Journal of E-Technology

DLINE JOURNALS

Print ISSN: 0976-3503
Online ISSN: 0976-2930

JET2025: 16 (4)
https://doi.org/10.6025/jet/2025/16/4/142-149

Multimodal Music Emotion Classification via Stacking-Based Fusion of Audio and Lyric Features with Transfer Learning

Xiaojuan Chen
School of Marxism, Chongqing Vocational College of Transportation
Chongqing, 402247. China
cxj3408@163.com

ABSTRACT

This paper proposes a novel multimodal music emotion classification algorithm that integrates audio and lyrical features to overcome the limitations of single modality approaches. Recognizing that music conveys emotion through both sound and text, the system employs deep learning techniques, specifically combining One D CNN and TwoD CNN models with C3D and I3D frameworks for audio processing, alongside text analysis using TF IDF and Word2vec. To effectively fuse these heterogeneous modalities, the study implements a stacking based decision level fusion strategy with a Soft Max secondary classifier, significantly outperforming feature-level and traditional decision fusion methods. Utilizing transfer learning on datasets like Sport-1M and Kinetics enhances model generalization, while Adam and SGD optimizers improve training efficiency. Experimental results on a dataset of 2000 songs (anger, happiness, relaxation, sadness) demonstrate that the proposed multimodal approach achieves a maximum classification accuracy of 78%, a 4% improvement over single-modal classifiers and a 2% gain over other fusion techniques. The method effectively mitigates data heterogeneity and over fitting via 5-fold cross validation and addresses challenges in classifying "relaxation" by leveraging complementary audio lyric cues. Evaluation metrics (accuracy, F1-score, ROC/AUC) confirm superior performance, validating that synergistic integration of audio spectral features and semantic lyric representations yields more precise, robust, and scalable music emotion recognition, even with limited labeled data.

Keywords: Multimodal Emotion Classification, Music Emotion Recognition, Audio-Lyric Fusion, Deep Learning, Transfer Learning, Stacking-Based Fusion, C3D/I3D Networks, TF-IDF/Word2vec

Received: 4 April 2025, Revised 27 June 2025, Accepted 10 July 2025

1. Introduction

In the realm of music, we have identified numerous essential emotional indicators. By organizing these indicators into distinct categories, we can enhance the way we manage and access this information. However, as music

plays on, its emotional characteristics become more extensive and challenging to pinpoint accurately. A significant number of studies on music emotion classification concentrate on single channel and text based methods, neglecting the intricate interactions between sound waves and lyrics. This oversight can compromise the accuracy of many datasets and lead to potential data gaps. To tackle this issue, we can create a new multichannel music emotion classification system that merges sound waves with lyrics, greatly enhancing classification precision [1].

With the progress of artificial intelligence, it can aid us in accurately identifying and processing various forms of information, allowing us to predict and recognize a broader spectrum of music with greater accuracy. To improve our understanding and processing of this information, researchers are actively investigating more effective data mining and recognition techniques to develop more precise lyric databases across different music genres. Conventional content segmentation models are becoming inadequate for today's fast evolving market due to their inefficiency and unreliability. Thus, the advancement of artificial intelligence technology offers us a more effective solution, employing natural language processing methods to achieve swift and precise emotion recognition based on diverse textual features. As scientific technology progresses, classification algorithms derived from machine learning and deep neural networks have been utilized in various contexts [2], including visual, auditory, and tactile domains, achieving impressive outcomes.

This paper seeks to investigate a groundbreaking multimodal music emotion classification algorithm that proficiently extracts features from various music modalities. The analysis and preprocessing of experimental data effectively address the analytical difficulties posed by the length of music. Through comparative studies, we discover that a new combination network classification algorithm can more effectively manage emotion-related challenges. This algorithm adeptly integrates emotional information from both music audio and lyrics, thus enhancing classification accuracy. In comparison to traditional multimodal fusion techniques, this method is more efficient. These strategies tackle the challenges of language recognition, such as how to effectively organize elements like sound and linguistic data into a unified whole. Furthermore, these techniques can facilitate rapid and precise language recognition. Through comprehensive research, we have found that utilizing multimodal approaches, particularly in emotion recognition, has produced favorable results. Therefore, we suggest that to overcome these challenges, we implement a combination of multiple modalities for emotion recognition.

2. Related Work

Multimodal data fusion techniques have found extensive applications across various sectors. For instance, in the domains of vision and hearing, they merge information from both modalities to achieve more accurate recognition and forecasting. Multimodal data can also be applied in music recognition and prediction, integrating sound and lyrics to capture a song's essence better and fulfill societal demands. By employing various strategies such as decision level (classifier) fusion and feature level (genomic) fusion, multiple modalities can be effectively combined.

In the study carried out by Phu V N, they examined the traits of lyrics through a bag of words model and sentiment lexicon. They also assessed the features of audio based on rhythm and timbre, integrating all this data to create a multimodal classifier. Through linear regression analysis, the benefits of this framework became clear, significantly lowering the bias found in single modal classifiers to a distortion rate as minimal as 31% [3]. In thorough investigations, Su-Zhi Z found that conventional voiceprint recognition methods could have been more effective in capturing the essence of a song. Consequently, they developed a novel technique that employs both

voiceprint recognition and text recognition technologies, merging various modalities of voiceprint recognition (including +text) [4] to accurately determine a song's essence. Tai Z S successfully executed multidimensional emotion recognition on 3,766 songs spanning diverse styles using the AdaBoost algorithm alongside decision trees, where 14 distinct emotions were meticulously classified with an accuracy rate of 78.19%, significantly outperforming the SVM algorithm [6]. Jing L I and colleagues extracted features from lyrics using TF-IDF and LSI techniques, mitigating noise by decreasing the dimensionality of the lyrics, and subsequently classifying them with SVM technology [5]. Moreover, to enhance audio classification, they employed a BP neural network integrating four crucial factors MFCC, pitch, duration, and timbre to develop an innovative LFSM algorithm that achieves highly cohesive and precise sound classification. Through the investigations of Phu V N and others, it was revealed that by merging low level information from two distinct acoustic patterns into a single context, hidden insights can be more effectively recognized and differences in acoustic patterns can be efficiently minimized, leading to improved recognizability. We adopt a new technique that conducts multidimensional analysis of audio emotions. Initially, we apply the Yuvaraj algorithm to determine the pitch of the audio and then utilize the Deep Boltzmann Machine algorithm [8] for a thorough exploration of the semantics behind these pitches, facilitating better identification of various emotions [7]. Additionally, we implement a new algorithm capable of accurately distinguishing different emotions, providing substantial support to our research. "Dat N D" combines diverse forms of information from audio and songs using the random forest algorithm, resulting in a reduction of their F-value by 56.8% [9]. Li Z Q and others utilise the melody and semantic features of "MIDI" music files and "MIDI" songs to classify them through supervised learning. The results of the experiments indicate that with only basic audio processing, their classification efficiency stands at just 44.3%. However, when applying "multimodal" processing, the efficiency can be elevated to 61.1%. In the advancement of this technology, the application of melody and audio has proven to be particularly important [10].

3. Classification Algorithm Design

3.1 Input Preprocessing

To effectively leverage multimodal deep learning techniques for enhancing model accuracy, we must implement preprocessing strategies like C3D or I3D on the input video data. The detailed procedures can be referenced in Figure 1. To achieve more precise outcomes, we can utilize the Fast Fourier Transform to extract the spectral characteristics of each frame. Following that, we can derive additional information by employing the equation to convert Mel and f into frequency.

$$Mel(f) = 2595 \lg\left(1 + \frac{f}{700}\right) \tag{1}$$

Prior to executing visual emotion recognition utilizing C3D and I3D video frameworks, it is essential to accurately configure the dimensions, quantity, and transmission pathways for all frames. To realize this aim, C3D and I3D networks employ a sequence of 32 frames featuring diverse colors.

3.2 Multimodal Classification

By integrating music CNN (OneDMCNN), CNN (TwoDMCNN), and other three dimensional music media, we can develop a sophisticated music transmission system encompassing four distinct modalities: C3D+OneDMCNN, I3D+OneDMCNN, C3D+TwoDMCNN, and I3D+TwoDMCNN. This enables us to manage the intricacies of music transmission and significantly enhance transmission efficiency. While processing audio and video, it is crucial to fine-tune their emotion classifiers. We will eliminate all Soft Max classifiers and utilize their classification outputs

in the multimodal data fusion. For the text classification model (M2), it is also necessary to refine its classifier to improve classification efficiency. We will enhance the accuracy of segmentation by employing training set P, test set T (with dimensions similar to the original dataset and comprising 2 columns of class label data), training set P_2, training set T_2, training set P_1, training set T_1, etc., as illustrated in the subsequent equation:

$$(P1)+(P2)=(P1P2)=(P) \quad (2)$$

$$(T1)+(T2)=(T1T2)=(T) \quad (3)$$

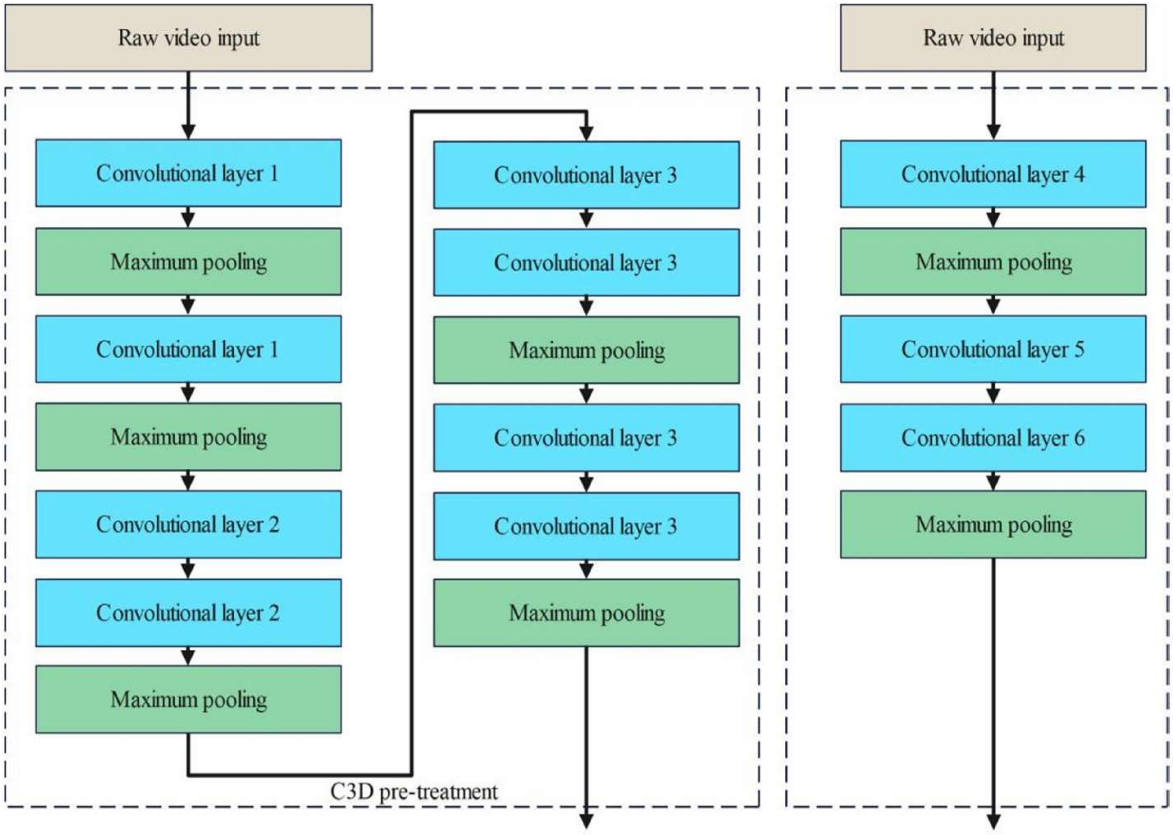

Figure 1. Input preprocessing

By incorporating transfer learning methodologies, we can effectively tackle the issue of missing data and achieve successful classification of music and videos. Specifically, we can optimize the weights of the algorithm during the initialization phase to better align with the characteristics of music and videos. Furthermore, we can extract a range of learning traits for individual emotion classifications to support multiple decision making scenarios. By applying three distinct datasets, sport-1M, RGBImageNet, and Kinetics, to C3D, c3d, I3D, CNN, and other two dimensional music models, we can efficiently synthesize this data to facilitate effective generalization of two dimensional music. Transfer learning allows us to acquire complex information from diverse data sources, promoting multimodal fusion that can occur at early, late, or mixed stages to optimize various phases of the dataset. By analyzing the learning outcomes of different single modal networks, we can compile a series of datasets to enhance the utilization of SoftMax for music or video emotion classification.

4. Experimental Design and Analysis

4.1 Experimental Design

In this research, we utilized a collection of datasets akin to single modal studies. After validation, audio from "angry" songs yielded strong performance, while "happy" lyrics text also demonstrated promising results. Conversely, the classification accuracy for the emotion "relaxed" was lacking. By combining data from both modalities, we can more effectively assess the precision of music emotion classification. In this experiment, we discovered that excitement typically denotes a state of elevated spirits stemming from happiness or stimulation, which can affect individuals' physical movements, social interactions, and their environment. This genre of music often features a pronounced rhythm, characterized by rapid melodies, powerful harmonies, smooth tempos, or diverse beats. Their creation is rooted in human responses to danger and fear, utilizing rapid, intense, and irregular beats to convey visual information. Relaxation can be perceived as a harmony of body and mind, aiding us in stress relief and promoting a sense of calm. This emotion can be expressed through the appreciation of natural landscapes, listening to music, and other activities. In contrast, sadness can be articulated through slow tempos and soft tones, enabling us to engage with the essence of life profoundly. Tense emotions may be depicted through violent imagery, typically expressed through loud volumes, quick tempos, and intense chords.

4.2 Experimental Results

By implementing transfer learning strategies, we can enhance the conventional CNN algorithm and more effectively manage different types of information, such as music and video in our research. To evaluate the algorithm's efficacy, we utilize several metrics including accuracy, F1 score, and the receiver operating characteristic (ROC) curve. The F1 score illustrates the algorithm's effectiveness by reflecting its precision, indicating how well each test result corresponds with the expected outcome. Moreover, the AUC assesses the algorithm's performance by showing the true positive rate of each test result, with its value dependent on the integration level of the test outcomes.

To boost efficiency, we apply two optimization algorithms separately in the two neural networks. These algorithms enhance the model's performance by adjusting its features. Following processing by the Adam and SGD optimizers, the one dimensional music CNN and I3D achieve commendable results, while the two dimensional music CNN and C3D also perform excellently with a learning rate of 0.001. By merging the Adam optimizer with the SGD optimizer, we can effectively unify the multimodal framework of audio and video. Notably, the performance of the CNN optimizer is exceptional, significantly enhancing the performance of the one-dimensional music CNN to fulfill our requirements better. Although the one dimensional music CNN is capable of capturing features such as time and spatial resolution, tonality, etc., it still falls short of the efficiency exhibited by the two-dimensional music CNN. Furthermore, to obtain more accurate Mel spectrograms, zero padding must be implemented during the 2D music CNN process to achieve superior audio quality. Mel spectrograms illustrate the vibrational attributes of an object, represented as a two dimensional graph and can be calculated based on their transient characteristics.

Combining single modal and SoftMax classifiers can effectively extract the best characteristics for music and video emotion analysis. Thus, we can utilize information obtained from single modal classifiers and achieve multimodal partitioning with the help of SoftMax classifiers. Through six fold cross validation, we combine multiple modalities as shown in Figure 2. This set of experiments validates the classification performance of different multimodal fusion approaches. For audio and lyrics features, we compare multimodal music emotion

classification using F1 feature level fusion, F2 decision level fusion, F3 improved dimensional decision fusion, and F4 stacking based multimodal ensemble methods. The experimental results indicate that due to the heterogeneity between different modalities, the improvement in classification accuracy through feature level fusion methods is not significant. The classification results are often related to the accuracy of the individual modalities and lack universality.The improved dimensional emotion decision fusion method utilizes the performance of different modalities on the Thayer emotional axis, resulting in a 2% increase in accuracy compared to single modal classification, with a maximum accuracy of 76%. However, this decision fusion method also needs more scalability in a multi classification system. The proposed stacking based multimodal ensemble method performs best by integrating audio and lyrics' emotion features from single modal classifiers and outputting a unified classification result using the sub classifier. Compared to single-modal data, it achieves a 4% increase in accuracy, with a maximum accuracy of 78%. Through decision level feature integration, we find that combining various audiovisual features with the SoftMax operator achieves superior results. By comparing and analyzing, we discover that using integrated multimodal classifiers exhibits high performance, providing strong decision support and perfectly integrating learning characteristics from various modalities, resulting in high accuracy performance.

To further assess the efficacy of the research model presented in this paper, various classification models suggested by researchers in the domain of music emotion classification in recent years are utilized on the experimental dataset for performance evaluation. A range of classification techniques are implemented in this series of experiments to verify their effectiveness. During the experiments, the Word2vec tool is employed to extract feature representations of word items in the text, and vector concatenation is conducted to derive the feature vector of the complete sample. CNN, LSTM, and CNN LSTM techniques are applied as classification models, and the resulting classification accuracy is illustrated in Figure 3. The experimental conditions are consistent with the foundational controlled experiment detailed in this paper. The five emotion labels (Patt1: anger, Patt2: happiness, Patt3: relaxation, Patt4: sadness, and Patt5: average) are extracted from the Last.fm tag subset, with audio and lyrics text files retrieved separately using downloading tools. Each emotion label extracts 500 songs, resulting in a total of 2000 songs. The complete dataset is randomly partitioned into 80% for training purposes and 20% for

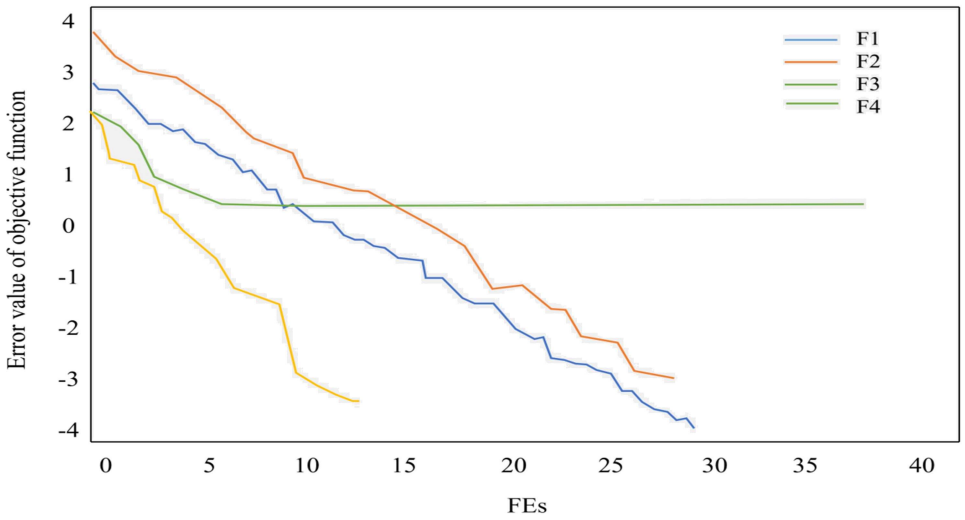

Figure 2. Evaluation Results of Single Mode Combination and Integrated Multimode

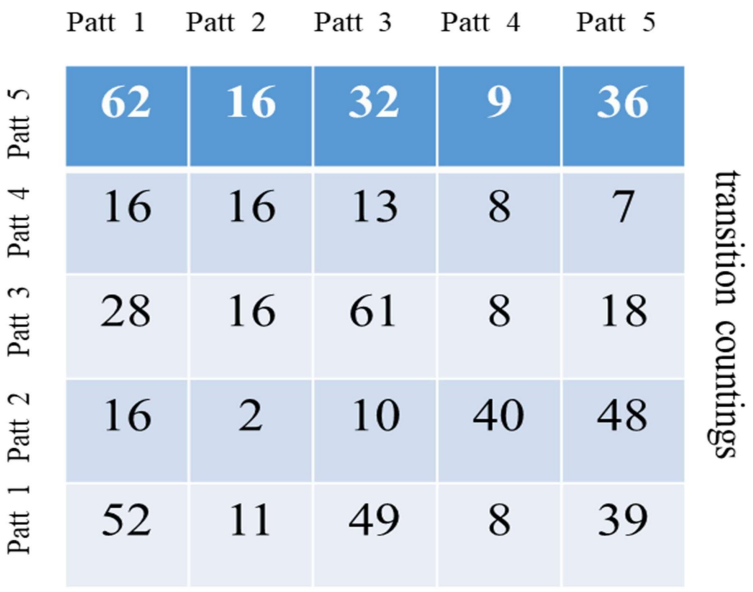

Figure 3. Confusion matrix of Integrated Multimode

testing. The experiment demonstrates that the overall performance of lyrics emotion classification surpasses that of audio classification, and the distribution of classification outcomes is more balanced. However, there are challenges regarding the classification performance of the “relaxation” emotion category. Specifically, concerning different feature extraction techniques, the TF-IDF method, which relies solely on word frequency, needs to show improved emotion classification performance when dealing with lyrics text samples that have implicit emotional semantics. On the other hand, the chi-square test method significantly improves the performance, with the improved chi-square test extraction method achieving a classification accuracy of 69%. At the same time, the distributed word vector feature representation extracted using the Word2vec tool achieves good accuracy in the SVM classifier. This distributed vector can serve as an input for deep network methods and provides a foundation for comparative experiments of different classification methods. According to the data in Figure 3, we observe that the performance of relaxation and sadness is most pronounced under different emotional conditions, while the performance of excitement and tension is most pronounced under different emotional conditions.

5. Conclusion

Music embodies a wealth of human emotional information, and exploring music emotion classification aids in organizing and retrieving vast amounts of music data. Music encompasses two modalities of emotional information: audio and lyrics. By establishing a multimodal music emotion classification system, we can significantly enhance classification performance. This paper selects the Thayer emotional model as the foundation for music emotion classification, categorizing music into four types: anger, happiness, relaxation, and sadness. Utilizing transfer learning and late stage decision level fusion, a music video emotion classification algorithm founded on multimodal deep learning is introduced. A compact music video dataset is created, separating the music and video components for pre training other audio and video CNNs. Results from experimental evaluation reveal that employing various modalities of classification methods as base classifiers, mitigating over fitting

through 5-fold cross validation, and generating a new dataset feature comprising emotion label results, along with fusing emotion output using a softmax secondary classifier, effectively addresses the heterogeneity issue. The integrated approach is straightforward to implement, fully leverages the classification capabilities of different algorithms to enhance each other's strengths, boosts the accuracy of music emotion classification, and demonstrates strong scalability. Multimodal fusion can significantly augment classification performance. These findings suggest that the proposed algorithm can learn the multimodal features of music videos and achieve precise and efficient emotion classification, even when the labeled data samples are limited.

References

[1] Shunmuganathan, K. L., Kalaivani, P. (2016). Feature selection based on genetic algorithm and hybrid model for sentiment polarity classification. *International Journal of Data Mining, Modelling & Management*, *8*(4), 315.

[2] Tripathy, A., Agrawal, A., Kumar, R. S. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, *3*, 120.

[3] Phu, V. N., Tran, V. T. N., Max, J. (2018). A CURE algorithm for Vietnamese sentiment classification in a parallel environment. *(Issue 7)*, 110.

[4] Su-zhi, Z., Pei-Feng, S. (2019). Sentiment classification of network comments based on KSVM. *Journal of Zhengzhou University of Light Industry (Natural Science)*, 11–14.

[5] Tai, Z. S., Fei, W. F., Fan, D., et al. (2018). Research on the majority decision algorithm based on WeChat sentiment classification. *Journal of Intelligent & Fuzzy Systems*, 1–10.

[6] Jing, L. I., Hongfei, L., Ruimin, L. I. (2012). Sentiment vector space model-based music emotion tag prediction. *Journal of Chinese Information Processing*, *26*(6), 45–44.

[7] Phu, V. N., Tran, V. T. N., Dat, N. D., others. (2017). STING algorithm used English sentiment classification in a parallel environment. *International Journal of Pattern Recognition and Artificial Intelligence*, *(8)*, 9–11.

[8] Yuvaraj, N., Sabari, S., others. (2017). Twitter sentiment classification using binary shuffled frog algorithm. *Intelligent Automation and Soft Computing*, *23*(2), 373–381.

[9] Dat, N. D., Phu, V. N., Tran, V. T. N., others. (2017). STING algorithm used English sentiment classification in a parallel environment. *International Journal of Pattern Recognition and Artificial Intelligence*, *31*(7), 1750021.

[10] Li, Z. Q., Yang, D. Q., Tan, Y., others. (2014). An improved Naive Bayesian classification algorithm for sentiment classification of microblogs. *Applied Mechanics and Materials*, *(543–547)*, 3614–3620.