

Lecture Notes in Electrical Engineering 940

Jimson Mathew
G. Santhosh Kumar
Deepak P.
Joemon M. Jose *Editors*

Responsible Data Science

Select Proceedings of ICDSE 2021

 Springer

Lecture Notes in Electrical Engineering

Volume 940

Series Editors

Leopoldo Angrisani, Department of Electrical and Information Technologies Engineering, University of Napoli Federico II, Naples, Italy
Marco Arteaga, Departament de Control y Robótica, Universidad Nacional Autónoma de México, Coyoacán, Mexico
Bijaya Ketan Panigrahi, Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, Delhi, India
Samarjit Chakraborty, Fakultät für Elektrotechnik und Informationstechnik, TU München, Munich, Germany
Jiming Chen, Zhejiang University, Hangzhou, Zhejiang, China
Shanben Chen, Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
Tan Kay Chen, Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore
Rüdiger Dillmann, Humanoids and Intelligent Systems Laboratory, Karlsruhe Institute for Technology, Karlsruhe, Germany
Haibin Duan, Beijing University of Aeronautics and Astronautics, Beijing, China
Gianluigi Ferrari, Università di Parma, Parma, Italy
Manuel Ferre, Centre for Automation and Robotics CAR (UPM-CSIC), Universidad Politécnica de Madrid, Madrid, Spain
Sandra Hirche, Department of Electrical Engineering and Information Science, Technische Universität München, Munich, Germany
Faryar Jabbari, Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA
Limin Jia, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China
Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
Alaa Khamis, German University in Egypt El Tagamoa El Khames, New Cairo City, Egypt
Torsten Kroeger, Stanford University, Stanford, CA, USA
Yong Li, Hunan University, Changsha, Hunan, China
Qilian Liang, Department of Electrical Engineering, University of Texas at Arlington, Arlington, TX, USA
Ferran Martín, Departament d'Enginyeria Electrònica, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain
Tan Cher Ming, College of Engineering, Nanyang Technological University, Singapore, Singapore
Wolfgang Minker, Institute of Information Technology, University of Ulm, Ulm, Germany
Pradeep Misra, Department of Electrical Engineering, Wright State University, Dayton, OH, USA
Sebastian Möller, Quality and Usability Laboratory, TU Berlin, Berlin, Germany
Subhas Mukhopadhyay, School of Engineering and Advanced Technology, Massey University, Palmerston North, Manawatu-Wanganui, New Zealand
Cun-Zheng Ning, Department of Electrical Engineering, Arizona State University, Tempe, AZ, USA
Toyooki Nishida, Graduate School of Informatics, Kyoto University, Kyoto, Japan
Luca Oneto, Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genova, Genova, Italy
Federica Pascucci, Dipartimento di Ingegneria, Università degli Studi "Roma Tre", Rome, Italy
Yong Qin, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China
Gan Woon Seng, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore
Joachim Speidel, Institute of Telecommunications, Universität Stuttgart, Stuttgart, Germany
Germano Veiga, Campus da FEUP, INESC Porto, Porto, Portugal
Haitao Wu, Academy of Opto-electronics, Chinese Academy of Sciences, Beijing, China
Walter Zamboni, DIEM—Università degli Studi di Salerno, Fisciano, Salerno, Italy
Junjie James Zhang, Charlotte, NC, USA

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering—quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact leontina.dicecco@springer.com.

To submit a proposal or request further information, please contact the Publishing Editor in your country:

China

Jasmine Dou, Editor (jasmine.dou@springer.com)

India, Japan, Rest of Asia

Swati Meherishi, Editorial Director (Swati.Meherishi@springer.com)

Southeast Asia, Australia, New Zealand

Ramesh Nath Premnath, Editor (ramesh.premnath@springernature.com)

USA, Canada:

Michael Luby, Senior Editor (michael.luby@springer.com)

All other Countries:

Leontina Di Cecco, Senior Editor (leontina.dicecco@springer.com)

**** This series is indexed by EI Compendex and Scopus databases. ****

Jimson Mathew · G. Santhosh Kumar · Deepak P. ·
Joemon M. Jose
Editors

Responsible Data Science

Select Proceedings of ICDSE 2021


 Springer

Editors

Jimson Mathew
Department of Computer Science
and Engineering
Indian Institute of Technology Patna
Patna, Bihar, India

G. Santhosh Kumar
Department of Computer Science
Cochin University of Science
and Technology
Cochin, Kerala, India

Deepak P.
School of Electronics, Electrical
Engineering and Computer Science
Queen's University Belfast
Belfast, UK

Joemon M. Jose 
School of Computing Science
University of Glasgow
Glasgow, UK

ISSN 1876-1100

ISSN 1876-1119 (electronic)

Lecture Notes in Electrical Engineering

ISBN 978-981-19-4452-9

ISBN 978-981-19-4453-6 (eBook)

<https://doi.org/10.1007/978-981-19-4453-6>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.

The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

End-to-End Hierarchical Approach for Emotion Detection in Short Texts	1
Georgios Hadjiharalambous, Kacper Beisert, and Joemon M. Jose	
Towards an Enhanced Understanding of Bias in Pre-trained Neural Language Models: A Survey with Special Emphasis on Affective Bias	13
Anoop K., Manjary P. Gangan, Deepak P., and Lajish V. L.	
Exploring Rawlsian Fairness for K-Means Clustering	47
Stanley Simoes, Deepak P., and Muiris MacCarthaigh	
Hybrid Explainable Educational Recommender Using Self-attention and Knowledge-Based Systems for E-Learning in MOOC Platforms	61
Mehbooba P. Shareef, Linda Rose Jimson, and Babita R. Jose	
An Improved Recommendation System with Aspect-Based Sentiment Analysis	75
Seema Safar, Babita R. Jose, and T. Santhanakrishnan	
Exploring Biomarker Identification and Mortality Prediction of COVID-19 Patients Using ML Algorithms	89
Rajan Singh and Prashant K. Srivastava	
COVID-19 Cases Prediction Based on LSTM and SIR Model Using Social Media	111
Aakansha Gupta and Rahul Katarya	
Joint Geometrical and Statistical Alignment Using Triplet Loss for Deep Domain Adaptation	119
R. Satya Rajendra Singh, Rakesh Kumar Sanodiya, and P. V. Arun	

Virtual Try-On Using Style Transfer	131
Ravi Ranjan Prasad Karn, Rakesh Kumar Sanodiya, Eswara Surya Chandaluri, S. Suryavardan, L Ranajith Reddy, and Leehter Yao	
Attention Mechanism in Convolutional Recurrent Neural Network for Improving Recognition Accuracy in Printed Devanagari Text	141
Shaheera Saba Mohd Naseem Akhter and Priti P. Rege	
Joint Learning for Multitasking Models	155
Ajai John Chemmanam and Bijoy A. Jose	
A CNN Approach for Detecting Red and White Lesions in Retinal Fundus Images	169
Rajesh Kumar and K. V. Pramod	
Predicting IMDB Movie Ratings Using RoBERTa Embeddings and Neural Networks	181
Anagha Jose and Sandhya Harikumar	
Domain-Specific Type-Safe APIs for Hierarchical Scientific Data with Modern C++	191
William F. Godoy, Addi Malviya Thakur, and Steven E. Hahn	
Kernelized Transfer Joint Matching for Unsupervised Domain Adaptation	205
A. K. Devika, Rakesh Kumar Sanodiya, and Babita R. Jose	

About the Editors

Jimson Mathew is currently a professor in the Department of the Computer Science and Engineering, Indian Institute of Technology Patna, India. He received a master's in computer engineering from Nanyang Technological University, Singapore, and a Ph.D. degree in computer engineering from the University of Bristol, Bristol, UK. He has held positions with the Centre for Wireless Communications, the National University of Singapore, Bell Laboratories Research Lucent Technologies North Ryde, Australia, Royal Institute of Technology KTH, Stockholm, Sweden, and Department of Computer Science, University of Bristol, UK. He is a Senior Member of IEEE. He has previously served as Guest Editor for ACM TECS. He also regularly serves on the program committee of top international conferences and holds multiple patents. His research interests include fault-tolerant computing, computer vision, machine learning, and IoT systems.

G. Santhosh Kumar is a full Professor at the Department of Computer Science, Cochin University of Science and Technology, Kerala, India. His research interests include cyber-physical systems, machine learning, and natural language processing. He is a senior member of the IEEE and the ACM, published several publications, and co-authored a book on Data Science.

Deepak P. is an Associate Professor of Computer Science at Queen's University Belfast (UK) and an adjunct faculty member at IIT Madras (India). His research interests include ethics for machine learning, natural language processing, and information retrieval. He is a senior member of the IEEE and the ACM and has authored over 100 publications, authored/edited three books, and is an inventor on over 10 patents.

Joemon M. Jose has been an active researcher in information retrieval (IR) since 1993 and has published over 300 journal and conference articles on information retrieval. He, along with co-authors, has received best paper/student paper awards at leading conferences, including ACM SIGIR, IiX, CHIIR, MMM, and the BCS ECIR. He has supervised, as primary supervisor, 20 Ph.D. students and over 20 RAs

and postdoctoral researchers. He has chaired several conferences, was one of the program committee chairs for the ECIR 2017 and 2020 conferences, regularly acts as a primary reviewer for A/A* conferences, and has attracted over 3M pounds in research funding.

End-to-End Hierarchical Approach for Emotion Detection in Short Texts



Georgios Hadjiharalambous, Kacper Beisert, and Joemon M. Jose

1 Introduction

Emotion significantly affects our decision-making process and plays an important role in our daily lives. As large amounts of textual documents are being created and circulated, there is a need to understand the sentiments and emotional orientation of the text. Real-life applications, such as reputation management, human–computer interaction, or understanding social responses to events [1], would benefit from emotion classification. Hence, the task of automatic identification of distinct emotions expressed in a text has been gaining increased attention by researchers [2–7]. The focus of emotion classification is to classify each sentence into one or several categories within a predefined emotion set.

In the last decade, a considerable amount of research has been directed at detecting sentiments in text, and several effective approaches have been developed [1]. However, emotion classification continues to pose a significantly greater challenge, especially in short texts. Social media streams, e.g., Twitter, generate vast amounts of real-time data, making them an important study area. Tweets are a maximum of 280 characters in length and are considered to capture meaningful and informative messages of the sender, which often contain indicators of emotions expressed by their senders. Detecting emotions in such short texts is a challenging issue, and few studies have aimed at it so far [1, 8, 9]. However, current approaches fail to effectively detect emotions from short texts due to the linguistic incompleteness of short social media texts. We argue that it is essential to develop an end-to-end model to identify

G. Hadjiharalambous · K. Beisert · J. M. Jose (✉)
School of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK
e-mail: joemon.jose@glasgow.ac.uk

G. Hadjiharalambous
e-mail: giorgos.hadjichi@gmail.com

K. Beisert
e-mail: kacper.beisert@gmail.com

the dominant emotions in short texts, and for this, we need to capture the context, sequential nature of the text, and the hidden linguistic nuances of expression. This paper developed a hierarchical approach to detecting emotions in short social media textual documents. Our contributions are as follows:

1. We have developed an end-to-end framework in which sentiments and emotions are detected in a hierarchical fashion. The framework can incorporate and use any state-of-the-art text understanding model.
2. We have tested our approach on four publicly available datasets and have shown that sentiment classification leads to more effective emotion classification.

2 Emotion Detection

To represent emotion, a number of models are proposed among which the most prominent one is Ekman’s universal emotion model [10], where six basic emotion types were identified: *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*. Since then, a number of alternative models are being proposed, for example, the circumplex model of Russel proposes two dimensions of mood: *valence and activity* and the OCC model suggests 22 emotional states. Similarly, Bollen et al. [11] extracted six dimensions of affections including *tension*, *depression*, *anger*, *vigor*, *fatigue*, and *confusion*. These different models highlight the variability in the application of textual emotion identification and potential ambiguity in identifying emotional states especially those in textual documents. The challenges in detecting and classifying emotions in social media text include casual writing style of text; semantic ambiguity of text messages; fuzzy boundaries of emotion classes; difficulty of generating labels; numerous emotional states; and inconsistent annotations [12].

Conventional machine learning approaches have been widely used to detect sentiment and emotions in text documents [9, 12]. In general, these approaches follow a similar pattern: Given a collection of documents, handcrafted features are extracted and models are trained and tested. Typically used features include bag-of-words features, lexicon, attributes from Knowledge Bases (KBs), and syntactic features of a given text (sentence or document). Wang et al. [9] used n-gram and part-of-speech features as well as several different lexicons and show that best results are produced when unigram, bigram, lexicons, and part-of-speech features are combined together.

Recently, a number of deep learning models have been applied to sentiment analysis and emotion classification. Recurrent Neural Network models and their variants have been applied extensively for emotion classification tasks, especially to capture sequential nature of text and the relationships between different words. Kumar et al. introduced a model [13] in which they claim emotion analysis improved sentiment classification. They used a distributed thesaurus to identify similar words and then used a two-layer attention model with a bidirectional LSTM representation. Unlike their work, our aim is to detect emotion in short text. Similarly, Gazi et al. [14] proposed a hierarchical approach involving emotion, neutrality, and polarity. However,

they worked with a two-stage process whereas we model this as an end-to-end deep learning approach. In addition, our experiments cover four large datasets, whereas Ghazi et al. used two smaller datasets.

In [8], Seyeditabari et al. proposed a GRU-based classifier for emotion detection from short text. They used fastText word embedding for each word in the input text and a bidirectional GRU to produce an intermediate representation. They used a max-pooling layer to extract the most important features from the GRU output and an average-pooling layer to create a generalized hidden representation for the text as a whole. The outputs of these layers were concatenated together to produce a final representation, which was then passed through a feed forward network with a final sigmoid layer to generate the classification output. Their experiments have shown superior performance for their proposed model over the state-of-the-art models in identifying emotion from short texts. However, their approach is used to train seven separate binary classifiers, which are then trained in a One-Versus-Rest fashion. However, we argue that we need an end-to-end classifier which can approach this problem from a multiclass perspective. Similarly, Chauhan et al. [15] proposed a multi-attention mechanism for detecting emotion from a multi-modal collection.

Summary: State-of-the-art approaches for short text emotion classification are limited [8, 16, 17] and have the following limitations: The performance of these models are poor; they fail to capture the context and sequential nature of text effectively [9]; and they often use a binary classification (instead of multiclass classification) approach to emotion detection [8, 13].

3 Hierarchical Approach

Given the fact that sentiment classification can achieve 80+% accuracy, it is our conjecture that sentiment classification will improve emotional classification, where current models' performance is less than 60% [9]. Hence, we propose a hierarchical approach: The first layer detects the sentiments and the second layer detects the emotions. The proposed hierarchical approach to text emotion detection is based on the idea that basic emotions can be grouped into more general class sets depending on their inherent semantic features (e.g., positive and negative emotions). These class sets and their component emotion classes can then be arranged hierarchically based on the semantic relationships between them. The hierarchical emotion structure used in our research as an abstract guide¹ is shown in Fig. 1.

Essentially, the hierarchical approach aims to exploit sentiment analysis to improve the effectiveness of emotion detection. Instead of classifying the input text directly into one of the considered emotion classes, the proposed method follows the hierarchical structure by first identifying its sentiment. The input text is then assigned into one of the emotion classes belonging to that branch of the hierarchy. If the input's sentiment is categorized as positive, then it is classified into one of the

¹ Our main experiments use only one positive emotion class.

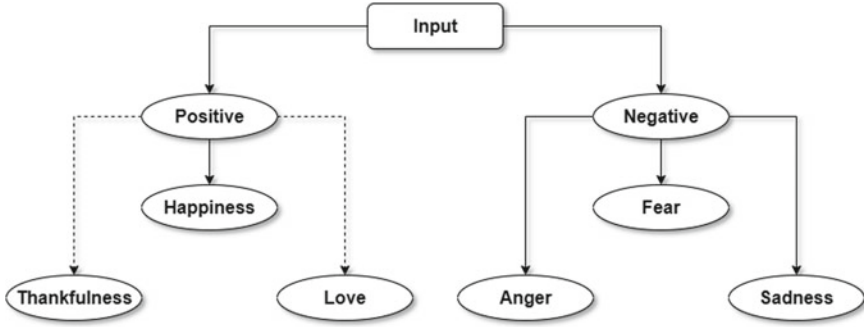


Fig. 1 The proposed hierarchical structure for emotion

positive classes. Conversely, if the input's sentiment is categorized as negative, then it is classified into one of the negative classes. By following the hierarchical structure, we are able to separate the instances of positive and negative emotion classes early in the classification process and with higher accuracy compared to the standard flat classification approaches. This reduces the proportion of instances of negative classes that get misclassified into a positive class and vice versa. Consequently, the overall classification performance is improved.

Our overall approach, instantiated with a BiLSTM, is shown in Fig. 2. In our approach, the input text is first processed in the embedding layer to generate an appropriate word embedding representation (Part A). This embedding fed directly into section Bi (Part B & C), which is responsible for binary (positive/negative) emotion classification (i.e., sentiment analysis). Based on the results of Bi, the embedding fed into section PMul (top branch) and NMul (bottom branch) concerns multiclass positive and negative emotion classification. Please note our adopted emotion representation [18] recognizes only a single positive emotion class (*happiness*), so we can classify it directly from the output of section Bi. Hence, in our main experiments we have no PMul branch. The hidden state embedding vector for the input text is passed through a separate dense layer (Part B), after which an appropriate activation function is applied (Part C) depending on the classification section (sigmoid for section Bi and log-softmax for section NMul). As a result, each section produces an output ($Output_{Bi}$ or $Output_{NMul}$) corresponding to its respective classification task. The generated outputs contain the predicted class labels ($Output_{Bi}$: *positive* or *negative*, $Output_{NMul}$: *anger*, *fear*, or *sadness*.) for the provided input text.

Emotion class labels vary between datasets. To make consistent comparisons, we make use of four emotion class labels. Jack et al. [18] recently specified four Universal emotions: *anger*, *happiness*, *fear*, and *sadness*. However, our framework can be reused to extend emotion classes, positive or negative, as shown later.

```

if  $Output_{Bi}$  == "Positive" then
   $Output_{Final}$  = "Happiness"
else
  
```

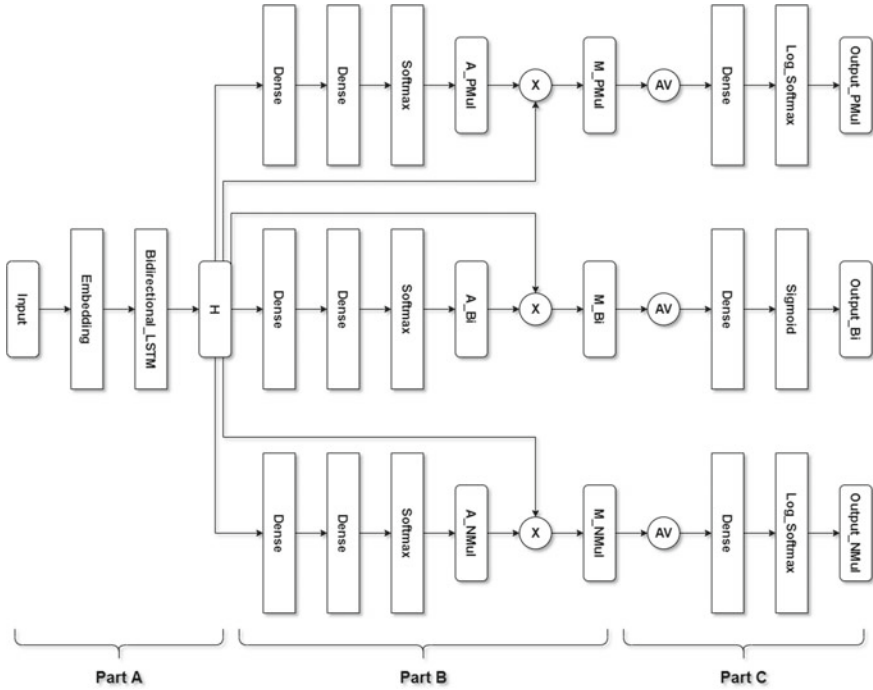


Fig. 2 Hierarchical architecture with a BiLSTM model

$Output_{Final} = Output_{NMul}$
end if

3.1 Loss Combination

We use 2 separate loss functions in our hierarchical framework with single positive emotion class: binary cross-entropy for the binary sentiment classification (Bi) and negative log-likelihood for the multiclass negative emotion classification (NMul). The per-section loss is calculated by applying the selected loss function to the classification output of its corresponding section. However, during training we minimize the total model loss, which is defined as the sum of per-section losses.

$$L_T = L_1(Output_{Bi}) + L_2(Output_{NMul}) \quad (1)$$

Here, L_T denotes the total model loss, while L_1 and L_2 represent the loss functions applied to the classification outputs of their corresponding sections.² This approach to

² L_2 component will be repeated in case of multiple positive emotions for PMul.

loss optimization allows the model to learn by considering the losses of its component classification sections both individually (as per-section losses) and collectively (as total model loss). It is possible to investigate optimal weight combination; however, in this work we used the same weight for both as this is still an initial investigation of the hierarchical model’s behavior.

3.2 *Instantiating the Approach with the State-of-the-Art Models*

We have used the following baseline models: The Self-Attentive Sentence Embedding (SASE) model [19] consists of a BiLSTM-based neural network supported by a self-attention mechanism; The Bidirectional GRU (BiGRU) model [8] consists of a BiGRU-based neural network with concatenated average-pooling and max-pooling layers; and pretrained Bert-based model with 12 layers [20]. In this section, we will provide the instantiation of the SASE model.

Given an input text $S = (w_1, w_2, \dots, w_n)$ with n words, each word is represented by a word embedding of size s . The output of the embedding layer is then forwarded to the bidirectional LSTM layer. This layer is essential to the model, as it allows it to understand and learn the context of the input text in both directions. The bidirectional LSTM produces a hidden state vector for each word in the input text. More formally, the generated hidden state vector for the i th word, after concatenating the forward and backward pass hidden state vectors for each of the n words in the input text, is of the form: $h_i = h_{i_f} \oplus h_{i_b}$. Thus, the hidden state matrix H with a shape of $(n, 2d)$ is produced, where d denotes the hidden dimension of the bidirectional LSTM layer, $H = [h_1, h_2, \dots, h_n]$.

Attention weight vector a is calculated by applying several transformations (represented by the 2 dense layers followed by the softmax layer in the architecture of the SASE model) to the hidden state matrix H : $a = \text{softmax}(w_2 \tanh(W_1 H^T))$.

The hidden state embedding vector m of the input text is then calculated as a weighted sum of the hidden state vectors comprising the matrix H , $m = aH$. The idea is to identify multiple aspects that collectively express the semantics of the whole text into vectors. The SASE model generates an attention weight matrix A of p such vectors, each focusing on a different part of the input. Thus, we have $A = \text{softmax}(W_2 \tanh(W_1 H^T))$. The hidden state embedding matrix M (corresponding to the hidden state embedding vector m described earlier) is then calculated through the multiplication of the attention weight matrix A and the hidden state matrix H , giving $M = AH$. The final classification output is produced through the application of a sigmoid function. Similar to this approach, BiGRU and Bert-based models are applied instead of SASE model for our experiments, with the ability of any SOTA text understanding model to be also used, creating hierarchical models each with its unique structure and capabilities.

Table 1 The statistics of the datasets—percentage of samples per emotion class and total number of samples

Dataset	Anger (%)	Fear (%)	Sadness (%)	Happiness (%)	Total
CF	6.5	42	25.7	25.8	20,062
TEC	9.5	17.1	23.3	50.1	16,429
Wang	27.0	6.7	30.6	35.7	1,078,878
Vent	10.6	12.4	48.2	28.8	2,259,748

4 Experiments

Datasets: For the evaluation, we use four different datasets, each with its own unique characteristics. The CrowdFlower (CF) was annotated by humans using crowdsourcing, and it is considered a noisy and difficult dataset [8]. Wang et al. [9] created a large dataset by annotating tweets using the hashtags present in the tweets. The twitter emotion corpus (TEC) [21] was also annotated using hashtags. The Vent dataset [22] uses data from the Vent social network, spanning from 2013–2018, where each user states its current emotional situation. This experimentation and use of Vent data is the first for the specific dataset, and we aim to create and establish benchmarks for the research community concerning text classification into emotions. All datasets are annotated with a lot different emotion labels, and to facilitate comparisons, we only use 4 basic emotion classes—*happiness*, *sadness*, *fear*, and *anger* [18]. Table 1 summarizes the statistics of each dataset.

Evaluation Protocols: We used a random set of training and testing samples split at 80% and 20%, respectively, and we run each experiment 10 times. A balanced training/testing size for each sentiment class is created, so as to avoid any biases. We use accuracy, recall, precision, and F1 score as our evaluation metrics.

Experimental Parameters: During our experiments, we used Glove word embeddings [23] of sizes 50 and 300 for SASE [24] and BiGRU [8], respectively, since they give the best performance. We set the maximum length of a text to 35 words, as tweets are usually short and use padding with zeros if a text is bigger. The number of epochs during training ranged from 5 to 10 epochs for all three models. The size of the vocabulary for SASE and BiGRU was set to 25,000 words. For the Bert model, we used the Bert-based implementation, which has its own special embeddings and does not require a limited word vocabulary.

4.1 Results and Discussion

We report the overall results for all models evaluated using the datasets with 4 emotions—“Happiness,” “Anger,” “Fear,” and “Sadness,” shown in Table 2. Hier-

Table 2 The overall results for multiclass classification

Model	Metric	CF	TEC	Wang	Vent
Bert	Accuracy	0.535	0.649	0.649	0.675
	Precision	0.530	0.651	0.650	0.678
	Recall	0.535	0.649	0.649	0.675
	F1	0.532	0.650	0.649	0.676
Bert-H	Accuracy	<u>0.642</u>	<u>0.706</u>	<u>0.711</u>	<u>0.730</u>
	Precision	0.639	0.704	0.710	0.737
	Recall	0.642	0.706	0.711	0.730
	F1	<u>0.640</u>	<u>0.704</u>	<u>0.710</u>	<u>0.733</u>
SASE	Accuracy	0.450	0.564	0.551	0.597
	Precision	0.461	0.574	0.553	0.607
	Recall	0.450	0.564	0.551	0.597
	F1	0.453	0.564	0.551	0.596
SASE-H	Accuracy	<u>0.559</u>	<u>0.632</u>	<u>0.618</u>	<u>0.662</u>
	Precision	0.575	0.636	0.625	0.685
	Recall	0.559	0.632	0.618	0.662
	F1	0.563	0.633	0.621	0.670
BiGRU	Accuracy	0.477	0.563	0.558	0.615
	Precision	0.482	0.574	0.562	0.619
	Recall	0.477	0.563	0.558	0.615
	F1	0.476	0.562	0.557	0.614
BiGRU-H	Accuracy	<u>0.530</u>	<u>0.607</u>	<u>0.616</u>	<u>0.622</u>
	Precision	0.554	0.629	0.632	0.655
	Recall	0.530	0.607	0.616	0.622
	F1	0.525	0.611	0.618	0.629

The underlined results show the best model for each dataset. Bold demonstrates statistical significance for each individual model between its hierarchical method and without it, respectively

archical models were named with the extension “-H” after the model’s name. The underlined results show the best model for each dataset. Bold results demonstrate statistical significance for each individual model between its hierarchical method and without it, respectively.

We compare our performance with the state-of-the-art emotion detection models. BiGRU [8], SASE [24], and pretrained transformer model—Bert [25] are trained to the downstream task of emotion classification. As shown in Table 2, the respective hierarchical methods achieve better results than its vanilla variation. Bert hierarchical model shows an increase of at least 6% for all metrics and datasets. Statistical tests (t -test) show that all hierarchical methods are significantly better than their simple variations (p -values < 0.023). Overall, Bert hierarchical approach gives superior performance.

We can also observe that all models that use Bert can achieve substantially better results than both the BiGRU and SASE. There is an increase of about 5% in all

metrics and datasets used, for both Bert-based models compared to both BiGRU and SASE. This was expected as Bert is built in such a way to understand and capture the underlying meanings of words and sentences through their context accurately. Besides this, Bert is a massive model with large depth (12 layers of transformer blocks) and is also trained on a massive corpus, compared to the training set used for BiGRU and SASE.

Multiple positive emotions: To explore the behavior of the hierarchical method, we conducted experiments with more than one positive emotion class (as shown in Fig. 2 where the architecture has one more layer for positive classification compared to the first experiment). Wang and Vent datasets have additional emotion classes, and hence, we used “Love” and “Thankfulness” as extra positive emotions called 3POS (3 positive and negative emotion classes). Similarly, we created a 2POS scenario with two positive classes “Love” and “Happiness,” and three negative classes (“Anger”, “Fear,” and “Sadness”). With these two additional datasets, we can explore and understand the role of the number of negative/positive emotion classes and the balance between them while using the hierarchical method. In Table 3 are the results for our models evaluated on Wang and Vent datasets consisting of extra positive emotions. The results from both datasets demonstrate that Bert-based models are the best among the three models, as expected from the previous results too. Bert’s models show a difference of near 10% and 8% in all metrics for Wang and Vent, respectively, compared to BiGRU and SASE. These major improvements once again prove the expressive power and capabilities that models such as Bert have.

With the said experimentation, we have the following results. On the Wang dataset, Bert-H (hierarchical) had consistent increase of about 2% on F1 and accuracy over simple Bert model on both scenarios. However, for the Vent data, Bert-H results were almost identical or slightly better than simple Bert (difference of F1 less than 1%). SASE-H model shows an increase of 1% in all metrics for both Wang and Vent 2POS setting as well as an increase of 2% in F1 on Vent 3POS compared to simple SASE model. However, on Wang 3POS scenario it exhibits a decrease of close to 0.5% F1 over its simple model. BiGRU-H model demonstrated a decrease of 1% and 2% in F1 score, respectively, for both 2POS and 3POS settings, compared to simple BiGRU. Overall, the hierarchical model performance improvements decreased when we consider more than one positive emotion class. We believe this is due to cross-talk between positive emotion classes, which we will explore below.

We observe no noticeable difference between the results produced by these models for data arrangements 3POS and 2POS. This is surprising, as the expected effect of reducing the number of considered emotion classes would be the general increase in classification performance, given the resulting increase in base random accuracy. The results suggest that there is a significant semantic overlap between the removed positive emotion class (“Thankfulness”) and the remaining two in the 2POS, given that its removal did not improve the classification performance greatly. As documented in [26, 27], “people rarely describe feeling a specific positive emotion without also claiming to feel other positive emotions”. As a result, the potential performance improvement introduced by the binary (positive/negative) emotion classifi-

Table 3 The overall results for multiclass classification with multiple positive emotion classes (3POS and 2POS) on Wang and Vent dataset

3POS				2POS			
Model	Metric	Wang	Vent	Model	Metric	Wang	Vent
Bert	Accuracy	0.588	0.579	Bert	Accuracy	0.602	0.619
	Precision	0.589	0.583		Precision	0.603	0.621
	Recall	0.588	0.579		Recall	0.602	0.619
	F1	0.588	0.579		F1	0.602	0.619
Bert-H	Accuracy	<u>0.599</u>	<u>0.580</u>	Bert-H	Accuracy	0.619	<u>0.623</u>
	Precision	0.598	0.583		Precision	0.620	0.629
	Recall	0.599	0.580		Recall	0.619	0.623
	F1	<u>0.599</u>	<u>0.580</u>		F1	<u>0.619</u>	<u>0.624</u>
SASE	Accuracy	0.494	0.483	SASE	Accuracy	0.499	0.535
	Precision	0.497	0.493		Precision	0.501	0.544
	Recall	0.494	0.483		Recall	0.499	0.535
	F1	0.494	0.482		F1	0.499	0.537
SASE-H	Accuracy	0.491	0.506	SASE-H	Accuracy	0.513	0.546
	Precision	0.489	0.513		Precision	0.513	0.556
	Recall	0.491	0.506		Recall	0.513	0.546
	F1	0.488	0.506		F1	0.511	0.548
BiGRU	Accuracy	0.496	0.512	BiGRU	Accuracy	0.505	0.555
	Precision	0.499	0.519		Precision	0.509	0.560
	Recall	0.496	0.512		Recall	0.505	0.555
	F1	0.495	0.512		F1	0.505	0.555
BiGRU-H	Accuracy	0.475	0.491	BiGRU-H	Accuracy	0.498	0.542
	Precision	0.478	0.501		Precision	0.502	0.552
	Recall	0.475	0.491		Recall	0.498	0.542
	F1	0.475	0.492		F1	0.496	0.543

The underlined results show the best model for each dataset. Bold demonstrates statistical significance for each individual model between its hierarchical method and without it, respectively

cation section of the hierarchical model is reduced when the model has to distinguish between 2 or more, highly semantically similar emotion classes. To sum up, due to the high semantic similarity of “Love,” “Thankfulness,” and “Happiness,” the performance improvements of hierarchical model decrease, when we consider multiple positive emotion classes. This opens up the need for creating more accurate test data, as unlike visual domain [10, 18], basic emotions in text data are not explored.

5 Conclusion

In this paper, we propose a novel end-to-end hierarchical approach to detect emotions in short text documents and evaluated it on four datasets. Our hierarchical approach can be instantiated with any of the state-of-the-art model, and we demonstrated with three state-of-the-art neural models (BiGRU [8], SASE [24], and Bert [25]). Our experiments clearly show the significant improvements brought in by the hierarchical approach. We have established strong baselines on these four datasets which can be used for comparison purposes in future research. We have also experimented with multiclass positive emotion categories. This helped us to understand the effect of the number of emotion classes within the hierarchical approach. Additionally, we have shown that the semantic overlap between the positive emotion classes affects the performance of the hierarchical approach and there is a need to develop appropriate test datasets. Nonetheless, the hierarchical approach we propose demonstrated great performance improvements and these observations are an encouraging first step toward the understanding and general adoption of this approach.

References

1. Yadollahi A, Shahraki AG, Zaiiane OR (2017) Current state of text sentiment analysis from opinion to emotion mining. *ACM Comput Surv* 50(2):25–12533
2. Deep KS, Ekbal A, Bhattacharyya P (2019) A deep neural framework for contextual affect detection. *ICONIP 2019: neural information processing*, vol 11955. Springer, Cham, pp 398–409
3. Sangwan S, Chauhan DS, Akhtar MS, Ekbal A, Bhattacharyya P (2019) Multi-task gated contextual cross-modal attention framework for sentiment and emotion analysis. *ICONIP 2019: neural information processing*, vol 1142. Springer, Cham, pp 662–669
4. Saha T, Patra AP, Saha S, Bhattacharyya P (2020) Towards emotion-aided multi-modal dialogue act classification. In: *ACL 2020*, pp 4361–4372
5. Chauhan DS, Dhanush SR, Ekbal A, Bhattacharyya P (2020) Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In: *58th ACL 2020*
6. Akhtar MS, Chauhan DS, Ekbal A (2020) A deep multi-task contextual attention framework for multi-modal affect analysis. *ACM Trans Knowl Discov Data* 14(3):32–13227
7. Chauhan DS, Dhanush SR, Ekbal A, Bhattacharyya P (2020) All-in-one: a deep attentive multi-task learning framework for humour, sarcasm, offensive, motivation, and sentiment on memes. In: *AAACL/IJCNLP 2020*
8. Seyeditabari A, Tabari N, Zadrozny W (2019) Emotion detection in text: focusing on latent representation. *AAAI*
9. Wang W, Chen L, Thirunarayan K, Sheth AP (2012) Harnessing twitter “big data” for automatic emotion identification. In: *2012 international conference on privacy, security, risk and trust, PASSAT 2012*, pp 587–592
10. Paul E (1999) Basic emotions. *Handbook of cognition and emotions*, pp 45–60
11. Bollen J, Mao H, Pepe A (2011) Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: *Proceedings of the fifth international conference on weblogs and social media, Spain*
12. Hasan M, Rundensteiner EA, Agu E (2019) Automatic emotion detection in text streams by analyzing twitter data. *Int J Data Sci Anal* 7(1)

13. Kumar A, Ekbal A, Kawahara D, Kur S (2019) Emotion helps sentiment: multi-task model for sentiment and emotion analysis. In: IJCNN 2019
14. Ghazi D, Inkpen D, Szpakowicz S (2010) Hierarchical versus flat classification of emotions in text. In: Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text. Association for Computational Linguistics, Los Angeles, CA, pp 140–146. <https://aclanthology.org/W10-0217>
15. Chauhan DS, Dhanush SR, Ekbal A, Bhattacharyya P (2020) Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In: Proceedings of the 58th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp 4351–4360 (Online). <https://www.aclweb.org/anthology/2020.acl-main.401>
16. Chen J, Hu Y, Liu J, Xiao Y, Jiang H (2019) Deep short text classification with knowledge powered attention. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 6252–6259
17. Zeng J, Li J, Song Y, Gao C, Lyu MR, King I (2018) Topic memory networks for short text classification. arXiv preprint [arXiv:1809.03664](https://arxiv.org/abs/1809.03664)
18. Jack R, Sun W, Delis I, Garrod O, Schyns P (2016) Four not six: revealing culturally common facial expressions of emotion. *J Exp Psychol: Gen* 145(6)
19. Lin Z, Feng M, dos Santos CN, Yu M, Xiang B, Zhou B, Bengio Y (2017) A structured self-attentive sentence embedding. CoRR abs/1703.03130
20. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019—2019 conference of the North American chapter of the Association for Computational Linguistics: human language technologies—proceedings of the conference 1 (MLM), pp 4171–4186. <https://arxiv.org/abs/1810.04805>
21. Mohammad SM, Turney PD (2013) Crowdsourcing a word-emotion association lexicon. *Comput Intell*. <https://doi.org/10.1111/j.1467-8640.2012.00460.x>
22. Lykousas N, Patsakis C, Kaltenbrunner A, Gómez V (2019) Sharing emotions at scale: the vent dataset. In: AAAI conference on web and social media, vol 13
23. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
24. Lin Z, Feng M, Santos CNd, Yu M, Xiang B, Zhou B, Bengio Y (2017) A structured self-attentive sentence embedding. arXiv preprint [arXiv:1703.03130](https://arxiv.org/abs/1703.03130)
25. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805)
26. Watson D, Clark LA (1992) On traits and temperament: general and specific factors of emotional experience and their relation to the five-factor model. *J Pers* 60(2):441–476
27. Posner J, Russell JA, Peterson BS (2005) The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev Psychopathol* 17(3):715

Towards an Enhanced Understanding of Bias in Pre-trained Neural Language Models: A Survey with Special Emphasis on Affective Bias



Anoop K. , Manjary P. Gangan , Deepak P. , and Lajish V. L. 

1 Introduction

Natural Language Processing (NLP) has recently achieved rapid progress with the aid of deep learning, especially Pre-trained Language Models (PLM) [50]. Large PLMs like BERT [31], GPT [87], etc., are highly efficient at capturing linguistic properties and producing representations of text with semantic and contextual information. Inclusion of contextual representations has led large PLMs to become popular towards addressing many downstream tasks such as Question Answering, Sentiment Analysis, and Neural Machine Translation [86]. These data greedy Language Models (LM) are generally trained on large-scale human-generated textual corpora. However, since ancient days, language has functioned as a channel to express and propagate unfairness toward marginalized social groups and assign power to oppressive institutions [29]. It is often very hard to analyze the quality of data in large corpora in context of such oppressive nature of language [117]. Yet, these human-generated textual corpora can carry plenty of harmful linguistic biases and social stereotypes that can lead NLP algorithms to produce unfair discrimination toward socially marginalized populations when deployed in real word [77]. A threatening scenario that was

The examples provided in this paper may be offensive in nature and may hurt your moral beliefs.

Anoop K. (✉) · Manjary P. Gangan · Lajish V. L.
University of Calicut, Malappuram, Kerala, India
e-mail: anoopk_dcs@uoc.ac.in

Manjary P. Gangan
e-mail: manjaryp_dcs@uoc.ac.in

Lajish V. L.
e-mail: lajish@uoc.ac.in

Deepak P.
Queen's University Belfast, Belfast, UK
e-mail: deepaksp@acm.org