Modeling Emotion Dynamics in Song Lyrics with State Space Models

Yingjin Song*
Department of Information and Computing Sciences
Utrecht University, Netherlands
y.song5@uu.nl

Daniel Beck
School of Computing and Information Systems
University of Melbourne, Australia
d.beck@unimelb.edu.au

Abstract

Most previous work in music emotion recognition assumes a single or a few song-level labels for the whole song. While it is known that different emotions can vary in intensity within a song, annotated data for this setup is scarce and difficult to obtain. In this work, we propose a method to predict emotion dynamics in song lyrics *without song-level supervision*. We frame each song as a time series and employ a State Space Model (SSM), combining a sentence-level emotion predictor with an Expectation-Maximization (EM) procedure to generate the full emotion dynamics. Our experiments show that applying our method consistently improves the performance of sentence-level baselines without requiring any annotated songs, making it ideal for limited training data scenarios. Further analysis through case studies shows the benefits of our method while also indicating the limitations and pointing to future directions.

1 Introduction

Music and emotions are intimately connected, with almost all music pieces being created to express and induce emotions (Juslin and Laukka, 2004). As a key factor of how music conveys emotion, lyrics contain part of the semantic information that the melodies cannot express (Besson et al., 1998). Lyrics-based music emotion recognition has attracted increasing attention driven by the demand to process massive collections of songs automatically, which is an important task for streaming and media service providers (Kim et al., 2010; Malheiro et al., 2016; Agrawal et al., 2021).

Vanilla emotion recognition studies in Natural Language Processing (NLP) assume the text instance expresses a static and single emotion (Mohammad and Bravo-Márquez, 2017; Nozza et al., 2017; Mohammad et al., 2018). However, emotion is non-static and highly correlated with the contextual information, making the single-label assumption too simplistic in dynamic scenarios, not just in music (Schmidt and Kim, 2011) but also in other domains such as conversations (Poria et al., 2019b). Figure 1 shows an example of this dynamic behavior, where the intensities of three different emotions vary within a song. Accurate emotion recognition systems should ideally generate the full emotional dynamics for each song, as opposed to simply predicting a single label.

A range of datasets and corpora for modeling dynamic emotion transitions has been developed in the literature (McKeown et al., 2011; Li et al., 2017; Hsu et al., 2018; Poria et al., 2019a; Firdaus et al., 2020), but most of them do not use song lyrics as the domain and assume discrete, categorical labels for emotions (either the presence or absence of one emotion). To the best of our knowledge, the dataset from Mihalcea and Strapparava (2012) is the only one that provides full fine-grained emotion intensity annotations for song lyrics at the verse[1] level. The lack of large-scale datasets for this task poses a challenge for traditional supervised methods. While previous work proposed methods for the similar sequence-based emotion recognition task, they all assume the availability of some levels of annotated data at training time, from full emotion dynamics (Kim et al., 2015) to coarse, discrete document-level labels (Täckström and McDonald, 2011b).

The data scarcity problem motivates our main research question: *''Can we predict emotion dynamics in song lyrics without requiring annotated lyrics?''* In this work, we claim that the answer is affirmative. To show this, we propose

*Work done when the first author was at The University of Melbourne.

[1]According to Mihalcea and Strapparava (2012), a ''verse'' is defined as a sentence or a line of lyrics.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 157–175, 2023. https://doi.org/10.1162/tacl_a_00541
Action Editor: Sara Rosenthal. Submission batch: 9/2022; Revision batch: 10/2022; Published 2/2023.
© 2023 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

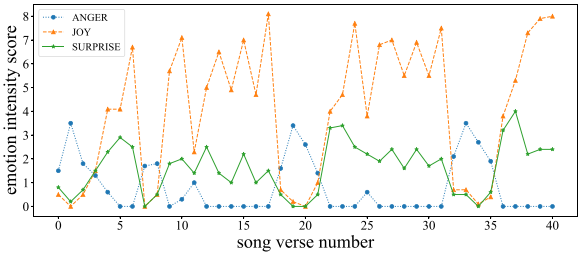

Figure 1: An illustration of emotion dynamics of a song in the LyricsEmotions dataset of Mihalcea and Strapparava (2012). Note the intensities of each emotion vary from verse to verse within the song.

a method consisting of two major stages: (1) a sentence or *verse-level* regressor that leverages existing emotion lexicons, pre-trained language models and other sentence-level datasets, and (2) a State Space Model (SSM) that constructs a full *song-level* emotional dynamics given the initial verse-level scores. Intuitively, we treat each verse as a time step and the emotional intensity sequence as a latent time series that is inferred without any song-level supervision, directly addressing the limited data problem. To the best of our knowledge, *this scenario was never addressed before either for song lyrics or other domains*.

To summarize, our main contributions are:

- We propose a hybrid approach for verse-level emotion intensity prediction that combines emotion lexicons with a pre-trained language model (BERT [Devlin et al., 2019] used in this work), which is trained on available sentence-level data.
- We show that by using SSMs to model song-level emotion dynamics, we can improve the performance of the verse-level approach without requiring any annotated lyrics.
- We perform a qualitative analysis of our best models, highlighting its limitations and pointing to directions for future work.

2 Background and Related Work

Emotion Models. Human emotion is a long-standing research field in psychology, with many studies aiming at defining a taxonomy for emotions. In NLP, emotion analysis mainly employs the datasets which are annotated based on the *categorical* or the *dimensional* model.

The categorical model assumes a fixed set of discrete emotions that can vary in intensity. Emotions can overlap but are assumed to be separate ''entities'' from each other, such as *anger*, *joy*, and *surprise*. Taxonomies using the categorical model include Ekman's basic emotions (Ekman, 1993), Plutchik's wheel of emotions (Plutchik, 1980), and the OCC model (Ortony et al., 1988). The dimensional models place emotions in a continuous space: The VAD (*Valence*, *Arousal*, and *Dominance*) taxonomy of Russell (1980) is the most commonly used in NLP. In this work, we focus on the Ekman taxonomy for purely experimental purposes, as it is the one used in the available data we employ. However, our approach is general and could be applied to other taxonomies.

Dynamic Emotion Analysis. Emotion Recognition in Conversation (ERC, Poria et al., 2019b), which focuses on tracking dynamic shifts of emotions, is the most similar task to our work. Within a conversation, the emotional state of each utterance is influenced by the previous state of the party and the stimulation from other parties (Li et al., 2020; Ghosal et al., 2021). Such an assumption of the real-time dynamic emotional changes also exists in music: The affective state of the current lyrics verse is correlated with the state of the previous verse(s) as a song progresses.

Contextual information in the ERC task is generally captured by deep learning models, which can be roughly categorized into sequence-based, graph-based and reinforcement learning-based methods. Sequence-based methods encode conversational context features using established methods like Recurrent Neural Networks (Poria et al., 2017; Hazarika et al., 2018a,b; Majumder et al., 2019; Hu et al., 2021) and Transformer-based architectures (Zhong et al., 2019; Li et al., 2020). They also include more advanced and tailored methods such as Hierarchical Memory Network (Jiao et al., 2020), Emotion Interaction Network (Lu et al., 2020), and Causal Aware Network (Zhao et al., 2022). Graph-based methods apply specific graphical structures to model dependencies in conversations (Ghosal et al., 2019; Zhang et al., 2019; Lian et al., 2020; Ishiwatari et al., 2020; Shen et al., 2021) using Graph Neural Networks

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

(Kipf and Welling, 2017). Reinforcement Learning (RL)-based methods (Zhang et al., 2021; Huang et al., 2021) model the influence of the previous emotional state on current utterance's emotion by using agent-environment nature of dialogue systems. In contrast to these methods, we capture contextual information using a SSM, mainly motivated by the need for a method that can train without supervision. Extending and/or combining an SSM with a deep learning model is theoretically possible but non-trivial, and care must be taken in a low-data situation such as ours.

The time-varying nature of music emotions has been investigated in music information retrieval (Caetano et al., 2012). To link the human emotions with the music acoustic signal, the emotion distributions were modeled as 2D Gaussian distributions in the Arousal-Valence (A-V) space, which were used to predict A-V responses through multi-label regression (Schmidt et al., 2010; Schmidt and Kim, 2010). Building on previous studies, Schmidt and Kim (2011) applied structured prediction methods to model complex emotion-space distributions as an A-V heatmap. These studies focus on the mapping between emotions and acoustic features/signals, while our work focuses on the lyrics component. Wu et al. (2014) developed a hierarchical Bayesian model that utilized both acoustic and textual features, but it was only applied to predict emotions as discrete labels (presence or absence) instead of fine-grained emotion intensities as in our work.

Combining Pre-trained Language Models with External Knowledge. Pre-trained language models (LMs) including BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), and GPT (Brown et al., 2020) have achieved state-of-the-art performance in numerous NLP tasks. Considerable effort has been made towards combining context-sensitive features of LMs with factual or commonsense knowledge from structured sources, including commonsense knowledge (Zhong et al., 2019; Ghosal et al., 2020), domain-specific knowledge (Ying et al., 2019), structured semantic information (Zhang et al., 2020), language-specific knowledge (Alghanmi et al., 2020; De Bruyne et al., 2021), and linguistic features (Koufakou et al., 2020; Mehta et al., 2020). This auxiliary knowledge is usually infused into the architecture by concatenating them with the Transformer-based representation before the prediction layer for downstream tasks. Our method proposes to utilize the rule-based representations derived from a collection of affective lexicons to improve the performance of BERT by incorporating task-specific knowledge. The motivation for our proposal is the hypothesis that the extension of lexicon-based information will compensate for BERT's lack of proper representations of semantic and world knowledge (Rogers et al., 2021), making the model more stable across domains.

State Space Models. In NLP tasks such as Part-of-Speech (POS) tagging and Named Entity Recognition, contextual information is widely acknowledged to play an important role in prediction. This led to the adoption of structured prediction approaches such as Hidden Markov Model (HMM, Rabiner and Juang, 1986), Maximum Entropy Markov Model (MEMM, McCallum et al., 2000), and Conditional Random Field (CRF, Lafferty et al., 2001), which relate a set of observable variables to a set of latent variables (e.g., words and their POS tags). State Space Models are similar to HMMs but assume continuous variables. The Linear Gaussian SSM (LG-SSM) is a particular case of SSM in which the conditional probability distributions are Gaussian.

Following the notation from Murphy (2012, Chap. 18), we briefly introduce the LG-SSM that we employ in our work. LG-SSMs assume a sequence of observed variables $\mathbf{y}_{1:T}$ as input, and the goal is to draw inferences about the corresponding hidden states $\mathbf{z}_{1:T}$, where T is the length of the sequence. Their relationship is given at each step t by the equations as:

$$\mathbf{z}_t = \mathbf{A}\mathbf{z}_{t-1} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \mathbf{Q})$$
$$\mathbf{y}_t = \mathbf{C}\mathbf{z}_t + \delta_t, \qquad \delta_t \sim \mathcal{N}(0, \mathbf{R})$$

where $\Theta = (\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R})$ are the model parameters, ϵ_t is the system noise and δ_t is the observation noise. The equations above are also referred to as *transition*[2] and *observation* equations, respectively. Given Θ and a sequence $\mathbf{y}_{1:T}$, the goal is to obtain the posteriors $p(\mathbf{z}_t)$ for each step t. In an LG-SSM, this posterior is Gaussian and can be obtained in closed form by applying the celebrated Kalman Filter (Kalman, 1960).

[2]We omit control matrix $\mathbf{B}$ and control vector $\mathbf{u}_t$ in the transition equation, assuming no external influence.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

There are other latent variable models to estimate temporal dynamics of emotions and sentiments in product reviews (McDonald et al., 2007; Täckström and McDonald, 2011a,b) and blogs (Kim et al., 2015). McDonald et al. (2007) and Täckström and McDonald (2011a,b) combined document-level and sentence-level supervision as the observed variables to condition on the latent sentence-level sentiment. Kim et al. (2015) introduced a continuous variable $\mathbf{y}_t$ to solely determine the sentiment polarity $\mathbf{z}_t$, while $\mathbf{z}_t$ is conditioned on both $\mathbf{y}_t$ and $\mathbf{z}_{t-1}$ for each t in the LG-SSM.

3 Method

We propose a two-stage method to predict emotion dynamics without requiring annotated song lyrics. The first stage is a verse-level model that predicts initial scores for each verse, where we use a hybrid approach combining lexicons and sentence-level annotated data from a different domain (§ 3.1). The second stage contextualizes these scores in the entire song, incorporating them into an LG-SSM trained in an unsupervised way (§ 3.2).

Task Formalization. Let d_x^y indicate the real-valued intensity of emotion y for sentence/verse x, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Note that $\mathcal{Y} = \{y_1, y_2, \ldots, y_c\}$ is a set of c labels, each of which represents one of the basic emotions (c = 6 for the datasets we used). Given a source dataset $\mathbb{D}_s = \{(x_1, E_1), (x_2, E_2), \ldots, (x_M, E_M)\}$, where x_i is a sentence, $E_i = \{d_{x_i}^{y_1}, d_{x_i}^{y_2}, \ldots, d_{x_i}^{y_c}\}$ and $M = |\mathbb{D}_s|$. The target dataset is $\mathbb{D}_t = \{S_1, S_2, \ldots, S_{|\mathbb{D}_t|}\}$, where $|\mathbb{D}_t|$ is the number of sequences (i.e., songs) and $S_i = \{(v_1, E_1), (v_2, E_2), \ldots, (v_{|S_i|}, E_{|S_i|})\}$ is a song consisting of $|S_i|$ verses. In the song S_i, the j-th verse v_j is also associated with c emotion intensities as $E_j = \{d_{v_j}^{y_1}, d_{v_j}^{y_2}, \ldots, d_{v_j}^{y_c}\}$. Given the homogeneity of label spaces of $\mathbb{D}_s$ and $\mathbb{D}_t$, the model trained by using $\mathbb{D}_s$ can be applied to predict $\mathbb{D}_t$ directly. The output of verse-level model is the emotion intensity predictions $\hat{\mathbf{Y}} \in \mathbb{R}^{N \times c}$, where N is the total number of verses in $\mathbb{D}_t$. Finally, we use $\hat{\mathbf{Y}}$ as the input sequences of the song-level model to produce optimized emotion intensity sequences $\hat{\mathbf{Z}} \in \mathbb{R}^{|\mathbb{D}_t| \times c}$.

3.1 Verse-Level Model

Emotion lexicons provide information on associations between words and emotions (Ramachandran and de Melo, 2020), which are beneficial in recognizing textual emotions (Mohammad et al., 2018; Zhou et al., 2020). Given that we would like to acquire accurate initial predictions at the verse level, we opted for a hybrid methodology that combines learning-based and lexicon-based approaches to enhance feature representation.

Overview. The verse-level model architecture is called BERTLex, as illustrated in Figure 2. It consists of three phases: (1) the embedding phase, (2) the integration phase, and (3) the prediction phase. In the embedding phase, the input sequence is represented as both contextualized embeddings from BERT and static word embeddings from lexicons. In the integration phase, contextualized and static word embeddings are concatenated at the sentence level by taking the pooling operations on the two embeddings separately. The prediction phase encodes the integrated sequence of feature vectors and performs the verse-level emotion intensity regression by using the $\mathbb{D}_s$ as the training/development set and the $\mathbb{D}_t$ as the test set.

Embedding Phase. The input sentence S is tokenized in two ways: one for the pre-trained language model and the other for the lexicon-based word embedding. These two tokenized sequences are denoted as T^{cxt} and T^{lex}, respectively. Then, T^{cxt} is fed into the pre-trained language model to produce a sequence of contextualized word embeddings $E^{cxt} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{|T^{cxt}|}\}$, where $E^{cxt} \in \mathbb{R}^{|T^{cxt}| \times D_{cxt}}$ and D_{cxt} is the embedding vector dimension.

To capture task-specific information, a lexicon embedding layer encodes a sequence of emotion and sentiment word associations for T^{lex}, generating a sequence of lexicon-based embeddings $E^{lex} = \{\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_{|T^{lex}|}\}$, where $E^{lex} \in \mathbb{R}^{|T^{lex}| \times D_{lex}}$ and D_{lex} is the lexical embedding vector dimension. We first build the vocabulary $\mathbb{V}$ from the text of $\mathbb{D}_s$ and $\mathbb{D}_t$. For each word $\mathbf{v}_i$ in $\mathbb{V}$ of T^{lex}, we use d lexicons to generate the rule-based feature vectors $\boldsymbol{\ell}_i = \{\ell_{i_1}, \ell_{i_2}, \ldots, \ell_{i_d}\}$, where ℓ_{i_j} is the lexical feature vector for word $\mathbf{v}_i$ derived from the j-th lexicon and $D_{lex} = |\boldsymbol{\ell}_i|$. Additionally, we perform a degree-p polynomial expansion on the feature vector ℓ_{i_j}.

Integration Phase. As BERT uses the WordPiece tokenizer (Wu et al., 2016) to split a number of words into a sequence of subwords, the contextualized embedding cannot be directly concatenated with the different-size static word embedding. Inspired by Alghanmi et al. (2020), we

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

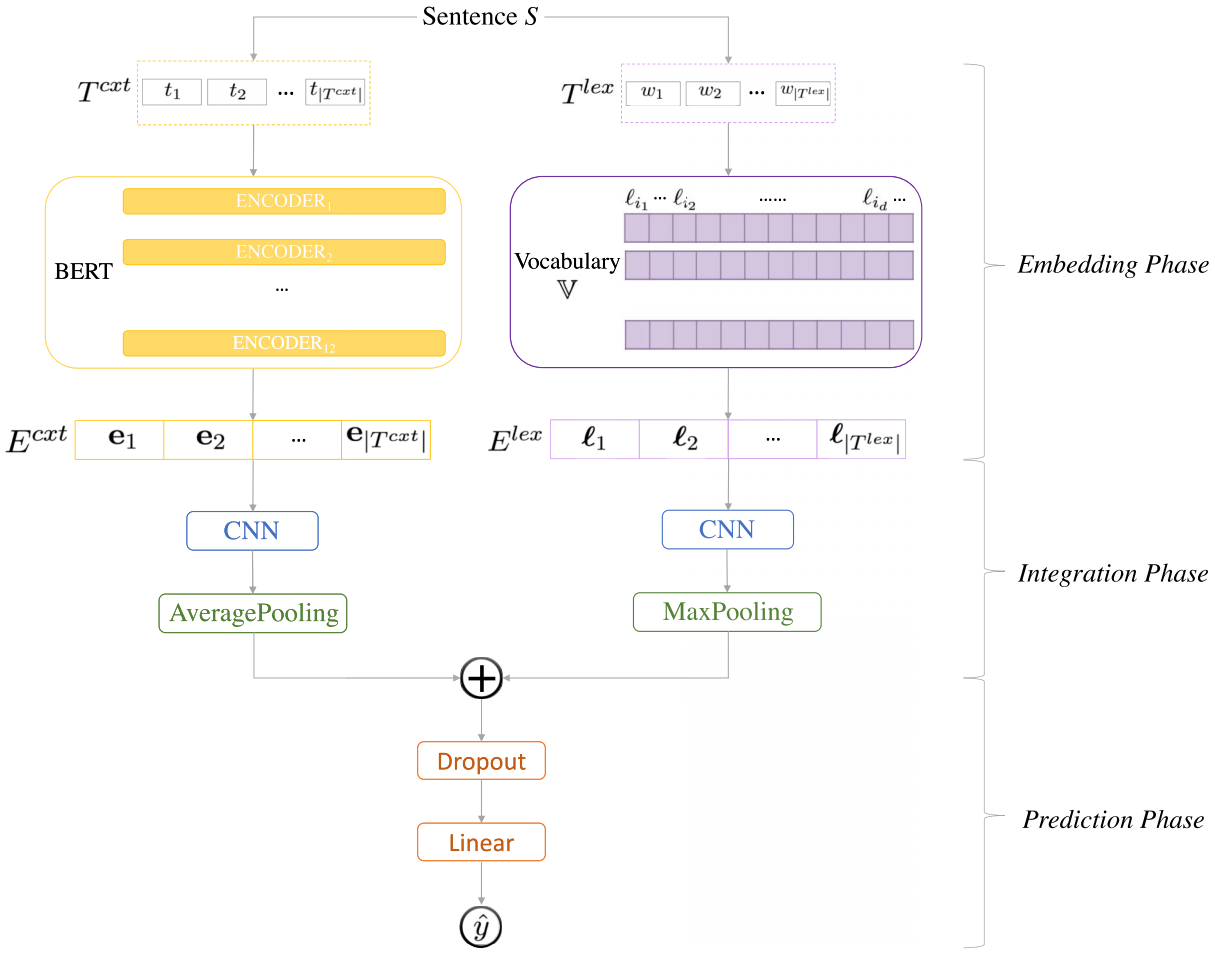

Figure 2: BERTLex architecture used for the verse-level model.

combine contextualized embeddings and static word embeddings at the sentence level by pooling the two embeddings E^{cxt} and E^{lex} separately. To perform initial feature extraction from the raw embeddings, we apply a single-layer 1D Convolutional Neural Network (Kim, 2014, CNN) with ReLU activation (Nair and Hinton, 2010) on each embedding as:

$$\mathbf{e}'_i = \text{ReLU}(\mathbf{W}_1[\mathbf{e}_i, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_{i+k-1}] + \mathbf{b}_1)$$

$$\boldsymbol{\ell}'_i = \text{ReLU}(\mathbf{W}_2[\boldsymbol{\ell}_i, \boldsymbol{\ell}_{i+1}, \ldots, \boldsymbol{\ell}_{i+k-1}] + \mathbf{b}_2)$$

where $\mathbf{W}_1$, $\mathbf{b}_1$, $\mathbf{W}_2$ and $\mathbf{b}_2$ are trainable parameters and k is the kernel size. We then apply the average pooling and max pooling on the feature maps, respectively:

$$\tilde{E}^{cxt} = \text{AvgPool}(\mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_{|T^{cxt}|-k+1}).$$

$$\tilde{E}^{lex} = \text{MaxPool}(\boldsymbol{\ell}'_1, \boldsymbol{\ell}'_2, \ldots, \boldsymbol{\ell}'_{|T^{lex}|-k+1}).$$

Finally, the contextualized embedding and the lexicon-based embedding are merged via a concatenation layer as $\tilde{E}^{cxt} \oplus \tilde{E}^{lex}$.

Prediction Phase. The prediction phase outputs the emotion intensity predictions $\hat{\mathbf{Y}} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\}$ by using a single dropout (Srivastava et al., 2014) layer and a linear regression layer. During training, the mean squared error loss is computed and backpropagated to update the model parameters.

3.2 Song-Level Model

After obtaining initial verse-level predictions, the next step involves incorporating these into a song-level model using an LG-SSM. We take one emotion as an example. Specifically, we consider the predicted scores of this emotion of each song as an *observed* sequence $\hat{\mathbf{y}}_i$. That is, we group the N predictions of $\hat{\mathbf{Y}}$ as $|\mathbb{D}_t|$ sequences of predictions as $\{\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_{|\mathbb{D}t|}\}$. For the i-th

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

Algorithm 1: Kalman Filter

Input : $\mathbf{y}_t, \hat{\mathbf{z}}_{t-1}, \mathbf{\Sigma}_{t-1}, \mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}$
Output: $\hat{\mathbf{z}}_t, \mathbf{\Sigma}_t$
PREDICTION:
$\hat{\mathbf{z}}_{t|t-1} = \mathbf{A}\hat{\mathbf{z}}_{t-1}$;
$\mathbf{\Sigma}_{t|t-1} = \mathbf{A}\mathbf{\Sigma}_{t-1}\mathbf{A}^\mathsf{T} + \mathbf{Q}$;
MEASUREMENT:
$\tilde{\mathbf{r}}_t = \mathbf{y}_t - \mathbf{C}\hat{\mathbf{z}}_{t|t-1}$;
$\mathbf{S}_t = \mathbf{C}\mathbf{\Sigma}_{t|t-1}\mathbf{C}^\mathsf{T} + \mathbf{R}$;
$\mathbf{K}_t = \mathbf{\Sigma}_{t|t-1}\mathbf{C}^\mathsf{T}\mathbf{S}_t^{-1}$;
$\hat{\mathbf{z}}_t = \hat{\mathbf{z}}_{t|t-1} + \mathbf{K}_t\tilde{\mathbf{r}}_t$;
$\mathbf{\Sigma}_t = (\mathbf{I} - \mathbf{K}_t\mathbf{C})\,\mathbf{\Sigma}_{t|t-1}$;
return $\hat{\mathbf{z}}_t, \mathbf{\Sigma}_t$

Algorithm 2: Kalman Smoother

Input : $\mathbf{y}_{1:T}, \mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}$
Output: $\hat{\mathbf{z}}_{t|T}, \mathbf{\Sigma}_{t|T}$
for $t \leftarrow 1$ **to** T **by** *1* **do**
  Apply the Kalman Filter (refer to Algorithm 1);
end
return $\hat{\mathbf{z}}_{T|T}, \mathbf{\Sigma}_{T|T}$;
for $t \leftarrow T$ **to** *1* **by** *1* **do**
  $\mathbf{J}_t = \mathbf{\Sigma}_{t|t}\mathbf{A}^\mathsf{T}\mathbf{\Sigma}_{t+1|t}^{-1}$;
  $\hat{\mathbf{z}}_{t|T} = \hat{\mathbf{z}}_{t|t} + \mathbf{J}_t(\hat{\mathbf{z}}_{t+1|T} - \hat{\mathbf{z}}_{t+1|t})$;
  $\mathbf{\Sigma}_{t|T} = \mathbf{\Sigma}_{t|t} + \mathbf{J}_t(\mathbf{\Sigma}_{t+1|T} - \mathbf{\Sigma}_{t+1|t})\mathbf{J}_t^\mathsf{T}$;
end
return $\hat{\mathbf{z}}_{T:1|T}, \mathbf{\Sigma}_{T:1|T}, \mathbf{J}_{T:1}$;

song, the observed sequence $\hat{\mathbf{y}}_i = \mathbf{y}_{1:T}$ is then used in an LG-SSM to obtain the latent sequence $\hat{\mathbf{z}}_{1:T}$ that represents the song-level emotional dynamics, where T is the number of verses in the song.

Standard applications of LG-SSM assume a temporal ordering in the sequence. This means that estimates of $p(\hat{\mathbf{z}}_t)$ should only depend on the observed values up to the verse step t (i.e., $\mathbf{y}_{1:t}$), which is the central assumption of the Kalman Filter algorithm. Given the sequence of observations, we recursively apply the Kalman Filter to calculate the mean and variance of the hidden states, whose computation steps are displayed in Algorithm 1.

Since we have obtained initial predictions for all verses in a song, we can assume that observed emotion scores are available for the sequence of an entire song a priori. In other words, we can include the ''future'' data (i.e., $\mathbf{y}_{t+1:T}$) to estimate the latent posteriors $p(\hat{\mathbf{z}}_t)$. This is achieved by using the Kalman smoothing algorithm, also known as RTS smoother (Rauch et al., 1965), shown in Algorithm 2.

As opposed to most other algorithms, the Kalman Filter and Kalman Smoother algorithms are used with already known parameters. Hence, learning the SSM involves estimating the parameters Θ. If a set of ground truth values for the complete $\mathbf{z}_{1:T}$ is available, they can be learned using a Maximum Likelihood Estimation (MLE). If only the noisy, observed sequences $\mathbf{y}_{1:T}$ are present, the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) provides an iterative method for finding the MLEs of Θ by successively maximizing the conditional expectation of the complete data likelihood until convergence.

4 Experiments

Our experiments aim to evaluate the method proposed to predict the emotional dynamics of song lyrics without utilizing any annotated lyrics data. We introduce datasets, lexicon resources, and the evaluation metric used (§4.1), and discuss the implementation details and experiment settings of the verse-level model (§4.2) and the song-level model (§4.3).

4.1 Datasets and Evaluation

LyricsEmotions. This corpus was developed by Mihalcea and Strapparava (2012), consisting of 100 popular English songs with 4,975 verses in total. The number of verses for each song varies from 14 to 110. The LyricsEmotions dataset was constructed by extracting the parallel alignment of musical features and lyrics from MIDI tracks. These lyrics were annotated using Mechanical Turk at verse level with real-valued intensity scores ranging from 0 to 10 of six Ekman's emotions (Ekman, 1993): `ANGER`, `DISGUST`, `FEAR`, `JOY`, `SADNESS`, and `SURPRISE`. Given that our goal is to predict emotions without relying on song-level dynamics, *we use this dataset for evaluation purposes only.*

NewsHeadlines. To train the verse-level model, we employ the NewsHeadlines[3] dataset

[3] `http://web.eecs.umich.edu/ mihalcea /affectivetext/`.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

	Scope	Size (PT)	Label	Reference
NRC-Emo-Int	Emotion	1 (4)	Numerical	Mohammad (2018)
SentiWordNet	Sentiment	2 (10)	Numerical	Esuli and Sebastiani (2007)
NRC-Emo-Lex	Emotion	1 (4)	Nominal	Mohammad and Turney (2013)
NRC-Hash-Emo	Emotion	1 (4)	Numerical	Mohammad and Kiritchenko (2015)
Sentiment140	Sentiment	3 (20)	Numerical	Mohammad et al. (2013)
Emo-Aff-Neg	Sentiment	3 (20)	Numerical	Zhu et al. (2014)
Hash-Aff-Neg	Sentiment	3 (20)	Numerical	Mohammad et al. (2013)
Hash-Senti	Sentiment	3 (20)	Numerical	Kiritchenko et al. (2014)
DepecheMood	Emotion	8 (165)	Numerical	Staiano and Guerini (2014)

Table 1: Lexicons used to build lexicon-based feature vectors: PT is the size of feature vector after polynomial feature expansion.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

(Strapparava and Mihalcea, 2007), which is a collection of 1,250 news headlines. Each headline is annotated with six scores ranging from 0 to 100 for each of Ekman's emotions and one score ranging from -100 to 100 for valence.

Lexicons. Following Goel et al. (2017) and Meisheri and Dey (2018), we use nine emotion and sentiment related lexicons to obtain the feature vectors from the text in NEWSHEADLINES and LYRICSEMOTIONS, summarized in Table 1.

Evaluation. In line with Mihalcea and Strapparava (2012), we use the Pearson correlation coefficient (r) as the evaluation metric to measure the correlation between the predictions and ground truth emotion intensities. To assess statistical significance, we conduct the Williams test (Williams, 1959) in the differences between the Pearson correlations of each pair of models.

For baselines, our method is unsupervised at the song level, and we are not aware of previous work that tackles similar cases. Therefore, we use the results of the verse-level model as our main baseline. We argue that this is a fair baseline since the SSM-based model does not require additional data.

4.2 Verse-level Experiments

Setup. For the pre-trained model, we choose the BERT$_{base}$ uncased model in English with all parameters frozen during training. All models are trained on an NVIDIA T4 Tensor Core GPU with CUDA (version 11.2).

BERTLex. The sequence of token embeddings, including [CLS] and [SEP] at the output of the last layer of the BERT$_{base}$ model, is fed into a Conv1D layer with 128 filters and a kernel size of 3, followed by a 1D global average pooling layer.

We concatenate nine vector representations for every word in the established vocabulary by using the lexicons in Table 1 in the identical order to form a united feature vector. As a result, the whole word embedding has shape (3309, 25), where 3309 is the vocabulary size and 25 is the number of lexicon-based features. To validate the benefit of adding polynomial features, we also perform a polynomial expansion with a degree of 3, extending the shape of vector representations to (3309, 267). Then, static word embeddings are fed a Conv1D layer with 128 filters and a kernel size of 3, followed by a global max-pooling layer.

The two pooled vectors are then concatenated through a Concatenate layer. The verse-level emotion intensities are predicted by using a Linear layer with a single neuron[4] for regression.

Training. Instead of using the standard train/dev/test split of the NEWSHEADLINES dataset, we apply 10-fold cross-validation to tune the hyperparameters of BERT-based models. Empirically tuned hyperparameters are listed in Table 2 and are adopted in the subsequent experiments. After tuning, the final models using this set of hyperparameters are trained on the full NEWSHEADLINES data. We use an ensemble of five runs, taking the mean of the predictions as the final output.

[4]We experimented with a multi-task model that predicted all six emotions jointly, but preliminary results showed that building separate models for each emotion performed better.

Parameters	Value
Dropout rate	0.1
Optimizer	Adam
Learning rate	2e-5
β_1 / β_2	0.9 / 0.999
Batch size	32

Table 2: Hyperparameter settings of BERT and CNN models.

4.3 Song-Level Experiments

We apply the library `pykalman` (version 0.9.2),[5] which implements the Kalman Filter, the Kalman Smoother, and the EM algorithm to train SSMs. We fix the initial state mean as the first observed value in the sequence (i.e., each song's first verse-level prediction) and the initial state covariance as 2. We then conduct experiments with several groups of parameters transition matrices $\mathbf{A}$, transition covariance $\mathbf{Q}$, observation matrices $\mathbf{C}$, and observation covariance $\mathbf{R}$ to initialize the Kalman Filter and Kalman Smoother. For parameter optimization, we experiment n_iter = $\{1,3,5,7,10\}$ to control the number of EM algorithm iterations. Additionally, we apply 10-fold cross-validation when optimizing parameters via EM, which means each fold (containing 10 songs) is processed by a Kalman Filter or Kalman Smoother defined by the optimal parameters that we obtained from training on the other folds (containing 90 songs).

5 Results and Analysis

In this section, we first compare the results of our lexicon-based, learning-based and hybrid methods at the verse level (§ 5.1). We then provide the results of the song-level models and investigate the impact of the initial predictions from verse-level models, SSM parameters, and parameter optimization (§ 5.2). We additionally show the qualitative case analysis results to understand our model's abilities and shortcomings (§ 5.3). Finally, we compare the results of supervised and unsupervised methods on LyricsEmotions (§ 5.4).

[5] `https://github.com/pykalman/pykalman`.

5.1 Results of Verse-level Models

Table 3 shows the results of verse-level models on NewsHeadlines (average of 10-fold cross-validation) and LyricsEmotions (as a test set).[6]

The domain difference is significant in news and lyrics, as we can observe from the different performance of the BERT-based models on the two datasets. Overall, our BERTLex method outperforms the lexicon-only and BERT-only baselines and reaches the highest Pearson score (0.503, BERTLex^{poly} for `JOY`) in LyricsEmotions.

Having a closer look at the results of LyricsEmotions, we also observe the following:

- The addition of lexicons for incorporating external knowledge consistently promotes the performance of BERT-based models.
- BERTLex models that add polynomial feature expansion are better than those that do not, when using LyricsEmotions as a test set (except for `DISGUST`). However, in the cross-validation of NewsHeadlines, the models without polynomial features outperform those with.

5.2 Results of Song-level Models

Extensive experiments confirm that our song-level models utilizing the Kalman Filter and Kalman Smoother can improve the initial predictions from verse-level models (see Table 4 and Table 5). The LG-SSMs with EM-optimized parameters always perform better than those without using EM. Furthermore, the performance improvements of the strongest SSMs from their corresponding verse-level baselines are statistically significant at 0.05 confidence (marked with *), except for `SURPRISE`.

Theoretically, the Kalman Smoother is supposed to perform better than the Kalman Filter, since the former uses all observations in the whole sequence. According to our experimental results, however, the best-performing algorithm depends on emotion. Furthermore, applying EM consistently improves the results of SSMs that use the initial values, except for `SURPRISE`.

[6] We perform five runs with different random seeds, using the mean, median, maximum or minimum to pool the results. Here we show the result of the best pooling method, but in practice we did not see any significant difference compared to the mean pooling.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

	Dataset	ANG	DIS	FEA	JOY	SAD	SUR
Lexicon only	NH_{cv}	0.197	0.106	0.231	0.219	0.112	0.056
	LE_{cv}	0.212	0.091	0.185	0.209	0.175	0.031
BERT only	NH_{cv}	0.740	0.651	0.792	0.719	0.808	0.469
	LE_{test}	0.311	0.261	0.314	0.492	0.306	0.071
BERTLex	NH_{cv}	0.865	0.828	0.840	0.858	0.906	0.771
	LE_{test}	0.340	**0.287**	0.336	0.472	0.338	0.066
$BERTLex^{poly}$	NH_{cv}	0.838	0.788	0.833	0.840	0.885	0.742
	LE_{test}	**0.345**	0.268	**0.350**	**0.503**	**0.350**	**0.089**

Table 3: Pearson correlations between ground truth labels and predictions of the verse-level models in the NEWSHEADLINES (NH) and LYRICSEMOTIONS (LE) datasets: The subscript cv means the average results of the 10-fold cross-validation experiments, and the subscript $test$ means the results of using the dataset as the test set.

	ANG	DIS	FEA	JOY	SAD	SUR
BERTLex	0.338	0.280	0.336	0.468	0.338	0.066
Filter	0.359*	0.287*	0.352*	0.498*	0.361*	**0.069**
Smoother	0.362*	0.282	0.352*	0.501*	0.366*	0.064
Filter-EM	0.396*	**0.293***	**0.357***	0.522*	**0.387***	0.069
Smoother-EM	**0.405***	0.280	0.339	**0.522***	0.385*	0.060
$BERTLex^{poly}$	0.315	0.261	0.350	0.503	0.347	0.083
Filter	0.334*	0.267	0.367*	0.538*	0.374*	**0.088**
Smoother	0.332*	0.258	0.368*	0.542*	0.380*	0.082
Filter-EM	**0.358***	**0.270***	**0.371***	0.568*	0.405*	0.087
Smoother-EM	0.356*	0.251	0.355	**0.570***	**0.405***	0.079

Table 4: Pearson correlations between ground truth emotion intensities and predictions of BERTLex models and SSMs, respectively. The default parameters in `pykalman` are used: $\mathbf{A}$ = 1, $\mathbf{Q}$ = 1, $\mathbf{C}$ = 1, $\mathbf{R}$ = 5, and n_iter = 10.

Combining the results in Table 3, Table 4, and Table 5, we observe that all models perform poorly when predicting the emotion intensities of SURPRISE ($r < 0.1$). The overall worst results for SURPRISE can also be observed from other work in LYRICSEMOTIONS and NEWSHEADLINES as well as similar work in different datasets annotated with the Ekman taxonomy. SURPRISE has significantly lower inter-annotator agreement than other emotions (Strapparava and Mihalcea, 2007; Schuff et al., 2017; Buechel and Hahn, 2017; Dang et al., 2021; Edmonds and Sedoc, 2021), which implies that SURPRISE is especially difficult to model and occurs less frequently (Mohammad et al., 2018; Bostan and Klinger, 2018). This might indicate the underlying problems in the definition of SURPRISE as an emotion category (Schuff et al., 2017).

Impact of Verse-level Predictions. The performance of applying Kalman Filter, Kalman Smoother, and EM algorithm are associated with the initial scores predicted by verse-level models. For the same emotion, we compare the results based on the mean predictions of the BERTLex models with and without polynomial expansion on lexical features, respectively (shown in Table 4). We observe that the higher the Pearson correlation between the ground truth and the verse-level predictions, the more accurate the estimates obtained after using LG-SSMs accordingly. The strongest SSMs also differ with the different types

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

	ANG	DIS	FEA	JOY	SAD	SUR
$\mathbf{A} = 0.5$						
Filter	0.265	0.223	0.350	0.455	0.351	0.074
Smoother	0.277	0.223	0.351	0.471	0.362*	0.071
Filter-EM	0.357*	**0.273***	**0.373***	0.563*	0.397*	**0.089**
Smoother-EM	**0.360***	0.262	0.364*	**0.569***	**0.402***	0.083
$\mathbf{A} = 2$						
Filter	0.354*	0.270	0.369*	0.560*	0.393*	0.085
Smoother	0.075	0.055	0.160	0.205	0.174	0.003
Filter-EM	0.355*	**0.272***	**0.375***	0.562*	0.399*	**0.089**
Smoother-EM	**0.358***	0.260	0.364*	**0.568***	**0.403***	0.083

Table 5: Pearson correlations between ground truth and SSMs with different values of transition matrices $\mathbf{A}$, based on BERTLex^{poly} models (as listed in the bottom half of Table 4). The other parameters are fixed as $\mathbf{Q} = 1$, $\mathbf{C} = 1$, $\mathbf{R} = 5$, and n_iter = 5.

of emotions and initial predictions, as denoted in boldface.

Impact of Initial Parameters. The results of Kalman Filter and Kalman Smoother are sensitive to the initial model parameters. As displayed in Table 5, when we only change the value of transition matrices $\mathbf{A}$ and fix the other parameters, running either the Filter or Smoother can actually decrease the performance. Fortunately, this kind of diminished performance can be diluted by optimizing the parameters with an EM algorithm.

Impact of Parameter Optimization. For either Kalman Filter or Kalman Smoother, using EM to optimize the parameters increases Pearson's r in most cases. Through experiments, the number of iterations does not significantly influence the performance of the EM algorithm, and $5 \sim 10$ iterations usually produce the strongest results.

5.3 Qualitative Case Studies

Domain Discrepancy. As displayed in Section 5.2, the Pearson scores between the ground-truth labels and estimates of SURPRISE are lower than 0.1, which means our verse-level and song-level models both underperform in predicting this emotion. Upon closer inspection, we observe that there are a great number of zeros in the ground-truth annotations of SURPRISE in the target domain dataset. For example, Figure 3 shows the emotion curves of *If You Love Somebody Set Them Free* by Sting, where all the ground-truth labels of SURPRISE are zeros in the whole song. Statistically, there are 1,933 zeros

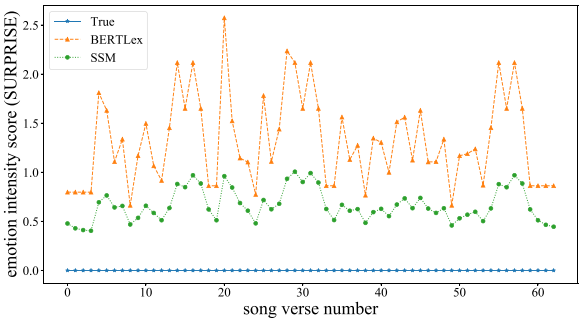

Figure 3: The SURPRISE emotion intensities of ground truth (all zeros), BERTLex model, and SSM in an example song.

out of 4,975 (38.85%) SURPRISE ground truth labels in LyricsEmotions but only 148 of 1,250 (11.84%) zeros in NewsHeadlines. The models trained on NewsHeadlines would not assume such a large absences of SURPRISE when predicting for LyricsEmotions. This domain discrepancy clearly affects the performance of our method.

Characteristics of Kalman Filter and Kalman Smoother. Our experiments indicate that initial predictions of at least 50 to 70 of 100 songs have been enhanced after modeling them with LG-SSMs. We summarize two trending types from the emotional dynamics of the songs whose predictions are weakened by LG-SSMs. One is that the ground truth emotional dynamics fluctuate more sharply than those of the verse-level predictions, as displayed in the first and the second sub-figures in Figure 4. The other is the opposite that verse-level models produce an emotion intensity curve with more sudden changes than the

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

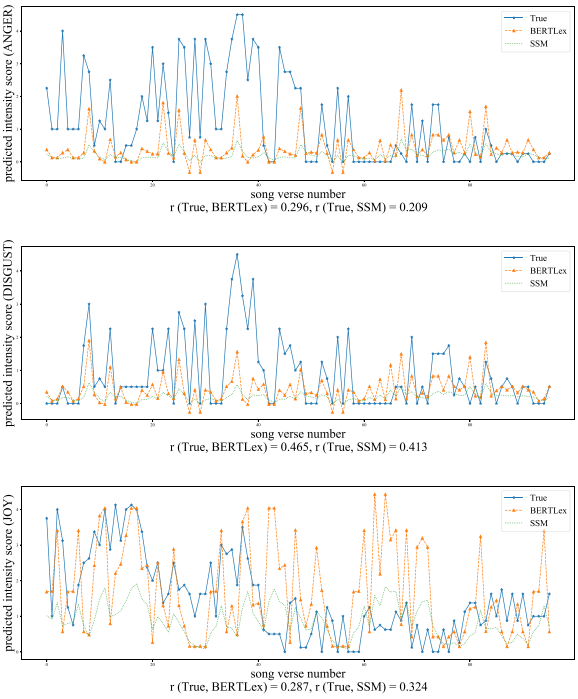

Figure 4: Emotional dynamics of ANGER, DISGUST and SURPRISE in *Bad Romance* by Lady Gaga: Pearson's r between ground truth and predictions of BERTLex^{poly}, estimates of Kalman Filter, are reported, respectively.

ground truth (see the third sub-figure in Figure 4). The emotional dynamics trend of estimates by song-level models is similar to verse-level models. Due to the Gaussian assumption, Kalman Filter and Kalman Smoother tend to flatten or smooth the curves of verse-level predictions. This means that applying LG-SSMs can somewhat reduce errors in the second type of emotion dynamic curves. For the first type, however, the Kalman Filter and Kalman Smoother make the results worse, as smoother estimations are not desirable in this situation.

Using Text Solely. The lyrics in LyricsEmotions are synchronized with acoustic features, where some verses with identical text are labeled as different emotional intensities. For instance, in Table 6, the verse *''When it rain and rain, it rain and rain''* repeats multiple times in the song *Rain* by Mika, and their ground truth SADNESS labels differ in different verses due to the melody. However, the verse-level models can only produce the same predictions since these verses contain the same text, and the models do not consider the context of the whole song. Consequently, the emotion scores of different verses

Verse ID	Truth	BERTLex	Smoother-EM
s55v15	4.33	8.65	1.68
s55v31	7.66	8.65	1.68
s55v32	7.33	8.65	1.63

Table 6: SADNESS scores of verses with the same lyrics verse *''When it rain and rain, it rain and rain''* but different ground truth labels in the song.

	ANG	DIS	FEA
BERTLex	0.837	0.736	0.790
SSM	0.405	0.293	0.375
	JOY	SAD	SUR
BERTLex	0.879	0.831	0.739
SSM	0.570	0.405	0.089

Table 7: Pearson correlations of predictions from supervised BERTLex models (10-fold cross validation) and predictions of the best SSMs.

predicted by LG-SSMs are close, as the results of song-level models are highly related to the initial predictions from BERTLex.

5.4 Comparison with a Supervised Model

Our last experiment aims to understand the degree of difficulty in solving the task by training a supervised model, which serves as a performance upper bound. We keep the same 10-fold cross-validation splits, but now use the training folds to fine-tune a BERTLex model at the verse level.

We compare the results of the supervised model with our best results of song-level models in Table 7, showing there is still a substantial performance gap in all emotions. In particular, the supervised model shows strong numbers for SURPRISE, the most challenging emotion to predict in our experiments. While our SSM models have the benefit of being readily applicable to new domains (such as songs in genres other than pop and languages other than English), this result demonstrates that practical systems could benefit with some level of annotations for SURPRISE. More generally, it also motivates extending SSMs to a semi-supervised setting, which we leave for future work.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

6 Conclusion and Future Work

This paper presents a two-stage BERTLex-SSM framework for sequence-labeling emotion intensity recognition tasks, especially in label-scarce scenarios. Combining the contextualized embeddings with static word embeddings and then modeling the initial predicted intensity scores as a State Space Model, our method can utilize context-sensitive features with external knowledge and capture the emotional dynamic transitions. Experimental results show that our proposed BERTLex-SSM effectively predicts emotion intensities in the lyrics without requiring annotated lyrics data.

Our findings and analysis point to a range of directions for future work:

Domain Adaptation. While our method could apply any general verse-level model, including a pure lexicon-based one, in practice, we obtained the best results by leveraging annotated sentence-level datasets. This naturally leads a domain discrepancy: in our particular case, between news and lyrics domains. Given that unlabeled song lyrics are relatively easy to obtain, one direction is to incorporate unsupervised domain adaptation techniques (Ramponi and Plank, 2020) to improve the performance of the verse-level model. Semi-supervised learning (similar to Täckström and McDonald, 2011b) is another promising direction, although methods would need to be modified to incorporate the continuous nature of the emotion labels.

SSM Extensions. Despite being able to optimize the estimates through Kalman Filter and Kalman Smoother, the simplicity of the LG-SSM makes it difficult to deal with the wide variations in emotion space dynamics, given that it is a linear model. We hypothesize that non-linear SSM extensions (Julier and Uhlmann, 1997; Ito and Xiong, 2000; Julier and Uhlmann, 2004) might be a better fit for modeling emotion dynamics.

Multimodal Grounding. Since the LyricsEmotions dataset is annotated on parallel acoustic and text features, using lyrics solely as the feature can cause inconsistencies in the model. Extending our method to a multi-modal setting would remedy this issue when the identical lyrics are companions with different musical features to appear in various verses. Taking the knowledge of song structure (e.g., Intro - Verse - Bridge - Chorus) into account has the potential to advance the modeling of emotion dynamics, assuming the way (up or down) that emotion intensities change is correlated with which part of the song the verses locate.

Acknowledgments

The authors would like to thank Rada Mihalcea for sharing the LyricsEmotions dataset with us and the anonymous reviewers and editors for their constructive and helpful comments.

References

Yudhik Agrawal, Ramaguru Guru Ravi Shanker, and Vinoo Alluri. 2021. Transformer-based approach towards music emotion recognition from lyrics. In *Advances in Information Retrieval, 43rd European Conference on IR Research (ECIR 2021)*, pages 167–175, Cham, Switzerland. Springer. https://doi.org/10.1007/978-3-030-72240-1_12

Israa Alghanmi, Luis Espinosa Anke, and Steven Schockaert. 2020. Combining BERT with static word embeddings for categorizing social media. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 28–33, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.wnut-1.5

Mireille Besson, Frederique Faita, Isabelle Peretz, A.-M. Bonnel, and Jean Requin. 1998. Singing in the brain: Independence of lyrics and tunes. *Psychological Science*, 9(6):494–498. https://doi.org/10.1111/1467-9280.00091

Laura-Ana-Maria Bostan 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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-2092

Marcelo Caetano, Athanasios Mouchtaris, and Frans Wiering. 2012. The role of time in music emotion recognition: Modeling musical emotions from time-varying music features. In *International Symposium on Computer Music Modeling and Retrieval*, pages 171–196. Springer. https://doi.org/10.1007/978-3-642-41248-6_10

Bao Minh Doan Dang, Laura Oberländer, and Roman Klinger. 2021. Emotion stimulus detection in German news headlines. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 73–85, Düsseldorf, Germany. KONVENS 2021 Organizers.

Luna De Bruyne, Orphee De Clercq, and Veronique Hoste. 2021. Emotional RobBERT and insensitive BERTje: Combining transformers and affect lexica for Dutch emotion detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 257–263, Online. Association for Computational Linguistics.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

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.

Darren Edmonds and João Sedoc. 2021. Multi-emotion classification for song lyrics. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 221–235, Online. Association for Computational Linguistics.

Paul Ekman. 1993. Facial expression and emotion. *American Psychologist*, 48(4):384. https://doi.org/10.1037/0003-066X.48.4.384, PubMed: 8512154

Andrea Esuli and Fabrizio Sebastiani. 2007. Sentiwordnet: A high-coverage lexical resource for opinion mining. *Evaluation*, 17(1):26.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4441–4453, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.393

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.224

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1435–1449, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.124

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1015

Pranav Goel, Devang Kulshreshtha, Prayas Jain, and Kaushal Kumar Shukla. 2017. Prayas at EmoInt 2017: An ensemble of deep neural architectures for emotion intensity prediction in tweets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 58–65, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5207

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1280

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1193, PubMed: 32219222

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7042–7052, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.547

Xiangdong Huang, Minjie Ren, Qiankun Han, Xiaoqi Shi, Jie Nie, Weizhi Nie, and An-An Liu. 2021. Emotion detection for conversations based on reinforcement learning framework. *IEEE MultiMedia*, 28(2):76–85. https://doi.org/10.1109/MMUL.2021.3065678

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7360–7370, Online. Association for Computational Linguistics.

Kazufumi Ito and Kaiqi Xiong. 2000. Gaussian filters for nonlinear filtering problems. *IEEE Transactions on Automatic Control*, 45(5):910–927. https://doi.org/10.1109/9.855552

Wenxiang Jiao, Michael Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 05, pages 8002–8009. https://doi.org/10.1609/aaai.v34i05.6309

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

Simon J. Julier and Jeffrey K. Uhlmann. 1997. New extension of the kalman filter to nonlinear systems. In *Signal Processing, Sensor Fusion, and Target Recognition VI*, volume 3068, pages 182–193. International Society for Optics and Photonics.

Simon J. Julier and Jeffrey K. Uhlmann. 2004. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422. https://doi.org/10.1109/JPROC.2003.823141

Patrik N. Juslin and Petri Laukka. 2004. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3):217–238. https://doi.org/10.1080/0929821042000317813

Rudolph Emil Kalman. 1960. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45. https://doi.org/10.1115/1.3662552

Seungyeon Kim, Joonseok Lee, Guy Lebanon, and Haesun Park. 2015. Estimating temporal dynamics of human emotions. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25–30, 2015, Austin, Texas, USA*, pages 168–174. AAAI Press.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

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 *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, volume 86, pages 255–266.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings*.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762. https://doi.org/10.1613/jair.4272

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.alw-1.5

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 – July 1, 2001*, pages 282–289. Morgan Kaufmann.

Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. HiTrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4190–4200, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Zheng Lian, Jianhua Tao, Bin Liu, Jian Huang, Zhanlei Yang, and Rongjun Li. 2020. Conversational emotion recognition using self-attention mechanisms and graph neural networks. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China,*

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

25–29 October 2020, pages 2347–2351. ISCA. https://doi.org/10.21437/Interspeech.2020-1703

Xin Lu, Yanyan Zhao, Yang Wu, Yijian Tian, Huipeng Chen, and Bing Qin. 2020. An iterative emotion interaction network for emotion recognition in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4078–4088, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.360

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825. https://doi.org/10.1609/aaai.v33i01.33016818

Ricardo Malheiro, Renato Panda, Paulo Gomes, and Rui Pedro Paiva. 2016. Emotionally-relevant features for classification and regression of music lyrics. *IEEE Transactions on Affective Computing*, 9(2):240–254. https://doi.org/10.1109/TAFFC.2016.2598569

Andrew McCallum, Dayne Freitag, and Fernando C. N. Pereira. 2000. Maximum entropy Markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 – July 2, 2000*, pages 591–598. Morgan Kaufmann.

Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 432–439, Prague, Czech Republic. Association for Computational Linguistics.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17. https://doi.org/10.1109/T-AFFC.2011.20

Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh Eetemadi. 2020. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1184–1189. IEEE. https://doi.org/10.1109/ICDM50108.2020.00146

Hardik Meisheri and Lipika Dey. 2018. TCS research at SemEval-2018 task 1: Learning robust representations using multi-attention architecture. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 291–299, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1043

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. 2018. Word affect intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1001

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327.

Saif M. Mohammad and Felipe Bravo-Márquez. 2017. WASSA-2017 shared task on emotion intensity. In *8th Workshop on Computational Approaches to Subjectivity, Sentiment and*

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

Social Media Analysis WASSA 2017: Proceedings of the Workshop, pages 34–49. The Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5205

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326. https://doi.org/10.1111/coin.12024

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Kevin P. Murphy. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 807–814.

Debora Nozza, Elisabetta Fersini, and Enza Messina. 2017. A multi-view sentiment corpus. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 273–280, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-1026

Andrew Ortony, Gerald L. Clore, and Allan Collins. 1988. The cognitive structure of emotions. https://doi.org/10.1017/CBO9780511571299

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Theories of Emotion*, pages 3–33. Elsevier.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1081

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1050

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953. https://doi.org/10.1109/ACCESS.2019.2929050

Lawrence Rabiner and Biinghwang Juang. 1986. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16. https://doi.org/10.1109/MASSP.1986.1165342

Arun Ramachandran and Gerard de Melo. 2020. Cross-lingual emotion lexicon induction using representation alignment in low-resource settings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5879–5890, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.517

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Herbert E. Rauch, F. Tung, and Charlotte T. Striebel. 1965. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450. https://doi.org/10.2514/3.3166

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866. https://doi.org/10.1162/tacl_a_00349

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178. https://doi.org/10.1037/h0077714

Erik M. Schmidt and Youngmoo E. Kim. 2010. Prediction of time-varying musical mood distributions using kalman filtering. In *2010 Ninth International Conference on Machine Learning and Applications*, pages 655–660. IEEE. https://doi.org/10.1109/ICMLA.2010.101

Erik M. Schmidt and Youngmoo E. Kim. 2011. Modeling musical emotion dynamics with conditional random fields. In *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR*, pages 777–782. Miami (Florida), USA.

Erik M. Schmidt, Douglas Turnbull, and Youngmoo E. Kim. 2010. Feature selection for content-based, time-varying musical emotion regression. In *Proceedings of the 11th ACM SIGMM International Conference on Multimedia Information Retrieval*, pages 267–274. https://doi.org/10.1145/1743384.1743431

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modeling 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, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5203

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1551–1560, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.123

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Jacopo Staiano and Marco Guerini. 2014. Depeche mood: A lexicon for emotion analysis from crowd annotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–433, Baltimore, Maryland. Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-2070

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics. https://doi.org/10.3115/1621474.1621487

Oscar Täckström and Ryan McDonald. 2011a. Discovering fine-grained sentiment with latent variable structured prediction models. In *European Conference on Information Retrieval*, pages 368–374. Springer.

Oscar Täckström and Ryan McDonald. 2011b. Semi-supervised latent variable models for sentence-level sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 569–574, Portland, Oregon, USA. Association for Computational Linguistics.

Evan J. Williams. 1959. The comparison of regression variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):396–399. https://doi.org/10.1111/j.2517-6161.1959.tb00346.x

Bin Wu, Erheng Zhong, Andrew Horner, and Qiang Yang. 2014. Music emotion recognition by multi-label multi-layer multi-instance multi-view learning. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 117–126.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025

Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Wenhao Ying, Rong Xiang, and Qin Lu. 2019. Improving multi-label emotion classification by integrating both general and domain-specific knowledge. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 316–321, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-5541

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10–16, 2019*, pages 5415–5421. ijcai.org.

Ke Zhang, Yuanqing Li, Jingyu Wang, Erik Cambria, and Xuelong Li. 2021. Real-time video emotion recognition based on reinforcement learning and domain knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1034–1047. https://doi.org/10.1109/TCSVT.2021.3072412

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware BERT for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. Cauain: Causal aware interaction network for emotion recognition in conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23–29 July 2022*, pages 4524–4530. ijcai.org. https://doi.org/10.24963/ijcai.2022/628

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176, Hong Kong, China. Association for Computational Linguistics.

Deyu Zhou, Shuangzhi Wu, Qing Wang, Jun Xie, Zhaopeng Tu, and Mu Li. 2020. Emotion classification by jointly learning to lexiconize and classify. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8–13, 2020*, pages 3235–3245. International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.288

Xiaodan Zhu, Svetlana Kiritchenko, and Saif Mohammad. 2014. NRC-Canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 443–447, Dublin, Ireland. Association for Computational Linguistics.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00541/2070877/tacl_a_00541.pdf by guest on 21 January 2025