust; additionally, alpha values were relatively consistent across emotions. 31.6% of annotations in Mihalcea and Strapparava's dataset failed to meet the Cohen's Kappa threshold, while 63.2% of annotations in the Edmonds Dance dataset failed to meet the threshold. Our results are summarized in Table A1, while Figures A1, A2, and A3 depict pictures of our HITs.

Next, we calculated the Pearson's correlation coefficient and related p-values between original annotations and the Turker annotations for both the Edmonds Dance and Mihalcea/Strapparava datasets. We were also unable to calculate correlation coefficients for Anticipation or Trust in the Rada dataset as the original dataset did not include annotations for these emotions. These results are summarized in Table A2.

While the relative strength of Pearson's Correlations across emotions was similar to that of our alpha values, correlation with fear was relatively higher than expected, and correlation with anger and surprise were lower than expected. Finally, we looked at Krippendorf's Alpha Values on an an-

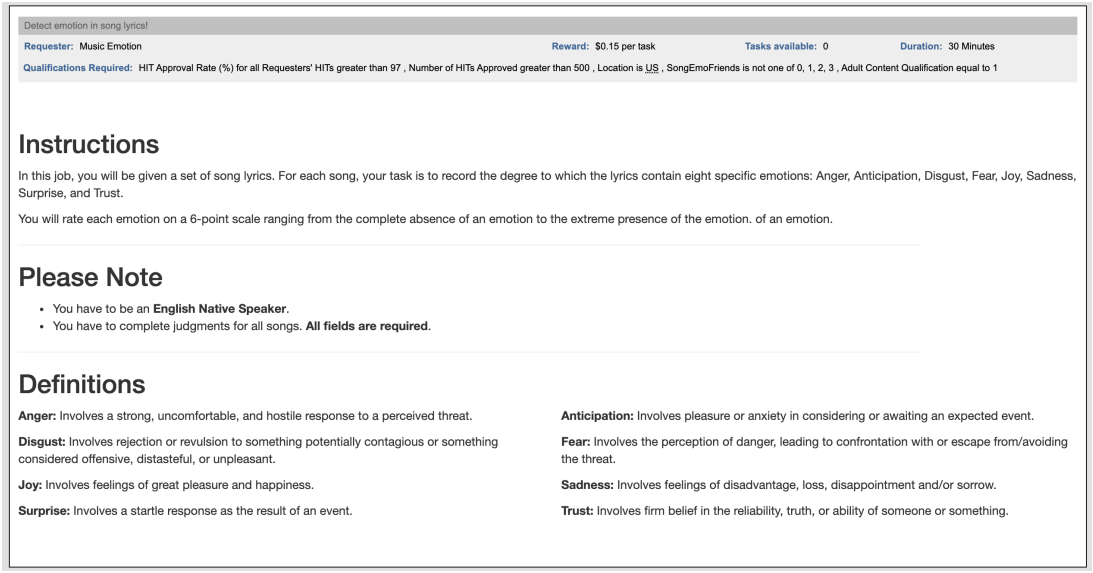

Figure A1: HIT Preliminary Instructions

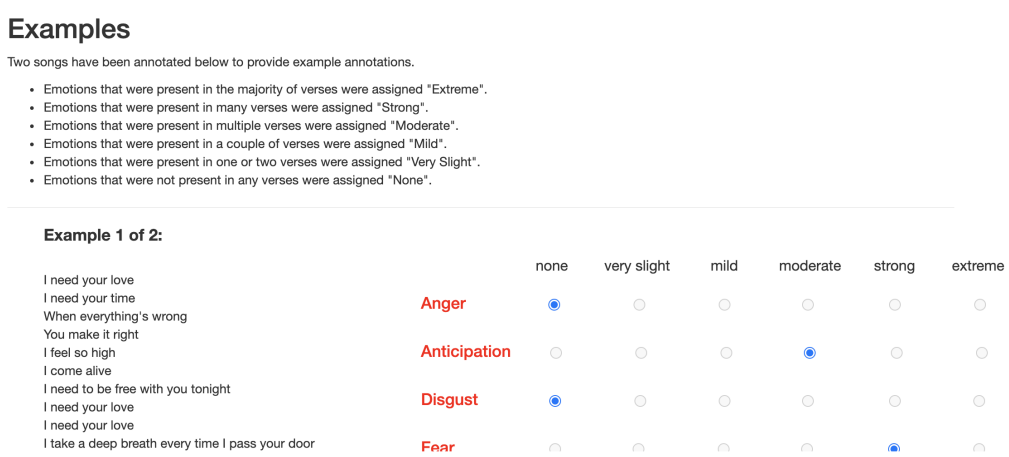

Figure A2: HIT Example Annotations

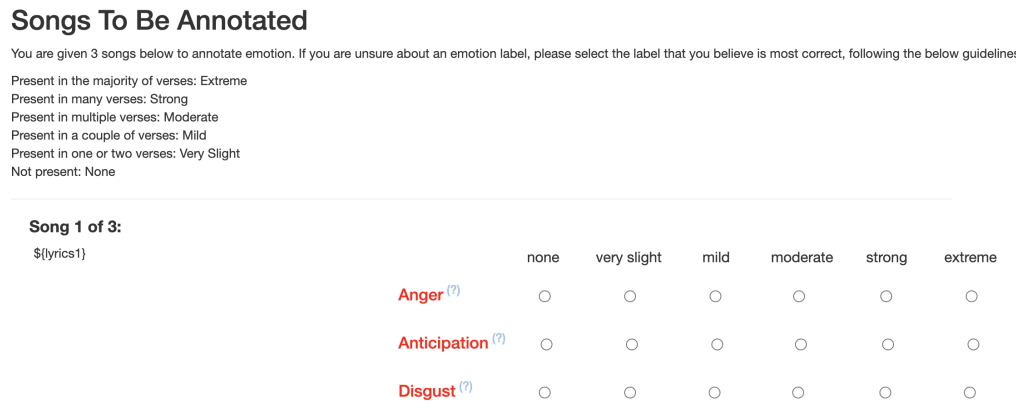

Figure A3: HIT Annotation Format

notation group level to better understand whether annotation agreement for specific emotions were consistently similar across songs. Our results, summarized in Table A3, provide evidence for a hierarchy of difficulty in emotion classification. Joy and Sadness have the most favorable distribution of alpha values with few low item-level alpha scores (<0.2), and greater numbers of medium (0.2-0.6) and high (>0.6) item-level alpha scores. Anger and Trust have the next most favorable distributions, while Anticipation, Disgust, and Fear have similar but lower agreement distributions. Finally, Surprise has the worst distribution, with almost all item-level Krippendorf's Alpha values being classified as low agreement.

Emotion	Agreement		
	Low	Moderate	High
Anger	25	17	8
Anticipation	27	15	3
Disgust	26	21	3
Fear	30	18	2
Joy	12	24	14
Sadness	10	23	17
Surprise	48	2	0
Trust	17	25	8

Table A3: Krippendorf's Alpha Values By Emotion Over 50 Annotator Groups

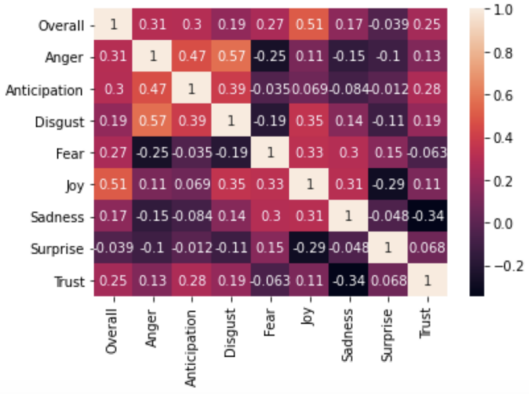

Figure A4: Correlation Heatmap of Krippendorf's Alphas

We then created a heat map of item-level Krippendorf's Alphas to explore correlation of interannotator agreement across emotions. Our results, visualized in Figure A4, reveal that alpha values are only slightly correlated across emotions. This implies that classification difficulty of a specific emotion varies depending on the song being annotated; indeed, the only emotions that have an inter-annotator agreement correlation above 0.4 are Anger/Disgust and Anger/Anticipation. We can also see that only joy has a moderate correlation with overall agreement across emotions, implying that songs with annotation agreement regarding joy may be easier to classify overall, but songs with annotation agreement regarding other emotions may not necessarily be easier to annotate. Consequently, the claim of a consistent hierarchy of difficulty is somewhat undermined and instead it seems that classification difficulty of a specific emotion varies depending on the song being annotated.

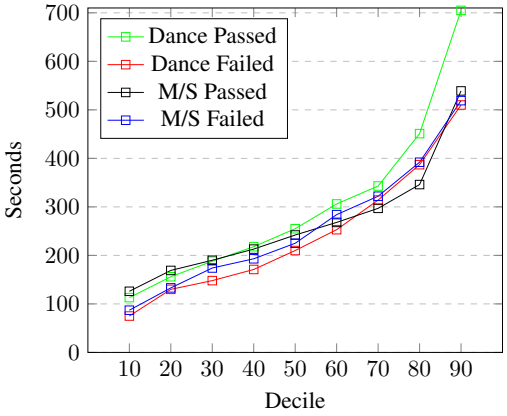

Figure A5: Annotation Completion Time by Quantile

A.3 Analysis of Crowd Workers

We analyzed the completion time of annotations across good and bad annotators for the Edmonds Dance and Mihalcea/Strapparava datasets, summarized in Figure A5. We can see that the distributions of completion times were very similar for bad annotators, while the distributions for good annotators were skewed upwards at higher deciles. In addition, the median completion time for good annotators was 31 seconds greater than the median completion time for bad annotators, and the mean completion time for good annotators was 37 seconds greater than that of bad annotators.

Next, we looked at differences between good and bad annotator groups regarding annotator enjoyment and familiarity of labeled songs. We found that bad annotators were more likely than good annotators to say that they were familiar with a song ($p<.00001$), or that they disliked a song ($p<.005$). Bad annotators also spent significantly less time than good annotators on labeling songs that they said they disliked ($p<0.0001$). These results are summarized in Figures A6 and A7.

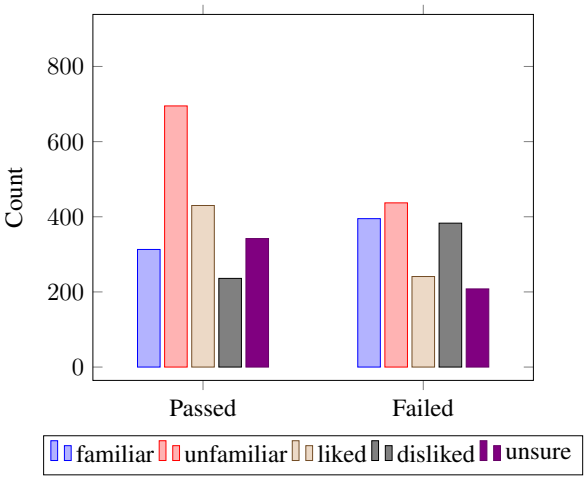

Figure A6: Annotator Enjoyment/Familiarity of Songs by Count and Annotation Quality

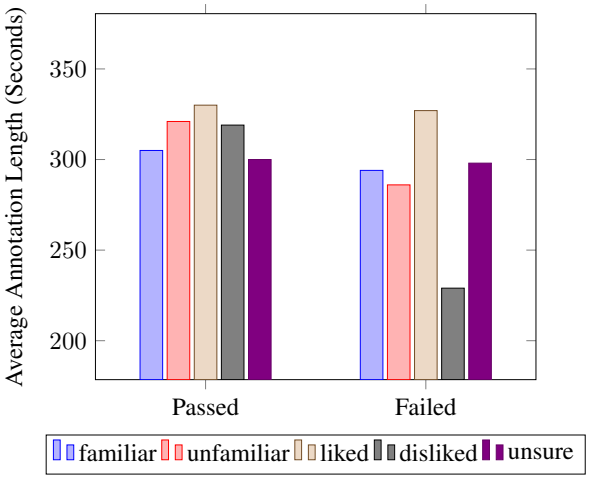

Figure A7: Annotator Enjoyment/Familiarity of Songs by Annotation Time and Quality

A.4 Additional Dataset Information

A.4.1 Data Augmentation and Transformation

To address misclassification of the minority class, we implemented oversampling techniques so classes would be more balanced. For each emotion in the CBET dataset, we added copies of tweets suggesting the presence of the emotion such that the new number of tweets with presence of emotion was between 40-60 percent of the total number of tweets. We then trained two BERT models for each emotion, one using the augmented CBET dataset and the other using the original. To confirm the quality of the original CBET dataset, we also trained and tested BERT models on subsets of CBET data that were randomly sampled without replacement. These results are shown in Table A4.

To explore another approach to the multi-emotion classification problem, lyrical data was then transformed into a feature vector of length 9 using the NRC Hashtag Emotion Lexicon, which contains binary indicators regarding the presence or absence of Plutchik's 8 core emotions in 14182 common English words (Mohammad and Turney, 2013). This occurred by iterating through a song's lyrics, counting each word present in the NRC Emotion Lexicon as well as its emotional classification, and storing this information in the feature vector. For example, the feature vector [5, 10, 1, 9, 4, 2, 2, 3, 28] would correspond to a song's lyrics that contained 28 words (not necessarily all distinct) which were present in the NRC Emotion Lexicon. Of these words, 5 were associated with joy, 10 with trust, 1 with fear, etc. This transformed dataset was generated for the purpose of exploring Random Forest methods for song emotion classification.

A.5 Evaluation

To gauge the quality of the CBET dataset, we first calculated the accuracies of BERT models trained and tested on randomly ordered subsets of CBET data, with an 80/20 train/test split. Emotion classification accuracies of these models were at least 90%, confirming the quality of the dataset. Next, we trained BERT models on the full CBET datasets, and evaluated them on the verse-based variation of Mihalcea and Strapparava's dataset, as well as the Edmonds Dance dataset. All BERT models were trained for 3 epochs, and used a sequence length of 128, batch size of 32, learning rate of $2e^{-5}$, and warmup proportion of 0.1. The performance of these models, depicted in Table A4, were then compared to the performance of baseline Naive Bayes and Random Forest models, shown in Tables A5 and A6. Only the baseline Naive Bayes model trained on augmented CBET data is depicted in Table A5, as the Naive Bayes model trained on normal CBET data had precision and recall of zero for each emotion.

It can be seen from Table A4 that BERT models trained on CBET did not generalize well to lyrical datasets. While models for joy and sadness improved on the performance of Naive Bayes and Random Forest classifiers, models for other emotions did not significantly improve on the baseline, and in some cases performed worse than baseline classifiers. BERT models for anger and fear had lower precision and recall than corresponding

Emotion	Train	Test	Accuracy	AUC	Prec	Rec
Anger	CBET sub	CBET sub	0.91	0.73	0.6	0.49
Anger	CBET	Dance	0.84	0.51	0.19	0.04
Anger	CBET aug	Dance	0.81	0.53	0.2	0.14
Anger	CBET	Dance Turk	0.81	0.5	0	0
Anger	CBET	M/S	0.85	0.05	0.1	0.07
Anger	CBET aug	M/S	0.90	0.55	0.42	0.12
Anger	CBET	M/S Turk	0.89	0.54	0.25	0.125
Disgust	CBET sub	CBET sub	0.93	0.79	0.7	0.61
Disgust	CBET	Dance	0.78	0.5	0	0
Disgust	CBET aug	Dance	0.78	0.5	0	0
Disgust	CBET	Dance Turk	0.94	0.49	0	0
Disgust	CBET	M/S	0.97	0.5	0	0
Disgust	CBET aug	M/S	0.97	0.5	0	0
Disgust	CBET	M/S Turk	0.96	0.5	0	0
Fear	CBET sub	CBET sub	0.95	0.86	0.82	0.74
Fear	CBET	Dance	0.77	0.55	0.34	0.19
Fear	CBET aug	Dance	0.8	0.52	0.5	0.06
Fear	CBET	Dance Turk	0.78	0.53	0.29	0.12
Fear	CBET	M/S	0.97	0.49	0	0
Fear	CBET aug	M/S	0.97	0.55	0.13	0.13
Fear	CBET	M/S Turk	0.87	0.53	0.33	0.09
Joy	CBET sub	CBET sub	0.9	0.76	0.67	0.57
Joy	CBET	Dance	0.61	0.56	0.82	0.16
Joy	CBET aug	Dance	0.58	0.53	0.74	0.07
Joy	CBET	Dance Turk	0.45	0.5	0	0
Joy	CBET	M/S	0.54	0.54	0.95	0.09
Joy	CBET aug	M/S	0.52	0.53	1	0.06
Joy	CBET	M/S Turk	0.56	0.52	1	0.05
Sadness	CBET sub	CBET sub	0.9	0.7	0.59	0.45
Sadness	CBET	Dance	0.7	0.64	0.6	0.45
Sadness	CBET aug	Dance	0.7	0.63	0.6	0.42
Sadness	CBET	Dance Turk	0.68	0.6	0.63	0.29
Sadness	CBET	M/S	0.69	0.59	0.57	0.03
Sadness	CBET aug	M/S	0.7	0.55	0.82	0.12
Sadness	CBET	M/S Turk	0.68	0.67	0.76	0.49
Surprise	CBET sub	CBET sub	0.92	0.78	0.66	0.6
Surprise	CBET	Dance	0.87	0.5	0	0
Surprise	CBET aug	Dance	0.87	0.5	0	0
Surprise	CBET	Dance Turk	0.88	0.5	0	0
Surprise	CBET	M/S	0.99	0.5	0	0
Surprise	CBET aug	M/S	0.99	0.5	0	0
Surprise	CBET	M/S Turk	0.9	0.49	0	0

Table A4: BERT Trained on CBET Variations

Emotion	Test	Accuracy	Precision	Recall
Anger	Dance Original	0.65	0.14	0.29
Anger	Dance Turk	0.63	0.14	0.31
Anger	M/S Original	0.7	0.12	0.34
Anger	M/S Turk	0.84	0.23	0.38
Disgust	Dance Original	0.77	0.43	0.2
Disgust	Dance Turk	0.84	0	0
Disgust	M/S Original	0.82	0.04	0.23
Disgust	M/S Turk	0.92	0.29	0.5
Fear	Dance Original	0.7	0.27	0.3
Fear	Dance Turk	0.73	0.35	0.53
Fear	M/S Original	0.84	0.03	0.25
Fear	M/S Turk	0.77	0.14	0.18
Joy	Dance Original	0.57	0.8	0.02
Joy	Dance Turk	0.45	0	0
Joy	M/S Original	0.5	0.67	0.02
Joy	M/S Turk	0.54	0	0
Sadness	Dance Original	0.61	0.47	0.77
Sadness	Dance Turk	0.53	0.42	0.79
Sadness	M/S Original	0.55	0.4	0.77
Sadness	M/S Turk	0.61	0.58	0.71
Surprise	Dance Original	0.86	0	0
Surprise	Dance Turk	0.87	0	0
Surprise	M/S Original	0.97	0	0
Surprise	M/S Turk	0.91	0	0

Table A5: Naive Bayes Trained on Augmented CBET

Emotion	Test	Accuracy	Precision	Recall
Anger	Dance Original	0.85	0.1	0.01
Anger	Dance Turk	0.84	0	0
Anger	M/S Original	0.86	0.04	0.02
Anger	M/S Turk	0.87	0	0
Disgust	Dance Original	0.78	0	0
Disgust	Dance Turk	0.95	0	0
Disgust	M/S Original	0.96	0	0
Disgust	M/S Turk	0.96	0	0
Fear	Dance Original	0.74	0.26	0.17
Fear	Dance Turk	0.76	0.27	0.18
Fear	M/S Original	0.93	0	0
Fear	M/S Turk	0.81	0.11	0.09
Joy	Dance Original	0.58	0.61	0.12
Joy	Dance Turk	0.48	0.71	0.1
Joy	M/S Original	0.53	0.73	0.11
Joy	M/S Turk	0.56	0.67	0.09
Sadness	Dance Original	0.6	0.36	0.18
Sadness	Dance Turk	0.65	0.54	0.21
Sadness	M/S Original	0.66	0.36	0.03
Sadness	M/S Turk	0.54	0.58	0.16
Surprise	Dance Original	0.87	0	0
Surprise	Dance Turk	0.86	0	0
Surprise	M/S Original	0.98	0	0
Surprise	M/S Turk	0.91	0	0

Table A6: Random Forest Trained on Transformed CBET

Naive Bayes and Random Forest models, while models for surprise were more or less equivalent

Emotion	Train	Test	Accuracy	AUC	Prec	Rec
Anger	TEC	Dance	0.85	0.58	0.39	0.22
Anger	TEC	M/S	0.89	0.55	0.29	0.15
Anger	TEC	Dance Turk	0.78	0.49	0.11	0.08
Anger	TEC	M/S Turk	0.92	0.56	1	0.13
Disgust	TEC	Dance	0.78	0.5	0	0
Disgust	TEC	M/S	0.97	0.5	0	0
Disgust	TEC	Dance Turk	0.95	0.5	0	0
Disgust	TEC	M/S Turk	0.96	0.5	0	0
Fear	TEC	Dance	0.74	0.58	0.33	0.32
Fear	TEC	M/S	0.9	0.76	0.1	0.63
Fear	TEC	Dance Turk	0.77	0.54	0.3	0.18
Fear	TEC	M/S Turk	0.8	0.61	0.27	0.36
Joy	TEC	Dance	0.67	0.66	0.65	0.54
Joy	TEC	M/S	0.66	0.66	0.68	0.62
Joy	TEC	Dance Turk	0.68	0.68	0.74	0.63
Joy	TEC	M/S Turk	0.66	0.66	0.62	0.65
Sadness	TEC	Dance	0.7	0.61	0.7	0.28
Sadness	TEC	M/S	0.73	0.6	0.83	0.23
Sadness	TEC	Dance Turk	0.75	0.69	0.76	0.47
Sadness	TEC	M/S Turk	0.51	0.49	0.46	0.13
Surprise	TEC	Dance	0.85	0.49	0	0
Surprise	TEC	M/S	0.97	0.49	0	0
Surprise	TEC	Dance Turk	0.88	0.5	0	0
Surprise	TEC	M/S Turk	0.9	0.49	0	0

Table A7: BERT Trained on TEC

Emotion	Train	Test	Accuracy	AUC	Prec	Rec
Anger	DD	Dance	0.76	0.65	0.28	0.51
Anger	DD	M/S	0.9	0.69	0.43	0.44
Anger	DD	Dance Turk	0.86	0.63	0.5	0.31
Anger	DD	M/S Turk	0.9	0.61	0.4	0.25
Disgust	DD	Dance	0.78	0.5	0	0
Disgust	DD	M/S	0.97	0.5	0	0
Disgust	DD	Dance Turk	0.95	0.5	0	0
Disgust	DD	M/S Turk	0.96	0.5	0	0
Fear	DD	Dance	0.8	0.5	0.5	0
Fear	DD	M/S	0.98	0.68	0.33	0.38
Fear	DD	Dance Turk	0.82	0.5	0	0
Fear	DD	M/S Turk	0.88	0.5	0	0
Joy	DD	Dance	0.61	0.57	0.72	0.2
Joy	DD	M/S	0.55	0.56	0.72	0.18
Joy	DD	Dance Turk	0.53	0.57	1	0.14
Joy	DD	M/S Turk	0.61	0.58	0.1	0.16
Sadness	DD	Dance	0.66	0.51	0.86	0.03
Sadness	DD	M/S	0.67	0.51	1	0.01
Sadness	DD	Dance Turk	0.66	0.53	1	0.06
Sadness	DD	M/S Turk	0.51	0.49	0	0
Surprise	DD	Dance	0.87	0.51	0.33	0.03
Surprise	DD	M/S	0.99	0.5	0	0
Surprise	DD	Dance Turk	0.88	0.5	0	0
Surprise	DD	M/S Turk	0.91	0.5	0	0

Table A8: BERT Trained on DailyDialog

to the baseline. Additionally, BERT and Random Forest models were unable to correctly identify disgust, while Naive Bayes models successfully identified multiple instances of disgust. As there was not a significant difference in balance between emotion classes within the CBET dataset, the fact that data augmentation did not significantly improve baseline precision and recall implies that class imbalance was not a main factor in discrepancies between classification accuracy of different emotions.

To compare emotion prediction accuracies across multiple text corpora, we then trained BERT models on the TEC and DailyDialog datasets, and tested them on the Edmonds Dance and Mihalcea/Strapparava datasets. The results are summarized in Tables A7 and A8. Both test accuracy for the TEC and the DailyDialog models were similar to those of the CBET models, implying that the dialog domain does not necessarily show more promise than the social media domain when considering the complex emotion classification problem in lyrics.

Finally, to compare in-domain model accuracy with our out of domain results, we trained and tested BERT models on the larger, original versions of the Edmonds Dance and Mihalcea/Strapparava datasets respectively, and vice versa. The results are summarized below in Table A9. We found that the accuracies of models trained and tested on the Edmonds Dance and Mihalcea/Strapparava datasets were on par with those of the out of domain models despite the much smaller training size and genre differences across the lyrical datasets, implying a significant advantage in using in-domain data to train models for complex emotion classification of songs.

It is important to note that precision and recall values for disgust, fear, and surprise remained very low, which could imply that certain emotions are generally more difficult than others to classify. This conclusion is supported by our Turker error analysis in Section 3.1.3, in which we found that emotions such as anticipation, disgust, fear and surprise had relatively lower inter-annotator agreement, while other emotions such as joy and sadness had relatively high agreement.

A.6 Miscellaneous: Emotion Magnitudes by Line

Mihalcea and Strapparava included a table in their paper with the number of lines that each of their 6 core emotions was present in, as well as the average magnitude for each emotion across all annotated lines. We used this information to calculate the average magnitude for each emotion across lines in which they were present, shown in Table A10.

As emotions were annotated on a scale from 0

Emotion	Train	Test	Accuracy	AUC	Prec	Rec
Anger	M/S	Dance	0.86	0.51	0.5	0.01
Anger	Dance	M/S	0.89	0.58	0.31	0.2
Disgust	M/S	Dance	0.78	0.5	0	0
Disgust	Dance	M/S	0.97	0.5	0	0
Fear	M/S	Dance	0.8	0.5	0	0
Fear	Dance	M/S	0.98	0.5	0	0
Joy	M/S	Dance	0.7	0.69	0.67	0.62
Joy	Dance	M/S	0.7	0.7	0.74	0.62
Sadness	M/S	Dance	0.72	0.66	0.64	0.46
Sadness	Dance	M/S	0.73	0.65	0.66	0.4
Surprise	M/S	Dance	0.87	0.5	0	0
Surprise	Dance	M/S	0.99	0.5	0	0

Table A9: BERT Trained and Tested on Lyrical Datasets

Emotion	Average Magnitude
Anger	1.88
Disgust	1.44
Fear	1.41
Joy	4.14
Sadness	2.94
Surprise	1.39

Table A10: Average Emotion Magnitude per Line in Mihalcea/Strapparava Dataset

to 10, we found it worthwhile to note that annotations for the presence of negative emotions such as anger, disgust, and fear were more likely to be mild than strong. We also found it interesting that only joy had an average magnitude greater than 3, which represented the cutoff for the presence of an emotion (Mihalcea and Strapparava, 2012).