Emotion Sentiment Analysis in Turkish Music

Nguyen Nguyen
Dept. of Computer Science
University of Wisconsin – Eau Claire
Eau Claire, Wisconsin, USA
nguyennt2507@uwec.edu

Naeem Seliya
Dept. of Computer Science
University of Wisconsin – Eau Claire
Eau Claire, Wisconsin, USA
seliyana@uwec.edu

***Abstract*—Music has become an indispensable element of human life. Its rhythms, melodies, and harmonies resonate deeply within us, touching our emotions and echoing our sentiments. In recent years, music emotion sentiment classifications in different languages have been studied. However, to be best of our knowledge, Turkish music has not been explored sufficiently using intelligent tools. We explore machine learning algorithms to classify Turkish music audio excerpts into distinct mood categories. We use two datasets: the Turkish Music Emotion (TME) dataset and the Turkish Emotional Voice Database (TurEV-DB) dataset. The Gradient Boosting, XGBoost, CatBoost, Random Forest, Decision Tree, and Gaussian Naïve Bayes machine learning algorithms are used for training and testing the learners. Our case study results demonstrate that the CatBoost learner has the best overall performance with an Area Under the ROC Curve of 0.948 and Accuracy of about 82% for the TME dataset, and an Area Under the ROC Curve of 0.989 and Accuracy of about 90% for the TurEV-DB dataset. In the context of the two datasets, the top two learners are CatBoost followed by XGBoost. The six learners, to the best of our knowledge, have not been explored elsewhere with these two datasets, making this work a unique addition to the related literature and state-of-the-art.**

Keywords—*Mood Classification, Sentiment Analysis, Turkish Music, Human Emotion, Machine Learning*.

I. Introduction

Music is a large part of human nature, as evidenced by the observation that every known culture has some form of music [1]. Music, with its diverse forms and cultural expressions, has the remarkable ability to evoke a wide spectrum of emotions, ranging from joy and serenity to melancholy and agitation [13]. Emotional classification of music is a multidisciplinary field of research [15]. The power to evoke emotions makes music a highly valuable tool for the investigation of emotions. In fact, given the ubiquity of music across cultures, and thus most probably throughout human history, our understanding of emotions would remain incomplete without a proper understanding of music-evoked emotions [1]. However, determining the emotional category (i.e., mood) of music using machine learning is a challenging problem, and several issues need to be addressed such as emotion labeling of music excerpts, feature extraction, and choice of the classification algorithm [2].

One of the primary reasons for this problem is that at its core, emotion perception is subjective, meaning that different people may experience different feelings when listening to the same song [14]. This subjectivity issue makes the performance evaluation of a music emotion identification system highly difficult. Another reason is that it is not easy to define emotions universally, because the attributes that are used to define the same emotion may vary among people. Finally, it is still unknown how music arouses emotions in humans, and it has not been understood fully how the inner element of music creates emotional reactions in the listener [10].

Turkish music, with its rich tapestry of melodies, rhythms, and cultural influences, stands as a testament to the country's vibrant musical heritage. While research on music emotion sentiment analysis has made some headway in various cultural contexts, there remains a notable gap in our understanding of Turkish music's emotional dimensions and its impact on the listeners of Turkish music.

In this paper, we focus on using categorical models, also known as discrete models, to use single words or phrases to classify emotion categories. This research builds upon these existing works by investigating music emotion analysis in Turkish music using two datasets: Turkish Music Emotion (TME) [3] and Turkish Emotional Voice Database (TurEV-DB) [4]. The TME dataset, comprising 400 samples, consists of 100 samples each representing happy, sad, angry, and relaxed emotional states. This dataset provides a balanced representation of various emotional categories within Turkish music, enabling a more thorough analysis and evaluation. Additionally, the TurEV-DB offers a substantial collection consisting of 1,735 vocal recordings. While the TME dataset offers a diverse collection of Turkish music samples spanning different genres, the TurEV-DB provides a rich repository of vocal recordings, enabling a comprehensive examination of emotional expressions in Turkish auditory sounds. We aim to contribute to the state-of-the-art by exploring the effectiveness of different machine learning models on these datasets and potentially identify avenues for further improvement in music emotion recognition for Turkish music.

Machine learning models can differentiate emotion sentiment based on the extracted features. These models can aid listeners in recommending similar music and audio pieces that they like to listen to. In this paper, we investigate six different machine learning models, including Gradient Boosting, XGBoost, CatBoost, Decision Tree, Random Forest, and Gaussian Naïve Bayes. Our empirical analysis yielded performance accuracies of the machine learning models as high as 95.2% for an individual emotion sentiment category.

2835-5776/24/$31.00 ©2024 IEEE
DOI 10.1109/IRI62200.2024.00024

The rest of the paper is structured as follows: Section II presents relevant related work; Section III discussed data preparation and preprocessing; Section IV summarizes the machine learners; Section IV presents and discussed empirical results; and Section VI concludes our work with a summary.

II. Related Literature

Automatic emotion recognition in music has attracted research interest in recent years [11]. This section explores existing approaches relevant to music emotion analysis in Turkish, focusing on the two datasets. Research on music emotion recognition in Turkish is an active investigation area.

Several studies have explored music emotion recognition using audio features. Hizlisoy et al. [2] proposed a Convolutional Long Short-Term Memory and Deep Neural Network architecture with correlation-based feature selection for music emotion recognition, achieving promising results, with 99.19% accuracy. Bilal Er and Aydilek [3] employed chroma spectrograms and deep visual features for music emotion classification. The best performance before data augmentation was obtained from the "Fc6" layer of the VGG-16 with SoftMax classifier with 76% accuracy; after data augmentation, the best classifier success was obtained from the "Fc7" layer of the VGG-16 with the SVM classifier with 89.2% accuracy.

While audio features offer valuable information, some research works incorporate lyrics for improved performance. Durahim et al. [5] investigated music emotion classification for Turkish songs using lyrics. They achieved a recall of 43.7% and a precision of 46.9% when using the Zemberek Long stemming method with the Multinomial Naïve Bayes classifier. This approach aligns with the growing interest in sentiment analysis of textual data, as explored in the broader sentiment analysis in the music domain by Shukla et al. [6]. The best performer is the combination of multi-type lyrics features with an accuracy of 63.8%. Biancofiore et al. [8] proposed an aspect-based sentiment analysis approach for music, highlighting the potential for emotion recognition beyond basic categories.

Canpolat et al. [4] introduced the Turkish Emotion Voice Database (TurEV-DB), a valuable resource for research on emotion recognition in Turkish speech. In addition to detailing the development of the TurEV-DB dataset, the authors performed mood classification using the SVM and CNN learners and compared their performances with that of human judgements. In contrast, our study focuses on investigating the dataset with six previously unexamined machine learners.

Our work builds upon these works by focusing on the consistency of performance across various machine learning models. We aim to achieve not only competitive accuracy but also demonstrate consistent performance across multiple trials using well-established Turkish music datasets. This approach will provide a more robust understanding of how well these models perform on unseen data and potentially identify models that are more suitable for real-world applications. To the best of our knowledge the six machine learners investigated in this paper have not been explored elsewhere with the TME and TurEV-DB datasets, making our contributions unique to the related literature and state-of-the-art.

III. Methodology

A. Case Study Data

In our study, we used two datasets to evaluate six machine learners. The first is the TME dataset [3], which comprises of four classes: happy, sad, angry, and relaxed. Verbal and nonverbal music from various Turkish music genres is chosen to prepare the dataset. To ensure that every class has an equal number of samples, a total of 100 musical pieces have been selected for each class. The TME dataset has 400 samples, with a 30-second interval with 50 different acoustic features, such as Mel Frequency Cepstral Coefficients, Temp, Chromagram, Spectral, and Harmonic, for each sample.

The second dataset is the TurEV-DB which involves a corpus of over 1735 tokens based on 82 words uttered by human subjects in four emotional states: angry, calm, happy, and sad. Within this database, emotional states are distributed across 487 samples of anger, 357 samples of happiness, 408 samples of calmness, and 483 samples of sadness. Each sample contains 1582 statistical features extracted using OpenSmile and IS10_paralig configuration [4].

B. Data Preparation and Empirical Settings

This section details the process undertaken to prepare the TME and TurEV-DB datasets for machine learning analysis and outlines the chosen empirical settings. A well-prepared dataset is important for effective model training and evaluation. The following steps were followed in our empirical study and analysis of the two datasets with the six machine learners.

- **Data Cleaning:** The initial step involved cleaning the dataset to identify and remove any duplicate entries or missing values. This step maintains data integrity and consistency by eliminating irrelevant information that could potentially affect the models' performance.
- **Feature Normalization:** Following cleaning, the data underwent normalization using StandardScaler from the scikit-learn library [9]. Normalization ensures all features contribute equally during training by scaling them to have a common mean and standard deviation. This mitigates biases introduced by inherent differences in the scale of various features.
- **Class Labeling:** After normalization, emotional categories in textual form {happy, sad, angry, relax} were converted into numerical labels {0, 1, 2, 3} using the Label Encoder from sklearn [9]. This step enables machine learning algorithms to effectively process the numeric class labels during their training process.
- **Data Splitting:** To establish the training and test datasets, the preprocessed data was randomly shuffled and split using a 70-30 ratio. This split allows for a substantial portion of the data to be used to train the models while reserving a separate set for evaluating their performances on unseen (during training) data.
- **Enhancing Robustness:** To strengthen the robustness and reliability of the results, the entire data preparation and splitting process was repeated five times. Each iteration involved random shuffling and splitting of the dataset, resulting in five distinct training and test datasets for model

performance evaluation. This approach helps account for potential biases introduced by using only a single random data split.

For additional and more specific details about the two datasets, including audio data translation and feature engineering, we refer the readers to the articles [3] and [4].

The performance of each classifier was evaluated using the following performance metrics: Accuracy, Precision, Recall, F1-measure, and Area Under the Receiver Operating Characteristic Curve (AUC) score. These metrics were computed for each mood category to assess the model's ability to identify specific emotions within both datasets. The metrics were also computed across all mood categories collectively. These results are averaged across the five independent trials.

IV. Machine Learners

We explore a diverse range of machine learning algorithms with a proven track record of success in various classification-based predictive tasks. While existing research on Turkish music emotion recognition has focused on specific models (refer to Related Literature section), our approach investigates a broader selection, including Gradient Boosting, XGBoost, CatBoost, Decision Tree, Random Forest, and Gaussian Naïve Bayes. These models offer a valuable opportunity to assess their generalizability to the domain of Turkish music sentiment or emotion analysis. Table I provides an overview of the specific machine learning algorithms utilized in our empirical analysis, along with their respective hyperparameter settings.

A. Gradient Boosting

Gradient Boosting combines multiple weak learners to create a strong predictive model. It works by sequentially adding models to an ensemble, with each new model trained to correct errors made by the previous ones. This iterative process gradually reduces errors, resulting in an effective predictive machine learning model.

B. CatBoost

CatBoost focuses on efficiently handling categorical features in the data. It uses methods like ordered boosting and gradient-based optimization to build accurate models while minimizing overfitting. CatBoost employs symmetric decision trees, where the same features are used at each level of the tree, making the modeling process faster and more robust. It dynamically adjusts the model during training, ensuring more accurate predictions even with complex datasets. CatBoost provides useful features like feature importance lists and decision tree plots, aiding in the interpretation of the model.

C. XGBoost

XGBoost enhances traditional gradient boosting by incorporating techniques to prevent overfitting and improve efficiency. It is designed to handle large datasets efficiently, making use of parallel processing for faster computations. XGBoost assigns importance to each feature in the data and adjusts them as it learns from the dataset. If a feature is found to be important in correcting mistakes made by previous models, it receives more emphasis in subsequent iterations. With its flexibility in tuning and built-in mechanisms for handling missing data, XGBoost is an effective classifier.

D. Decision Tree

Decision Tree is a relatively straightforward yet effective algorithm for learning from data [16]. It creates a series of if-then decision rules based on the input features. At each step, the algorithm selects the feature that best splits the data, aiming to create homogeneous groups. Decision trees are easier to understand and visualize, making them useful for exploring patterns in the data.

E. Random Forest

Random Forest is an ensemble-based learning method that constructs multiple decision trees during training. Each tree is built from a random subset of the training data and features, reducing overfitting, and improving generalization performance. During prediction, each tree independently classifies the input data, and the most common prediction among all trees is chosen as the final output. Random Forest is robust and scalable, making it suitable for a wide range of applications in machine learning.

F. Gaussian Naïve Bayes

Gaussian Naive Bayes is a relatively simple, yet effective classifier based on Bayes' theorem and the assumption of feature independence [12]. It is particularly useful for classification tasks with continuous features. The algorithm calculates the probability of each class given the input features and makes predictions based on these probabilities.

TABLE I. Machine Learning Algorithms and Hyperparameters (Scikit-Learn Python Package [9])

Machine Learner	Hyperparameters
Gradient Boosting	random_state=42, min_samples_split=5
CatBoost	Default settings
XGBoost	use_label_encoder=False, eval_metric='mlogloss'
Decision Tree	random_state=42, max_depth=100, min_samples_split=7
Random Forest	Default settings [9]
GaussianNB	Default settings [9]

V. Results and Discussion

For ease of comparison, the highest and lowest values for accuracy, precision, recall, and F1-measure in Tables II, III, IV, and V are presented in bold font. Table II provides a detailed breakdown of the performance metrics for each mood category while Table III provides the performance metrics for the overall machine-learning model across all five trials for the TME dataset. As observed in the table, CatBoost emerged as the best-performing model, consistently achieving the highest average accuracy of 0.821 and an F1-measure of 0.814 across all categories. Conversely, Decision Tree exhibited the lowest overall F1-measure of 0.638.

Table IV provides a detailed breakdown of the performance metrics for each mood category, while Table V provides the performance metrics for the overall machine-learning model across all five trials for the TurEV-DB dataset. As shown in the tables, CatBoost outperformed all other models, yielding the highest average accuracy of 0.898 and an F1-measure of 0.900. This impressive performance across all emotions suggests that CatBoost may be particularly adept at capturing the nuances of emotional expression within Turkish music data. On the other hand, Gaussian Naïve Bayes performed the worst overall, with an average accuracy of 0.618 and an F1-measure of 0.610.

While all models achieved reasonable performance on some emotions, the difficulty level varied across categories. The happy emotion exhibited the highest consistency in performance across all models and trials for the TME dataset; in the TurEV-DB, the angry emotion displayed the highest consistency in performance across all models except Gaussian Naïve Bayes. On the other hand, the sad emotion proved to be the most challenging to model, with significant variations in performance in the TME dataset. In the TurEV-DB, the happy emotion reveals the lowest consistency in performance across all models except Gaussian Naïve Bayes (see Table IV).

TABLE II. AVERAGE PERFORMANCE FOR EACH EMOTION CATEGORY IN FIVE TRIALS WITH THE TME DATASET

Classifier	Category	Accuracy	Precision	Recall	F1-measure
XGBoost	happy	**0.952**	0.850	**0.952**	0.895
	sad	0.639	0.665	0.639	0.650
	angry	0.820	0.866	0.820	0.841
	relax	0.784	0.820	0.784	0.796
CatBoost	happy	0.931	**0.893**	0.931	**0.911**
	sad	0.673	0.710	0.673	0.684
	angry	0.868	0.875	0.868	0.869
	relax	0.810	0.780	0.810	0.791
Gradient Boosting	happy	0.862	0.868	0.862	0.864
	sad	0.621	0.608	0.621	0.610
	angry	0.829	0.813	0.829	0.818
	relax	0.693	0.731	0.693	0.707
Decision Tree	happy	0.811	0.764	0.811	0.775
	sad	**0.439**	**0.570**	**0.439**	**0.491**
	angry	0.755	0.683	0.755	0.708
	relax	0.592	0.564	0.592	0.577
Random Forest	happy	0.924	0.833	0.924	0.874
	sad	0.581	0.732	0.581	0.643
	angry	0.833	0.874	0.833	0.850
	relax	0.816	0.699	0.816	0.749
Gaussian Naïve Bayes	happy	0.943	0.789	0.943	0.858
	sad	0.546	0.681	0.546	0.601
	angry	0.781	0.826	0.781	0.800
	relax	0.772	0.757	0.772	0.763

TABLE III. AVERAGE PERFORMANCE FOR ALL EMOTION CATEGORIES IN FIVE TRIALS WITH THE TME DATASET

Classifier	Accuracy	Precision	Recall	F1-measure	AUC-score
XGBoost	0.799	0.801	0.799	0.795	0.945
CatBoost	**0.821**	**0.815**	**0.821**	**0.814**	**0.949**
Gradient Boosting	0.752	0.755	0.752	0.750	0.931
Decision Tree	**0.649**	**0.645**	**0.649**	**0.638**	**0.787**
Random Forest	0.789	0.785	0.789	0.779	0.939
Gaussian Naïve Bayes	0.761	0.763	0.761	0.756	0.925

TABLE IV. AVERAGE PERFORMANCE FOR EACH EMOTION CATEGORY IN FIVE TRIALS WITH THE TUREV-DB DATASET

Classifier	Category	Accuracy	Precision	Recall	F1-measure
XGBoost	happy	0.875	0.879	0.875	0.877
	sad	0.899	0.912	0.899	0.905
	angry	0.914	0.912	0.914	0.913
	relax	0.889	0.868	0.889	0.877
CatBoost	happy	0.860	0.923	0.860	0.890
	sad	0.899	0.907	0.899	0.902
	angry	**0.953**	**0.913**	**0.953**	**0.932**
	relax	0.882	0.867	0.882	0.874
Gradient Boosting	happy	0.834	0.903	0.834	0.866
	sad	0.901	0.886	0.901	0.893
	angry	0.934	0.897	0.934	0.915
	relax	0.869	0.867	0.869	0.868
Decision Tree	Happy	0.683	0.623	0.683	0.652
	sad	0.676	0.697	0.676	0.685
	angry	0.757	0.725	0.757	0.740
	relax	0.611	0.681	0.611	0.643
Random Forest	happy	0.770	0.860	0.770	0.811
	sad	0.861	0.847	0.861	0.853
	angry	0.902	0.817	0.902	0.857
	relax	0.802	0.844	0.802	0.821
Gaussian Naïve Bayes	happy	0.704	**0.502**	0.704	**0.586**
	sad	**0.493**	0.845	**0.493**	0.623
	angry	0.593	0.681	0.593	0.633
	relax	0.679	0.536	0.679	0.597

TABLE V. AVERAGE PERFORMANCE FOR ALL EMOTION CATEGORIES IN FIVE TRIALS WITH THE TUREV-DB DATASET

Classifier	Accuracy	Precision	Recall	F1-measure	AUC-score
XGBoost	0.894	0.893	0.894	0.893	0.985
CatBoost	**0.898**	**0.903**	**0.898**	**0.900**	**0.989**
Gradient Boosting	0.884	0.888	0.884	0.886	0.983
Decision Tree	0.682	0.681	0.682	0.680	**0.797**
Random Forest	0.834	0.842	0.834	0.836	0.969
Gaussian Naïve Bayes	**0.618**	**0.641**	**0.618**	**0.610**	0.829

VI. CONCLUSION

This study investigated the effectiveness of six different machine learning models in classifying emotional sentiment within Turkish music. We employed two distinct datasets and a rigorous data preparation methodology to ensure the quality and suitability of the data for learning and analysis.

The CatBoost learner emerged as the best-performing model, consistently achieving the highest average accuracy and F1-measure across all emotion categories. Additionally, the happiness emotion exhibited highest consistency in performances across all models and trials in the TME dataset, while the angry emotion had highest consistency in performances across that of the TurEV-DB dataset.

This research provides a unique contribution to the field of music emotion recognition by focusing on the analysis of Turkish music. We will continue to expand upon this research by evaluating the performance of these models on additional Turkish music datasets encompassing a wider range of musical styles and emotional expressions. To further explore the generalizability of our findings, we will conduct experiments with music emotions data from different languages. The underlying characteristics of the TME and TurEV-DB datasets will be investigated for their influence, if any, upon the performances of the CatBoost and XGBoost machine learners.

REFERENCES

[1] S. Koelsch, "A coordinate-based meta-analysis of music-evoked emotions," *NeuroImage*, vol. 223, p. 117350, Dec. 2020, doi: https://doi.org/10.1016/j.neuroimage.2020.117350.

[2] S. Hizlisoy, S. Yildirim, and Z. Tufekci, "Music emotion recognition using convolutional long short term memory deep neural networks," *Engineering Science and Technology, an International Journal*, Nov. 2020, doi: https://doi.org/10.1016/j.jestch.2020.10.009.

[3] M. Bilal Er and I. B. Aydilek, "Music Emotion Recognition by Using Chroma Spectrogram and Deep Visual Features," *International Journal of Computational Intelligence Systems*, 2019, doi: https://doi.org/10.2991/ijcis.d.191216.001.

[4] Salih Firat Canpolat, Zuhal Ormanoglu, and Deniz Zeyrek, "Turkish Emotion Voice Database (TurEV-DB)," pp. 368–375, May 2020.

[5] A. O. Durahim, A. Coşkun Setirek, B. Başarır Özel, and H. Kebapçı, "Music emotion classification for Turkish songs using lyrics," *Pamukkale University Journal of Engineering Sciences*, vol. 24, no. 2, pp. 292–301, 2018, doi: https://doi.org/10.5505/pajes.2017.15493.

[6] S. Shukla, P. Khanna and K. K. Agrawal, "Review on sentiment analysis on music," *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, Dubai, United Arab Emirates, 2017, pp. 777-780, doi: 10.1109/ICTUS.2017.8286111.

[7] C. Oflazoglu and S. Yildirim, "Recognizing emotion from Turkish speech using acoustic features," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, Dec. 2013, doi: https://doi.org/10.1186/1687-4722-2013-26.

[8] G. M. Biancofiore, T. Di Noia, E. Di Sciascio, F. Narducci, and P. Pastore, "Aspect-based sentiment analysis in music," *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, Apr. 2022, doi: https://doi.org/10.1145/3477314.3507092.

[9] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[10] 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, Feb. 2008, doi: https://doi.org/10.1109/tasl.2007.911513.

[11] N. S. Suhaimi, J. Mountstephens, and J. Teo, "EEG-Based Emotion Recognition: A State-of-the-Art Review of Current Trends and Opportunities," *Computational Intelligence and Neuroscience*, vol. 2020, pp. 1–19, Sep. 2020, doi: https://doi.org/10.1155/2020/8875426.

[12] X. Liu, "Research on Music Genre Recognition Based on Improved Naive Bayes Algorithm," *Mobile Information Systems*, vol. 2022, pp. 1–8, Jun. 2022, doi: https://doi.org/10.1155/2022/1909928.

[13] E. Altenmüller, S. Siggel, B. Mohammadi, A. Samii, and T. F. Münte, "Play it again sam: brain correlates of emotional music recognition," *Frontiers in Psychology*, vol. 5, 2014, doi: https://doi.org/10.3389/fpsyg.2014.00114.

[14] M. Allen, "People who deeply grasp the pain or happiness of others also process music differently in the brain," *SMU Research*. https://blog.smu.edu/research/2018/06/11/people-who-deeply-grasp-the-pain-or-happiness-of-others-also-process-music-differently-in-the-brain/

[15] W. Na and F. Yong, "Music Recognition and Classification Algorithm considering Audio Emotion," *Scientific Programming*, vol. 2022, p. e3138851, Jan. 2022, doi: https://doi.org/10.1155/2022/3138851.

[16] M. Akamine and J. Ajmera, "Decision tree-based acoustic models for speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2012, no. 1, Feb. 2012, doi: https://doi.org/10.1186/1687-4722-2012-10.