Artificial Intelligence Review (2021) 54:5789–5829
https://doi.org/10.1007/s10462-021-09958-2

Transformer models for text-based emotion detection: a review of BERT-based approaches

Francisca Adoma Acheampong[1] · **Henry Nunoo-Mensah**[2] · **Wenyu Chen**[1]

Published online: 8 February 2021

Abstract
We cannot overemphasize the essence of contextual information in most natural language processing (NLP) applications. The extraction of context yields significant improvements in many NLP tasks, including emotion recognition from texts. The paper discusses transformer-based models for NLP tasks. It highlights the pros and cons of the identified models. The models discussed include the Generative Pre-training (GPT) and its variants, Transformer-XL, Cross-lingual Language Models (XLM), and the Bidirectional Encoder Representations from Transformers (BERT). Considering BERT's strength and popularity in text-based emotion detection, the paper discusses recent works in which researchers proposed various BERT-based models. The survey presents its contributions, results, limitations, and datasets used. We have also provided future research directions to encourage research in text-based emotion detection using these models.

Keywords Natural language processing · Sentiment analysis · Text-based emotion detection · Transformers

1 Introduction

Natural Language Processing (NLP) is a branch of Computer Science and Artificial Intelligence that focuses on the computational treatment of human language with the core intent of making machines understand and generate human languages. NLP's favored applications, such as translation systems, search engines, natural language assistants, sentiment, and opinion analysis, are resolving societal issues at an unprecedented rate (Yue et al. 2019; Gobinda 2003; Wang et al. 2019). Although data sources are available in diverse forms, i.e., images, voice/speech, body language, texts, the use

✉ Henry Nunoo-Mensah
hnunoo-mensah@knust.edu.gh

Francisca Adoma Acheampong
francaadoma@gmail.com

[1] Computational Intelligence Lab, School of Computer Science and Technology, University of Electronic Science and Technology of China, Chengdu, China

[2] Connected Devices Lab, Department of Computer Engineering, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

of texts has recently gained traction due to the enormity of text data available because of the emergence of Web 2.0 and the social media (Murugesan 2007). Textual data is sequential; hence the order of words and their relation in a sentence, i.e., context, play a vital role in deriving complete meaning from the sentence. Not only does the traditional unsupervised machine learning models disregard the occurring order or relationship of words in written texts but are also limited by the relatively small fixed input size. Thus, the motivation for applying computationally deep approaches to textual data. The Recurrent Neural Network (RNN) is a sequential model capable of handling sequential data (Du and Swamy 2013) but limited by the slow training time and the long-range sequence dependencies. However, a variant of RNN, the Long Short-Term Memory (LSTM), is paramount in mitigating some of these limitations. The LSTM provided a fair solution to the long-range sequence dependency problem but disregarded parallel computations and was even slower than the RNN (Sundermeyer et al. 2012).

Attention networks were proposed to alleviate some of the challenges associated with RNNs and LSTMs. The goal of an attention network is to focus on sections in data that are of high interest. With attention in focus, the ensemble of attention with other sequential models accelerated as identified in the work by Akhtar et al. (2020), where the intensities of emotions were detected using a stacked ensemble of convolutional neural network (CNN), LSTM, and gated recurrent unit (GRU). However, this approach required a ground-up architectural design for each domain-specific task and required the passing of single inputs. In 2017, Vaswani et al. (2017) proposed the transformer model that combined the ensemble as a single-structured architecture. The proposed transformer model by Vaswani et al. utilized attention together with encoders and decoders to extract relational context better, ameliorated the long-range sequence dependencies problem, and further allowed the input sequences to be passed in parallel. Using the transfer learning paradigm that centers on gaining knowledge from one task and applying it to other related tasks, the model eliminated the segregated learning paradigms that are often built from the ground up. It achieved this by training a new task with already saved weights from a related task in a pre-training process; thus, making training periods relatively shorter and yielding better accuracy (Ruder et al. 2019). In 2018, the transformer was used for language modeling (Al-Rfou et al. 2019), making it applicable for a substantive number of NLP applications.

Emotion Detection (ED) is a branch of sentiment analysis (SA) that seeks to extract fine-grained emotions from either speech/voice, image, or text data. Detecting emotions from texts has suffered great paucity regardless of the quantum of text data available (Acheampong et al. 2020). This issue has been partly due to the absence of voice modulation, facial expressions, etc., which may present cues to aid in context and relation extraction. Another is partly due to the unavailability of a practical context extraction approach for texts. Besides, the need for emotion-conveying words disambiguation to verify classified emotions as real emotions presents a significant hitch in the field because some texts convey multiple emotional expressions. Recently, state of the art (SOTA) results are being obtained in the field using pre-trained transformer-based models.

Thus, the objective of this paper is to review some of these transformer-based models that are being pre-trained to achieve the SOTA in various NLP tasks, specifically in text-based emotion detection. Initially, the paper highlights the concept of transformers and some transformer-based models, then, their pros and cons. The paper then focuses on how some pre-trained transformer-based models impact the field of text-based emotion detection by reviewing some of the recent works with regards to their contributions, datasets used, model architecture, results obtained, and limitations. Additionally, open issues and

interesting future research directions are presented to readers to encourage research in the field using these pre-trained transformer-based models.

In Sect. 2 the models of emotions and the concept of text-based emotion detection are discussed. Section 3 presents a brief introduction to the concept of transformers. Section 4 details some transformer-based models, their pros and cons. Section 5 reviews existing text-based emotion detection works that utilized transformer-based pre-trained models. Section 6 concludes the paper and points out some recommendations to improve research in the field of text-based emotion detection.

2 Emotion models and Text-based emotion detection

The models of emotion fundamentally define how an emotion differs from the other. Emotions are generally represented in the discrete and dimensional form. Discrete representations place emotions into finite categories. Prominent in this representation scheme is the Paul Ekman's model (Ekman 1999), which places emotions into six basic categories, i.e., happiness, sadness, anger, disgust, surprise, and fear. Ekman posited that these emotions are independent of each other, basic, and can produce complex emotions by combinations. The Robert Plutchik model (Plutchik 1980) assented to some of Ekman's postulates; the existence of primary emotions, and their combinations resulting in a complex emotion. He however named 8 of such primary emotions instead of 6 by Ekman which occured in opposite pairs. The 8 primary emotions as proposed by Plutchik occurring in opposite pairs were joy versus sadness, trust versus disgust, anger versus fear and surprise versus anticipation.

Orthony, Clore, and Collins (OCC) disagreed with the analogy of 'basic emotions' as presented by Ekman and Plutchik. However, they agreed that emotions arose due to how individuals perceived events, and those emotions varied according to their degree of intensity. They discretized emotions into 22, adding 16 emotions to the emotions Ekman posited as essential, thus spanning a much more comprehensive representation of emotions, with additional classes of relief, envy, reproach, self-reproach, appreciation, shame, pity, disappointment, admiration, hope, fears-confirmed, grief, gratification, gloating, like and dislike (Ortony et al. 1990).

The dimensional models, on the other hand, places emotions into a uni- or multi-dimensional space. The spatial order depicts the relationship between emotions and their relative degree of occurrence. The Circumplex of Affect proposed by Russell (1980) is a profound example of the dimensional emotion model that places emotions on a 2-dimensional circular wheel in an arousal-valence domain. Russell and Mehrabian (1977) also presented a 3-D emotional model made up of valence, arousal, and dominance with valence differentiating emotions by pleasantness and unpleasantness, arousal differentiating emotions by activations and deactivations, and dominance describing the degree to which experiencers control their emotions.

The Hourglass of Emotions revisited model presented by Susanto et al. (2020) is an improvement over the original Hourglass of Enotions model proposed by Cambria et al. (2012). Some issues identified by Susanto et al. inspired the revisit of the earlier proposal. The identified issues were the uncanny color associations of the original model, the presence of neutral and ambiguous emotions, the absence of polar emotions such as calmness and eagerness. Also, an advantage of having emotion categorization models was the ability to classify unknown concepts based on known features, however, the

original model contained wrong associations of antithetic emotions such as anger and fear (which were both negatives) or surprise and anticipation (which were opposite in terms of meaning but not in terms of polarity). The low polarity scores for compound emotions and the absence of self-conscious or moral emotions also contributed to the revisit. The Hourglass of Emotions revisited model together with other emotion categorization models was tested on three sentiment benchmarks (i.e., Blitzer dataset, Pang and Lee dataset, and Amazon dataset). The revisited hourglass model demonstrated the highest scores in all benchmarks. The achieved results were 94.72%, 93.29%, and 89.85% for the Blitzer, Pang and Lee, and Amazon datasets, respectively. Figure 1 illustrated the revised hourglass model.

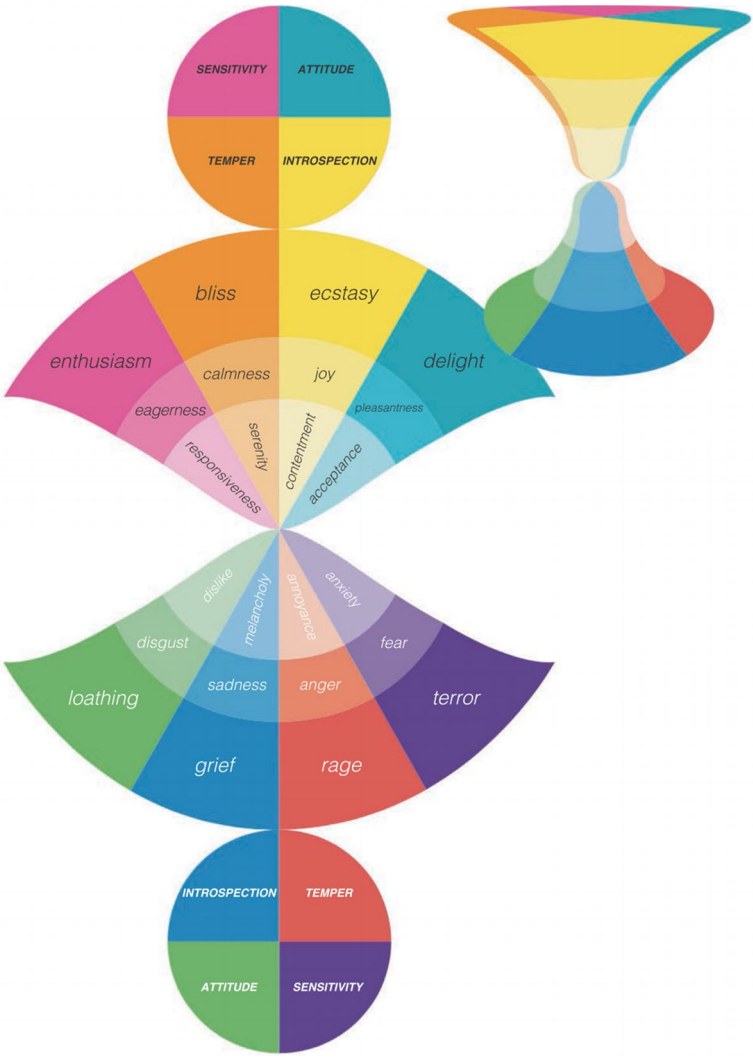


Fig. 1 The Hourglass of Emotions (Susanto et al. 2020)

ED is the extraction of finer-grained user sentiments. Text-based emotion detection is a sub-branch of emotion detection that focuses on extracting fine-grained emotions from written texts. Defining the text-based emotion recognition problem mathematically as,

$$r : A * T \rightarrow E \tag{1}$$

where A is an author that writes down a specific text, T is the written text from which emotions are to be extracted, and r is the relationship between the author and their written texts (Kao et al. 2009), Fig. 2 conveys the processes involved in recognizing emotions present in texts.

The approaches for recognizing emotions from texts as outlined by Acheampong et al. (2020) inferred that Machine learning approaches are currently achieving the state of the art with deep learning techniques in the lead. However, there remained the hurdle of adequate knowledge about the appropriate embedding techniques for extracting the relationship between long term dependent texts, the parallel processing of text sequence, and the integration of logical reasoning within these techniques. The transformer model provides a substantive solution to some of these problems.

3 Transformers

The transformer model, providing a substantive solution to the long-standing problems faced in sequential manipulations, breaks grounds at a breath-taking pace in the NLP research space. It has yielded many SOTA results in some NLP applications since its birth in 2017, particularly due to its constituents. Figure 3 shows the architecture of the transformer model.

The model consists of the encoder and decoder blocks with a softmax activation function for normalizing output probabilities. The input to the model is a sequence of data. The input words are embedded and passed through the positional encoders, which assign vectors to words based on their positioning in a sentence; thus, extracting input words' contextual meaning. The encoder blocks consisting of the multi-head attention and a feed-forward network receive the embeddings. The multi-head attention layers compute attention vectors for each input to represent how each word is related to other words in the same sentence, i.e., further capturing the contextual relationship between the words in the sentence. The attention vectors are passed through a feed-forward network one vector at a time to the decoder block. It is worth stating that parallelization is reached in the multi-head attention layer because the attention networks are independent.

The decoder has the positional encoders and masked multi-head attention layers that work similarly to the encoder block. The attention vectors from its masked multi-head attention layers and that from the encoder block are passed into another multi-head attention block. Each vector represents the relation with other words in the entire document. The vectors are then passed to a feed-forward network and then to the linear layer and finally to a softmax activation function that converts it into a probability distribution for the output. For the detailed operations of the transformer model, readers are encouraged to refer to the paper (Vaswani et al. 2017).

The transformer model, initially designed for machine translation, is being used for language modeling, making it applicable for other NLP tasks such as text classification, document summarization, question answering, etc. (Al-Rfou et al. 2019). The model used

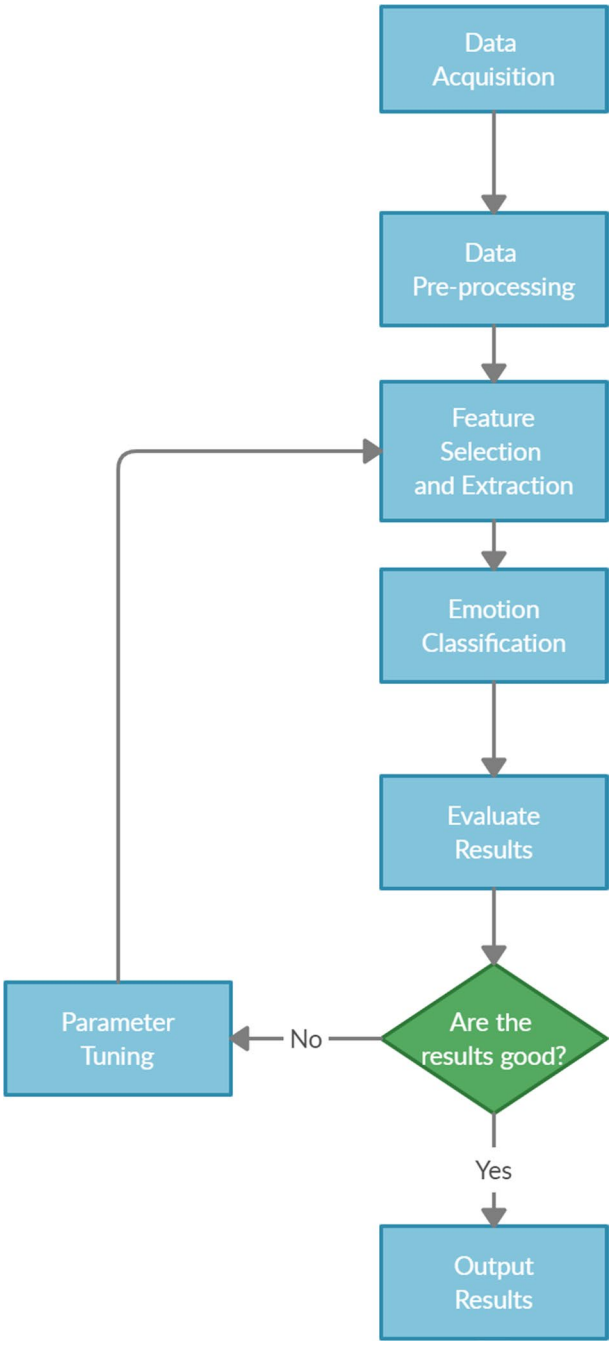


Fig. 2 The Emotion Recognition Process

a network of 64 transformer layers and causal attention to predict the next character in a fixed-length input. Limited by its fixed input length of 512 characters, the model segments input data and learns from each segment separately. This characteristic makes the model

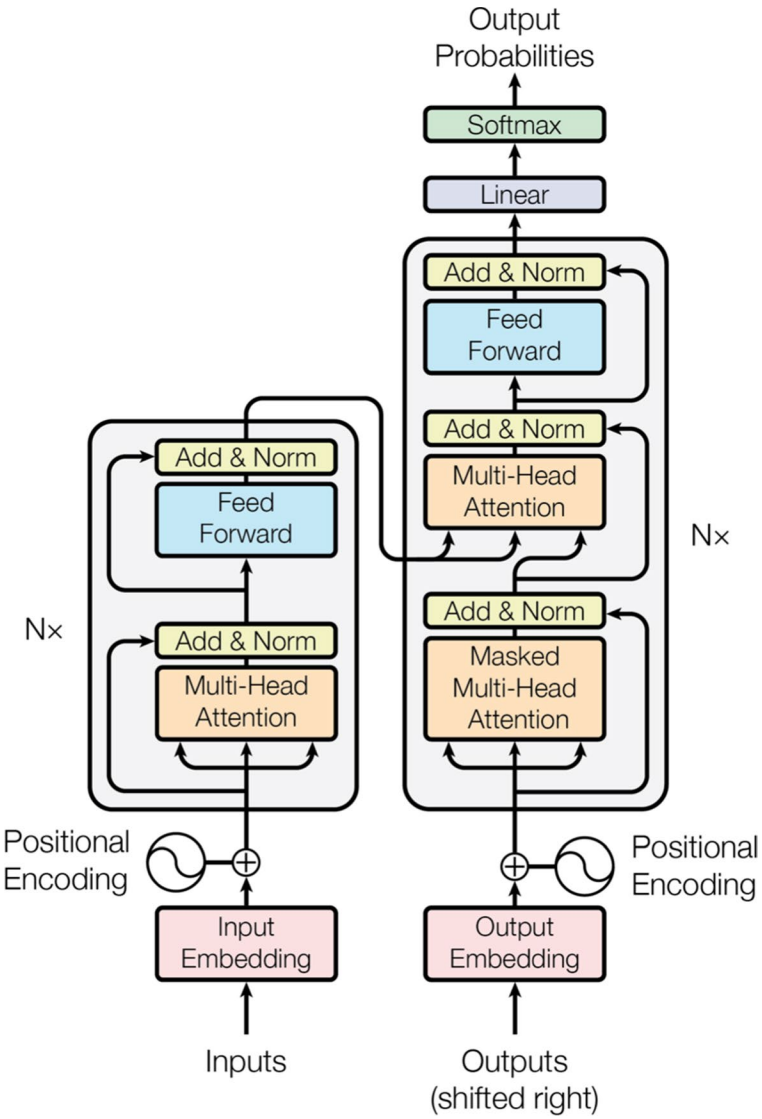


Fig. 3 Transformer model architecture (Vaswani et al. 2017)

suffer from context fragmentation and context-dependency limitations. From the transformer concept, other architectures have evolved.

4 Transformer-based Models

This section expounds on some models that have evolved from the transformer architecture, strengths, and weaknesses. The models discussed in this section include Transformer-XL, Generative Pre-Training (GPT), Bidirectional Encoder Representations from Transformers

(BERT), Cross-Lingual Language Model (XLM), XLNet, Robustly Optimized BERT pre-training Approach (RoBERTa), and DistilBERT.

4.1 Transformer-XL

Transformer-XL was proposed to solve the existing limitations associated with the vanilla transformer (Al-Rfou et al. 2019). The Transformer-XL (Dai et al. 2019) proposes an architecture beyond the fixed-length context. It achieves this using the segment level recurrence and the relative positional encoding. The recurrence technique at the segment level allows previously computed segment representations to be reused as context extension when new segments are processed. Thus, allowing the flow of contextual information across fixed segment boundaries. The relative positional encoding scheme validates the recurrence technique, and it is based primarily on the relative distance between tokens.

Transformer-XL can better capture longer-term dependencies than RNN and Al-Rfou's (Al-Rfou et al. 2019) transformer. The model is also capable of capturing short-term dependencies. Also, model inference for more protracted contexts is faster using Transformer-XL. The model achieves SOTA results for language modeling and audio analysis tasks. Notwithstanding the strengths, the model can only be used in limited application areas, i.e., audio analysis and machine translations.

4.2 Generative pre-training (GPT)

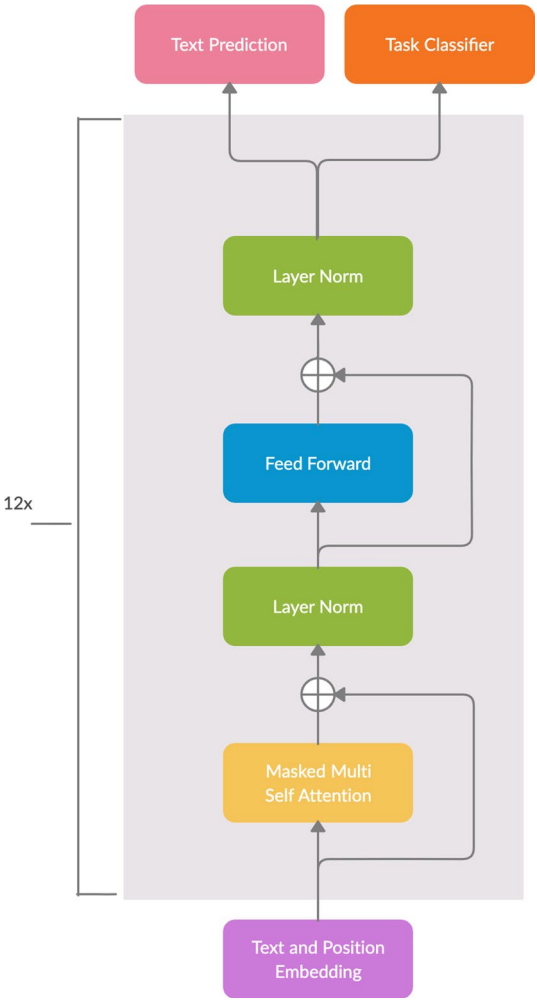
The use of structurally labeled data compared to unlabeled data yields better results in most machine learning tasks. This effect has put supervised learning at the frontiers in the most successful works in Machine Learning (ML). However, with the advent of Web 2.0, there exists a vast quantum of unlabeled data available for various tasks. When structured, labeled, and used in ML tasks, these unstructured data have the potency to achieve good results. Nevertheless, the manual efforts and long periods needed to label these unstructured data make the task challenging. To overcome this challenge, GPT leverages the semi-supervised learning approach to model language using transformer decoders (Radford et al. 2018). Mainly used for text representation, the GPT is made up of 12 transformer layers, 12 attention heads transformer decoder that uses the massive unlabeled datasets, i.e., the BooksCorpus dataset (Zhu et al. 2015) through pre-training and fine-tunes them on the limited supervised datasets.

The sole task of the GPT is to predict the next token in the sequence. It achieves this using the architecture shown in Fig. 4. The GPT's input is the input texts' weight embeddings plus their positional embeddings for context extraction. The input is passed to the multi-head attention layer in the 12 layered transformer decoder blocks, a feed-forward layer, and then the softmax outputs a probability distribution. Mathematically, this is expressed as:

$$\begin{aligned} h_0 &= UW_e + W_p \\ h_l &= \mathbf{transformer\_block}(h_{l-1}) \, \forall i \in [1, n] \\ P(u) &= \mathbf{softmax}(h_n W_e^T) \end{aligned} \tag{2}$$

where $U = (u-k, \ldots, u-1)$ is the context vector of tokens, k is the context window size, n is the number of layers, W_e is the token embedding matrix, and W_p is the position embedding matrix. The GPT model can be pre-trained and fine-tuned for other tasks. Scaling up the GPT model has produced the GPT-2 and GPT-3 architectures.

Fig. 4 GPT-2 architecture



Like the GPT model, the second version, i.e., GPT-2 (Radford et al. 2019), was designed to predict the next sentence in sentences and establishes that language models can learn tasks without direct supervision. It possesses an architecture like that of the GPT but with a normalization layer to the input of each sub-block and after the final self-attention layer. GPT-2 has four model sizes with the large and small models designed from a 40 gigabyte (GB) text with approximately 1.5 billion and 117 million trainable parameters respectively, thus, offering a scale up to the GPT model as shown in Table 1. Its larger data composition makes it suitable for diverse NLP tasks such as question answering, natural language inference, text classification, semantic similarity assessments, etc.

The GPT-3 model (Brown et al. 2020) scales up on the GPT-2 model even further with 175 billion trainable parameters. Its model architecture is the same as that of the GPT-2 except that the transformer layers have alternating dense and locally banded sparse attention patterns. A total of eight different sizes of the model was trained with a size range of 125 million to 175 billion parameters, as shown in Table 2. In the table, n_params

Table 1 Hyper-parameters for the four model sizes

Parameters	Layers	d(underscore model)
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

represents the total number of trainable parameters, n_layers represents the total number of layers, and d_model represents the number of units in each layer.

The strengths of the GPT model are highlighted as follows. There is improved lexical robustness when GPT is applied. The pre-trained GPT model can be fine-tuned to perform other tasks without model customization. The GPT model also out-performs various models trained on domain-specific datasets and produces SOTA results on a diverse range of domain-specific language modeling tasks. Concerning GPT-2 and GPT-3, they require no fine-tuning at all. The limitations of the GPT model are the resource-intensive nature of the model rendering the pre-training step expensive. Also, the model tends to suffer from the inability to model dependencies longer than designated fixed lengths.

4.3 Bidirectional encoder representations from transformers (BERT)

The Bidirectional Encoder Representations from Transformers (BERT) by Devlin et al. (2019), used the encoders in a transformer as a sub-structure to pre-training models for NLP tasks such as SA, Question Answering (QA), Text Summarization (TS), etc. BERT's execution for these tasks is in two phases viz., pre-training for language understanding, and fine-tuning for a specific task.

BERT can understand language by training on the Masked Language Modeling (MLM) and the Next Sentence Prediction (NSP) mechanisms. BERT assumed a blinder with MLM to learn bi-directional contexts in sentences. Thus, it takes in as input some random sentences, masks some of the words in the sentences, and reconstruct the masked words from the surrounding texts at the output. Its ability to input two sentences at once and determine if the second sentence comes after the first makes it achieve NSP. This ability helps the model to maintain long-distance relationships between texts. BERT was trained on 16GB texts from the BooksCorpus datasets (Zhu et al. 2015) and the English Wikipedia.

Table 2 Model hyper-parameters for GPT-3's 8 model sizes

Model Name	n_params	n_layers	d_model
GPT-3 Small	125M	12	768
GPT -3 Medium	350M	24	1024
GPT-3 Large	760M	24	1536
GPT-3 XL	1.3B	24	2048
GPT-3 2.7B	2.7B	32	2560
GPT-3 6.7B	6.7B	32	4096
GPT-3 13B	13.0B	40	5140
GPT-3 175B	175.0B	96	12,288

After pre-training, the model is then trained on an NLP task by performing supervised training on a dataset and replacing the BERT's fully connected output with a new set of output layers. The BERT model trains faster since the other model parameters are only fine-tuned aside from the output parameters learned from scratch (Fig. 5).

BERT has two models, i.e., the BERT-base and BERT-large models. The BERT-base model comprises 12-layered transformer encoder blocks with each block containing 12-head self-attention layers and 768 hidden layers and producing ≈ 110 million parameters in total. On the other hand, the BERT-large is made up of 24-layered transformer encoder blocks with each block containing 24-head self-attention layers and producing a total of ≈ 340 million parameters. The performance of BERT is dependent on the model type, i.e., BERT-large can achieve higher accuracies than BERT-base. However, this improvement in accuracy using the BERT-large comes at the cost of requiring more extensive resources to complete. The general architecture of the BERT model is shown in Fig. 6

The pros of BERT include its ability to handle contextual information extraction due to its bi-directional ability; it trains faster and has been used in a wide range of language modeling applications. However, the following cons persist for BERT: it is limited to monolingual classifications, the length of input sentences also limits it, it suffers from pragmatic inference, and the use of BERT-large may tend to be computationally expensive.

4.4 Cross-lingual language models

The BERT model, achieving promising results in NLP tasks, required monolingual data. The Cross-Lingual Language Models (XLMs) (Conneau and Lample 2019) improves BERT for classification and translation tasks by pre-training cross-lingual models for NLP. The model provides an extension to the MLM represented in BERT using the model translation concept proposed by Lample et al. (2018). It involves training both MLM and TLM and alternating between them; Fig. 7. The corpora are made up of the same sentences in two different languages. The sentences, one from each language corpus, are fed in parallel to the model's input. The model learns context by exploring the texts in each language and both, to predict masked words.

The strengths of XLM include: the model considers bilingual classification tasks, parallel text input is made possible with XLM, and XLM can be pre-trained to achieve SOTA

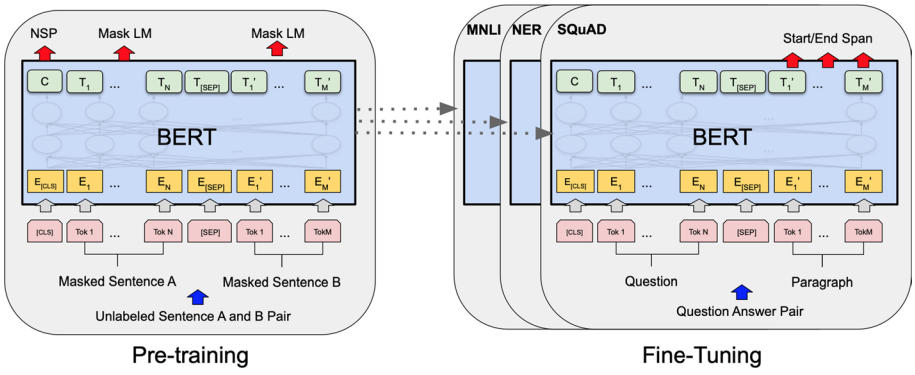
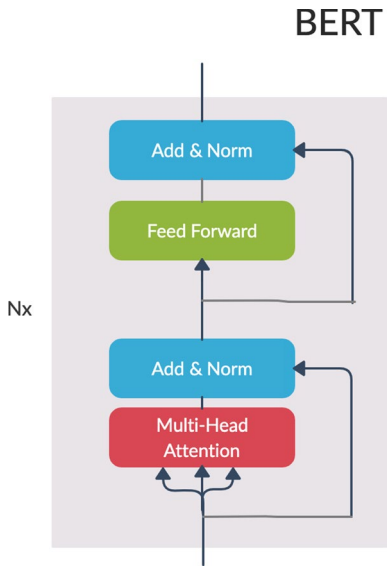


Fig. 5 Pre-training and fine-tuning BERT

Fig. 6 BERT model architecture



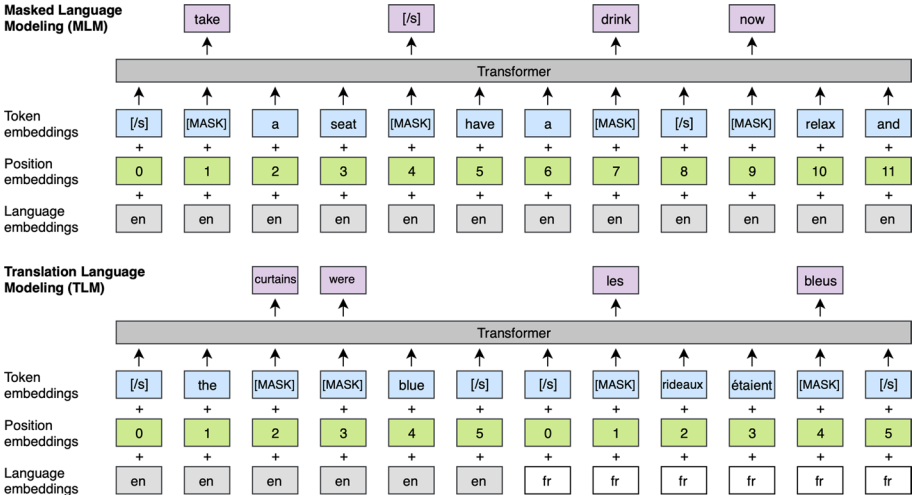


Fig. 7 The XLM model architecture

results in bilingual tasks. However, XLM suffers from the fixed-length limitation and has been applied in a limited number of NLP tasks.

4.5 XLNet: generalized auto-regression pre-training for language understanding

XLNet (Yang et al. 2019) is an auto-regressive language model that utilizes the concepts of the Permutation Language Model (PLM) and the Transformer-XL model to achieve the SOTA. As a BERT variant, the significant difference between the BERT and the XLNet has to do with their training objective. XLNet uses the permutation objective, whereas

BERT masks the data and tries to predict the masked data using a bi-directional context. With PLM, the model can learn bi-directional context by training all possible permutations of words in a sentence. Then using the positional encoding and recurrence mechanisms in Transformer-XL eradicates the fixed-length problem in BERT. The architecture presents a two-stream self-attention for target-aware representations, as shown in Fig. 8.

The strengths of the XLNet model include: the ability to extract contextual information due to the PLM implementation in the model, the model is known to perform better than BERT in a broader range of language modeling applications. Most importantly, it eradicates BERT's fixed-length limitation and can also be pre-trained to achieve the SOTA. However, all these advantages bring about some computational complexities, thus making XLNet computationally intensive.

4.6 Robustly optimized BERT pre-training approach (RoBERTa)

The Robustly Optimized BERT pre-training Approach (RoBERTa) (Liu et al. 2019) is a BERT variant that seeks to ultimately optimize BERT by tweaking various methodological parameters in the initial version of BERT. The tweaks helped to investigate hyper-parameters' essence, such as the effect of larger pre-training datasets, the influence of static over dynamic MLM, the contributions of batch sizes, and the relevance of text encoding and the importance of the NSP technique in BERT. A total of 160GB of uncompressed texts from five English language datasets: the data sources for the BERT model, the CC-News data, the Stories dataset (Trinh and Le 2018), and the Open Web data were trained on the model. They contained 16GB, 75GB, 31GB, and 38GB of sentences respectively. RoBERTa was trained with dynamic masking, FULL-SENTENCES without NSP loss, large mini-batches, and a larger byte-pair encoding (BPE). Dynamic masking was used instead of the static masking utilized for BERT; this helped avoid using a single mask for each epoch's training instance. The use of the dynamic mask offered slight improvement in multitasks but not in single tasks.

Different cases were defined for segment-pair + NSP, sentence-pair + NSP, full sentences, and document sentences to investigate the necessity for NSP. The decreasing order of performance was document sentences, full sentences, segment pairs, and

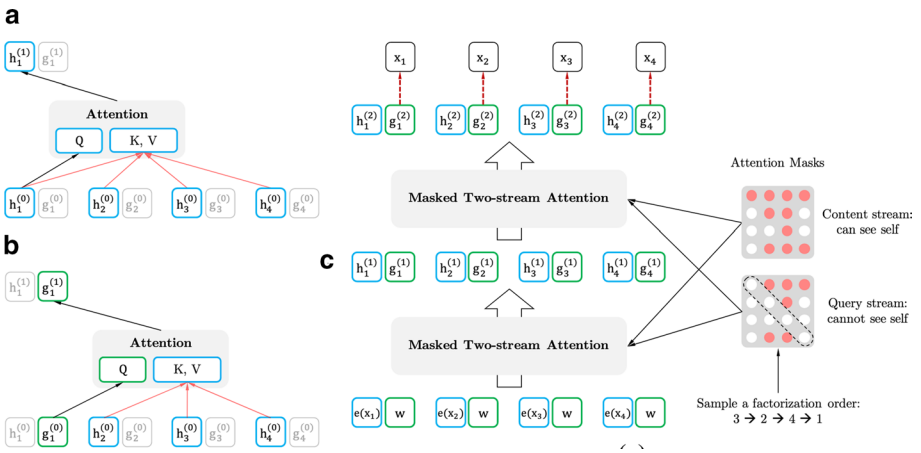


Fig. 8 The XLNet architecture

sentence pairs accordingly. However, the case of the full sentences was realized as the most effective. The authors reported that removing NSP did not affect performance heavily. The batch size of 2K was ultimate for the model in comparison with other batch sizes. Investigations were made into the performance of the word/character level encoding and BPE. BPE out-performed the word/character level encoding. The performance of RoBERTa compared with other BERT variants is shown in Table 3.

The advantages of RoBERTa are as follows; the more massive pre-training data used yields better performance in a variety of tasks. RoBERTa also out-performs XLNet and BERT in downstream NLP tasks. However, the cons of RoBERTa are resource-intensive nature as data requirements are extensive and increased computational complexity. The higher the pre-training step, the better the performance, but it becomes computationally expensive, requiring more time to execute.

4.7 DistilBERT

The concept of distillation in neural networks, as proposed by Tang et al. (2019), aims at speeding up models. It achieves that by replacing more massive architectures with broad parameters with a lightweight version of the same architecture that possesses fewer parameters. The DistilBERT operates in that manner. It takes the architecture of the initial version of BERT, reduces the number of layers in the BERT-base model by a factor of 2, removes token embeddings and poolers to yield a much smaller and faster version of BERT for general-purpose use. It applies dynamic masking and ignores the next sentence predictions as proposed by Liu et al. (2019) for better inference.

The General Language Understanding Evaluation (GLUE) benchmark was used to evaluate the model's performance on downstream tasks and performed comparatively well. DistilBERT specifically retained 97% of BERT's performance even when 40% fewer parameters were used. DistilBERT was also 60% faster than BERT.

DistilBERT is capable of language modeling and can be pre-trained on other language modeling tasks. The model is faster and lighter than the original BERT. The qualities showcase the strengths of DistilBERT. However, the issue of the fixed-length limitation of BERT persists in DistilBERT.

Table 3 Performance of BERT and other BERT variants

Model	Data	Bsz	Steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
With BOOKS + WIKI	16 GB	8 K	100 K	93.6/87.3	89.0	95.3
+ Additional data	160 GB	8 K	100 K	94.0/87.7	89.3	95.6
+ Pre-train longer	160 GB	8 K	300 K	94.4/88.7	90.0	96.1
+ Pre-train even longer	160 GB	8 K	500 K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
With BOOKS + WIKI	13 GB	256	1 M	90.9/81.8	86.6	93.7
$\text{XLNet}_{\text{LARGE}}$						
With BOOKS + WIKI	13 GB	256	1 M	94.0/87.8	88.4	94.4
+ Additional data	13 GB	2 K	500 K	94.5/88.8	89.8	95.6

The bold text shows the best results for various works

5 Recent advances of BERT-based models for text-based emotion detection

This section discusses some SOTA for detecting text-based emotions using BERT and its variants. Their approaches, contributions, accuracies, and model weaknesses or limitations are elucidated.

Yang et al. (2019), proposed a contextual emotion classifier, EmotionX-KU, which was applied to the EmotionX 2019 challenge shared task for detecting emotions in dialogue utterances. The EmotionLines dataset (Chen et al. 2019) consisting of two data subsets (i.e., the Friends and EmotionPush datasets) was released for the challenge; emotions were classified into four labels (i.e., neutral, joy, sadness, and anger). Table 4 shows the data and label distribution of the Friends and EmotionPush datasets. The dataset's sample distribution showed neutral having the most classes, followed by joy, sadness, and anger, respectively. The two datasets suffered from imbalanced class distributions, as shown in Table 4.

The authors considered the contextual emotion classification task a sequence problem whereby each utterance had a context in a dialogue. Their model leveraged transfer language modeling and dynamic max pooling to consider utterances and their corresponding contextual information effectively. Their model sought to advance the work by Khosla (2018). In performing input embedding, the authors initially lower-cased and tokenized the input utterances based on a BPE tokenizer. They then appended a special token '[SPE]' between utterance tokens belonging to different dialogues. The tokens were also embedded using WordPiece embedding (Wu et al. 2016). For example, the input embedding represented by E_{i,l_i}, (where i is the index of utterance and l_i is the length of the ith utterance), were the summation of the token and positional embedding. The language model by Vaswani et al. (2017) was adopted for the work. The selection was made due to the model's ability to alleviate the long-term dependency problem, thus effectively capturing worthy context information. To further enhance the adaptability of the chosen language model, the authors post-trained the model by Xu et al. (2019) via MLM and NSP. A dynamic max-pooling technique was applied to curb the heterogeneous sized utterance representation to a more homogeneous representation while still maintaining important information in each dimension. These representations from the encoders were eventually passed through a classifier consisting of two linear layers, a scaled exponential linear unit (SELU) activation function and a dropout layer. An overview of Yang et al. is illustrated in Fig. 9. They presented results for the case where two datasets were trained on their model together and the case where they were trained separately, as depicted in Tables 5 and 6. From the table, the authors hypothesized that post-training the self-attention-based transferable language model, i.e., uncased BERT-base model, produced more accurate contextual emotion predictions. The limitation associated with their work was that their model could not consider all the utterances when the number of input tokens exceeded the present maximum length.

The proposal by Huang et al. (2019a) for detecting emotions in dialogue utterances used the EmotionLines dataset. The dataset was the same as the one highlighted in Table 4 for the EmotionX 2019 challenge. The proposed model by Huang et al. consisted of three main phases viz., the casual utterances, pre-training, and fine-tuning phases, as illustrated in Fig. 10.

The casual utterances phase's objective was to preserve emotion information and utterances by merging two consecutive utterances in a dialogue together as a single utterance. The authors applied the casual utterances phase to each of the EmotionLines dataset (i.e., Friends and EmotionPush). In the Friends dataset, it was observed that dialogues were

Table 4 Corpus statistics and label distributions of Friends and EmotionPush datasets

Dataset	Set	#Dialogues /#Utterances	#Avg. utterances per dialogue	#Avg. length of dialogues	Neutral (%)	Joy (%)	Sadness (%)	Anger (%)	Out-Of-Domain (%)
Friend	Train	4000/58,012	14.50	160.92	45.0	11.8	3.4	5.2	34.6
	Test	240/3296	13.73	156.38	31.4	15.3	3.7	4.3	45.3
EmotionPush	Train	4000/58,968	14.74	114.96	66.8	14.2	3.5	0.9	14.6
	Test	240/3536	14.73	92.43	60.7	17.0	3.1	0.8	18.4%

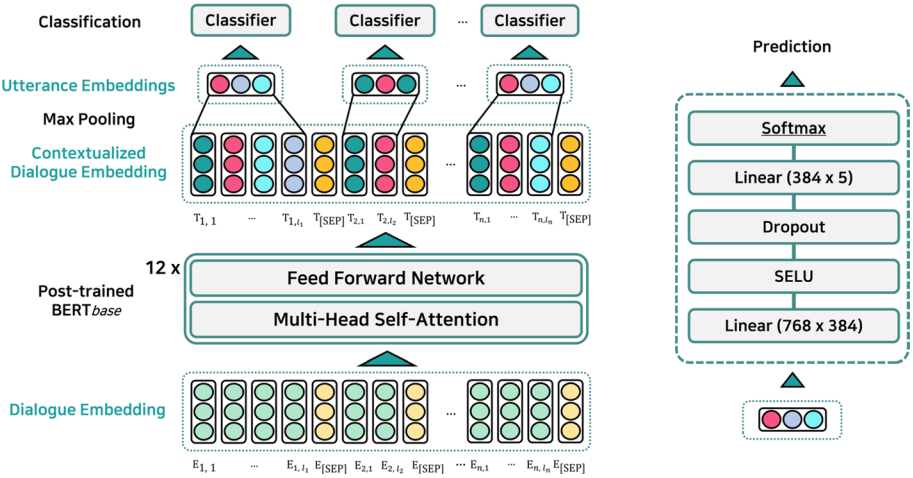


Fig. 9 EmotionX-KU model overview (Yang et al. 2019)

Table 5 F1 scores of emotions with friends and EmotionPush datasets trained separately

Dataset	Model	Micro-F1	Neutral	Joy	Sadness	Anger
Friends	Base + Mean	77.5	85.3	72.3	50.0	53.5
	Base + Max	77.1	85.0	71.5	49.7	59.9
	Post + Mean	**78.4**	85.3	**73.3**	**58.1**	**61.4**
	Post + Max	77.5	**85.5**	71.8	49.5	57.3
EmotionPush	Base + Mean	83.7	90.4	71.3	59.0	18.9
	Base + Max	85.0	90.6	**73.8**	61.1	29.8
	Post + Mean	84.1	90.5	71.3	61.5	20.0
	Post + Max	**85.6**	**91.1**	73.5	**63.4**	**30.6**

The bold text shows the best results for various works+

Table 6 F1 scores of emotions with friends and EmotionPush datasets trained together

Dataset	Model	Micro-F1	Neutral	Joy	Sadness	Anger
Friends	Base + Mean	74.0	83.9	63.6	49.6	53.8
	Base + Max	76.1	85.1	65.1	51.2	57.3
	Post + Mean	76.2	85.4	67.5	**54.7**	55.9
	Post + Max	**77.5**	**85.4**	**70.9**	52.0	**59.7**
EmotionPush	Base + Mean	84.4	90.6	71.5	52.9	33.3
	Base + Max	86.0	**91.6**	73.5	**61.9**	29.9
	Post + Mean	85.8	91.1	71.7	60.1	24.6
	Post + Max	**86.3**	91.5	**74.7**	61.0	**36.2**

The bold text shows the best results for various works

centered towards individual personalities. As a result of these characteristics, the authors applied personality tokenization to the Friends dataset to extract their character features.

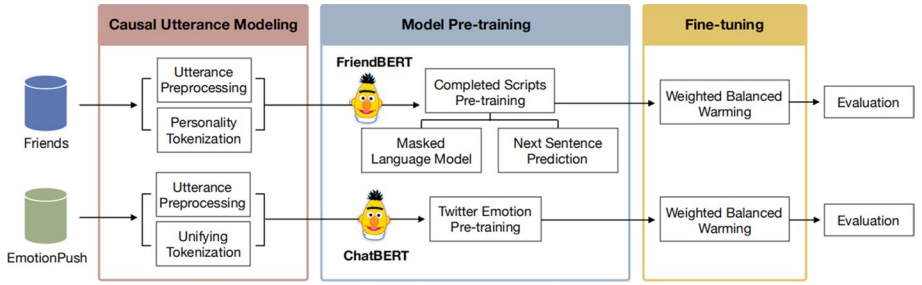


Fig. 10 Architecture of the EmotionX-IDEA: emotion BERT–an affectional model for conversation (Huang et al. 2019a)

During the personality tokenization process, speakers of interest had their utterances concatenated together with utterances from previous speakers, as shown in Table 7. Similarly, the EmotionPush dataset samples were pre-processed, and both outputs were forwarded to the pre-training stage. During pre-training, the utterances from the Friends dataset, which they called FriendsBERT after pre-processing, were prepared for the MLM pre-training and NSP tasks, respectively. EmotionPush, which they named as ChatBERT on the other hand, saw data from Twitter being used for pre-training. The processed data was then fine-tuned following the BERT architecture and then evaluated using 240 dialogues, including 3296 and 3536 utterances in Friends and EmotionPush, respectively. Their model achieved a micro F1 score of 81.5 and 88.5 for the Friends and EmotionPush datasets, respectively. These micro F1 scores were higher than those achieved by Yang et al. (2019), which were 77.5 and 86.3 for the Friends and EmotionPush datasets, respectively.

The SemEval 2019 contextual emotion detection in text (EmoContext) task received a submission from Huang et al. (2019b). The competition consisted of classifying the emotion of utterances from Twitter conversations as happy, sad, angry, or others (Chatterjee et al. 2019) given their conversational context. Huang et al. (2019b) proposed an ensemble approach made up of a Hierarchical LSTMs for contextual emotion detection (HRLCE) and BERT models. The main contribution of the authors was the proposal of the HRLCE model used for the classification. The dataset was composed of a reasonable number of emotion-carrying emojis. The authors used the *ekphrasis* package (Baziotis et al. 2017) to convert the emojis to their textual aliases and further cleaned the converted texts. The proposed architecture implemented by Huang et al. (2019b) is illustrated in Fig. 11.

The authors examined several pre-trained models to encode each user's utterance semantically and emotionally at the word level. A key component of the HRLCE is the Hierarchical or Context recurrent encoder-decoder (HRED) (Sordoni et al. 2015). The HRED architecture was used to capture the context information of dialogue exchanges effectively. The HRED architecture consisted of two RNN units: *encoder* RNN and *context* RNN. The *encoder* RNNs were used to map each utterance to an utterance vector while the *context* RNN further processed it. Due to the use of *context* RNNs, the Global Vectors for Word Representations (GloVe) (Pennington et al. 2014), the Embeddings for Language Models (ELMo) (Peters et al. 2018), and DeepMoji (Felbo et al. 2017) for embedding emojis were used to extract the syntactic, semantic information, and the emotional content of the input utterances. These embeddings were then encoded into vectors, and context LSTMs used to process these vectors further. A multi-head self-attention layer (Vaswani et al. 2017) was applied to focus on relevant feature vectors, and from the attended features, emotions were classified. To validate their model, Huang et al. (2019b) compared HRLCE with

Table 7 Training data distribution for the TRAC-1 workshop at COLING 2018

Speaker	Utterances	Emotion	Representation (with personality tokenization)	Label
Janice	I'm sorry.	Sadness	[CLS] I'm sorry. [SEP] [None] [SEP]	Sadness
Chandler	Ohhh. Don't go.	Sadness	[CLS] [Chandler] [says] ohhh. Don't go. [SEP] I'm sorry.[SEP]	Sadness
Janice	No, I gotta go.	Non-neutral	[CLS] No, I gotta go.[SEP] [Chandler] [says] ohhh. Don't go. [SEP]	Non-neutral

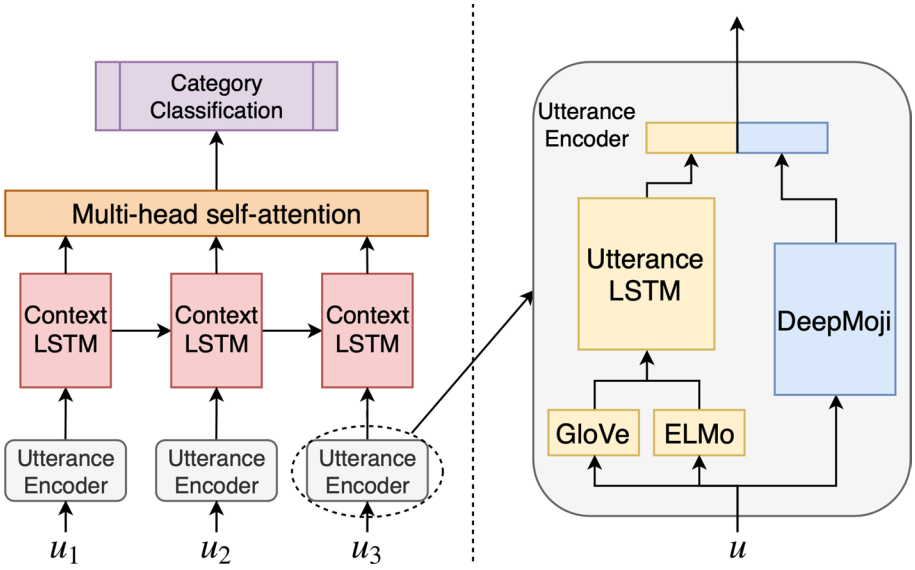


Fig. 11 The HRLCE architecture (Huang et al. 2019b)

some baseline models (i.e., self-attention LSTM (SL) and self-attention-LSTM-DeepMoji (SLD)) and applied BERT fine-tuning. HRLCE obtained an F1 score of 0.7709 after BERT fine-tuning. The limitation of the HRLCE was the poor miss-classification of the '*Others*' class. The authors proposed that researchers design a binary classifier first to classify the two classes '*Not-Others*' and '*Others*' before applying HRLCE on the '*Not-Others*' class. The authors believe this would help improve the accuracy of their work.

Malte and Ratadiya (2019) suggested using a deep learning approach using BERT to detect cyber abuse in English and Hindi texts. With data from the shared task on aggression identification organized at the Trolling, Aggression, and Cyberbullying (TRAC-1) workshop (Kumar et al. 2018), texts were classified into three categories: overtly aggressive (OAG), covertly aggressive (CAG) or non-aggressive (NAG). The distribution of the data used to train the model over these specified categories and languages is shown in Table 8. The data contained emojis, which were converted to text and pre-processed to eliminate noise. The training dataset was then passed onto an encoder and then to the decoder before being classified. The encoder was made up of a multi-head attention layer, residual connections, a normalization layer, and a feed-forward layer. The decoder had a masked attention layer also. Inside the encoder, input embeddings were concatenated, and each placed in a positional encoding of position and order. An attention mechanism was then applied bi-directionally so that each word's multiple attention features could be learned in the

Table 8 Training data distribution for TRAC-1

Category	English	Hindi
NAG	5052	2275
CAG	4240	4869
OAG	2708	4856

two-way context. The process was followed by residual connections so that weights could also be learned directly. The vectors were normalized and passed through the feed-forward network. The model was pre-trained for the MLM and the NSP model. The action helped the model perform better due to the bi-directional language model context. The Slanted Triangular Learning Rate (STLR) (Howard and Ruder 2018) was used to help the model converge. Compared with baseline methods, their approach attained an F1 score of 0.6596 for the Hindi Language and 0.6244 for the English language. Their model's drawback was the out of vocabulary limitation the WordPiece embedding presented and the need for a better technique for English slang representations.

Park et al. (2019) proposed a framework to learn dimensional valence arousal dimension (VAD) scores from the corpus with categorical emotion labels. The authors demonstrated their concepts by utilizing a fine-tuned, pre-trained BERT model. From their work, the authors were able to categorically classify emotions and predict VAD scores' dimensional emotions simultaneously. They utilized the NRC-VAD lexicon (Mohammad 2018) to categorically map labels to the VAD dimensional space to learn VAD distributions from the input sentences categorical labels. For instance, the emotion label 'joy' was mapped to (0.980, 0.824, 0.794) whereas 'sad' was mapped to (0.225, 0.333, 0.149) in the VAD space so that using these points, the distribution in Eq. (3)

$$P(e|X) = P(v, a, d|X) \tag{3}$$

where E is a set of pre-defined categorical emotion (i.e., joy, sad, happy, anger), e is an instance of a categorical emotion, and X, the written text/input text may be predicted. The variables v, a, and d represent the valence, arousal, and dominance dimensions. The authors trained their model by sorting categorical emotions in the E space into v, a, d scores so that target conditions for each $P(v|X)$, $P(a|X)$, and $P(d|X)$ may be obtained. Considering the categorical emotions of joy, sad, happy, and anger with valence scores (0.980, 0.225, 1000, 0.167), the categorical emotion space, E, were rearranged in ascending order to obtain anger, sad, joy, happy depending on their scores. It was carried out to obtain the target conditional $P(v|X)$. The distance between the true and predicted $P(.|X)$ was minimized using the Earth Movers Distance (EMD) loss (Hou et al. 2016). The categorical emotion labels were predicted by computing the product of predicted $P(v|X)$, $P(a|X)$, $P(d|X)$, to obtain the predicted $P(v, a, d|X)$, assuming conditional independence and an emotion label chosen using Eq. (4).

$$\underset{\{v,a,d\}=e\in E}{\arg\max}\ P(v, a, d|X) \tag{4}$$

The emotion label(s) with the highest probability over the designated threshold of 0.125 was chosen by comparing the joint probabilities. In predicting the VAD scores, the expectations of the predicted conditionals $P(v|X)$, $P(a|X)$, $P(d|X)$ were computed according to Eq. (5)

$$v_X = E(P(v|X)), a_X = E(P(a|X)), d_X = E(P(d|X)) \tag{5}$$

The overview of their model is shown in Fig. 12

Park et al. applied their model to the following datasets: SemEval 2018 E-c (Mohammad et al. 2018), ISEAR (Scherer and Wallbott 1994), and EmoBank (Buechel and Hahn 2017). The SemEval 2018 E-c contains 10, 983 tweets labeled for 'neutral or no emotion', or as one or more of the 11 emotions that best describe twitter users' mental state. The ISEAR data consists of 7 emotion labels, i.e., joy, sadness, fear, anger, guilt, disgust, and shame.

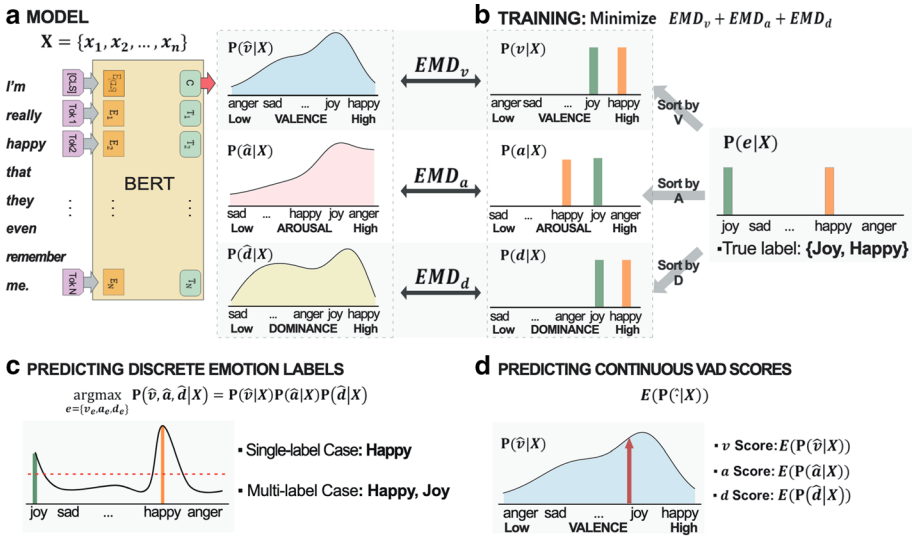


Fig. 12 The overview of the model (Park et al. 2019)

The final dataset reports a total of 7665 sentences labeled with emotions. The EmoBank data consists of over 10, 000 sentences annotated dimensionally according to the VAD emotion representation model. A subset of the dataset had been annotated categorically using Ekman's basic emotion model, making it suitable for dual representational designs. With the addition of a linear transformation layer with sigmoid activation, the authors used the BERT model to train the ISEAR and some of the SemEval data and then fine-tuned using the SemEval data. The proposed Park et al. model was evaluated using the EmoBank data and obtained a micro F1 score of 0.688 and 0.695 when fine-tuned on the ISEAR and SemEval datasets, respectively. When the model was trained on the SemEval dataset and fine-tuned on the EmoBank dataset, it obtained high VAD results of 0.659, 0.327, and 0.287, respectively, in comparison with the state of the art VAD regression models. However, the authors highlighted the need for models that gave sensible VAD scores without VAD annotations.

Participating in the CLPsych-2019 shared task, (Zirikly et al. 2019) Matero et al. (2019) assessed suicide risks using multi-level dual-context language and BERT. The data released for the shared task was obtained from Reddit. This data was annotated for four suicide risk categories, i.e., no risk, low risk, moderate risk, and severe risk. These were aggregated into single labels representing their highest suicide risk across the collection, as posited by Shing et al. (2018). Sub-dividing the process into three phases (i.e., Tasks A, B, and C) according to their data levels, Tasks A and B contained 496 training and 128 testing sets. Task C contained 993 training and 248 testing sets. Their dual-context modeling approach aimed at predicting suicide post separately from other posts. An attention-based RNN architecture, together with logistic regression, was implemented on Tasks A and B. For task C, only logistic regression was used and was implemented using the *DLATK* Python package (Schwartz et al. 2017). The model acting on the task A data is referred to as Deep-Att, the input to this model followed the 'OpenTheory' assertions. The input was concatenated and fed into an LSTM with hierarchical post-level attention (Yang et al. 2016) to extract details about the neural model and evaluated using BERT. For Task B, using their

Table 9 The multi-level dual context architecture

Model	Train		Test	
	Acc	F1	Acc	F1
Open	0.54	0.44	–	–
Theory	0.48	0.33	–	–
Single Context OpenTheory	0.50	0.35	–	–
Dual Context OpenTheory	**0.58**	**0.47**	0.56	0.46
DualDeepAtt	0.47	0.41	0.51	0.44
DualContextBert	0.53	0.43	**0.57**	**0.50**

The bold text shows the best results for various works

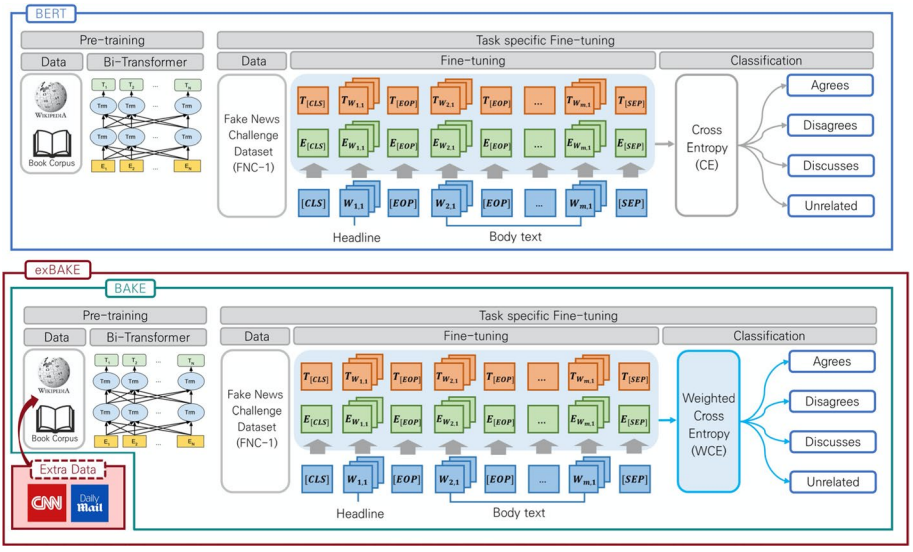


Fig. 13 Proposed BAKE and exBAKE model (Jwa et al. 2019)

'DualOpenTheory' logistic regression-based model, suicide and non-suicide features were processed separately, and BERT used for evaluation in their 'DualContextBert' approach. For Task C, they built a logistic regression model, 'OpenTheorySubr', using BERT embeddings without user traits of personality, age/gender, anxiety, anger, and depression scores. Table 9 highlights their results in comparison with other models.

Jwa et al. (2019) designed the BAKE and exBAKE model to automatically detect fake news using BERT through the scrutiny of headlines and body texts of news. Data was gathered from the Daily Mail and CNN sites (https://github.com/abisee/cnn-dailymail). The CNN articles collected between April 2005 and April 2017 consisted of approximately 90, 000 documents and 380, 000 questions. The Daily Mail articles, collected from June 2010 to April 2015, contained 197, 000 documents and 879, 000 questions. The authors made use of the fake news challenge stage 1 (FNC-1) data https://github.com/FakeNewsChallenge/fnc-1 to fine-tune their model. Their model was in two phases viz., the fine-tuning and pre-training phases, as shown in Fig. 13. The fine-tuning phase involved the use of weighted cross-entropy (WCE) to classify data into four classes; agrees (AGR), disagrees (DSG), discusses (DSC), and unrelated (UNR). This process they named as BAKE.

Table 10 Model performance of BAKE and exBAKE in comparison with other models

Models	F1	AGR	DSG	DSC	UNR
Majority vote	0.210	0.0	0.0	0.0	0.839
TalosComb (Baird et al. 2017)	0.582	0.539	0.035	0.760	0.994
TalosTree (Baird et al. 2017)	0.570	0.520	0.003	0.762	0.994
TalosCNN (Baird et al. 2017)	0.308	0.258	0.092	0.0	0.882
Athene (Hanselowski et al. 2018)	0.604	0.487	0.151	0.780	**0.996**
UCLMR (Riedel et al. 2017)	0.583	0.479	0.114	0.747	0.989
FeatMLP (Riedel et al. 2017; Davis and Proctor 2017)	0.607	0.530	0.151	0.766	0.982
stackLSTM (Riedel et al. 2017; Hermans and Schrauwen 2013)	0.609	0.501	0.180	0.757	0.995
BERT	0.656	0.651	0.145	**0.839**	0.989
BAKE	0.734	0.667	0.463	0.822	0.986
exBAKE	**0.746**	**0.684**	**0.501**	0.813	0.988
Upper bound	0.754	0.588	0.667	0.765	0.997

The bold text shows the best results for various works

Table 11 Data distribution for the Yang et al. model

Dataset	Train	Dev	Test
XIAONIU (Task 1)	16,420	1026	4106
XIAONIU (Task 2)	16,671	1042	4172
FUN	251,415	N/A	61,794
HAHA	24,000	N/A	6000

The pre-training phase involved experimenting with additional news articles from CNN and Daily Mail on the BAKE model to obtain the exBAKE model. Evaluations carried out indicated a macro-averaged F1 score of 0.746, showing an improvement on the SOTA, as shown in Table 10.

The work by Yang et al. (2019) detected humor in paragraph texts by implementing a three-level approach of data augmentation with paragraph decomposition (PD), fine-tuning BERT with task-specific label and ensemble for inference. Three datasets, viz., CCL2019 Chinese Humor Computation (XIAONIU) dataset, the FUN (Blinov et al. 2019) dataset, and the authors used the HAHA (Humor Analysis based on Human Annotation) (Chiruzzo et al. 2019) dataset in evaluating their model. The XIAONIU dataset is a Chinese humor dataset made up of 21, 552 human written jokes mixed with machine written ones. It is also composed of 21, 885 jokes labeled in three levels for tri-classification problems. The FUN dataset is a Russian dataset containing 313, 210 samples collected from Russian social media websites and labeled binary classifications. The HAHA is a Spanish dataset that contains 30, 000 tweets, out of which 11, 595 is labeled as humorous. The data distribution as used in their work is shown in Table 11 Due to the unavailability of publicly accessible large datasets for the task, Yang et al. augmented their data by decomposing paragraphs into pairs of paragraphs while preserving joke labels. The decomposition also helped optimize BERT's performance, as long sequences of texts could not be effectively processed. BERT was adopted during the fine-tuning phase. The concatenation of the first token from

Table 12 Performance of the Yang et al. model with and without PD

Method	Xiaoniu		Fun	Haha
	Task 1	Task 2		
BERT (w/o PD)	0.8930	0.4889	0.9081	**0.7932**
BERT (w/ PD)	**0.8968**	**0.4936**	**0.9102**	0.7926

The bold text shows the best results for various works

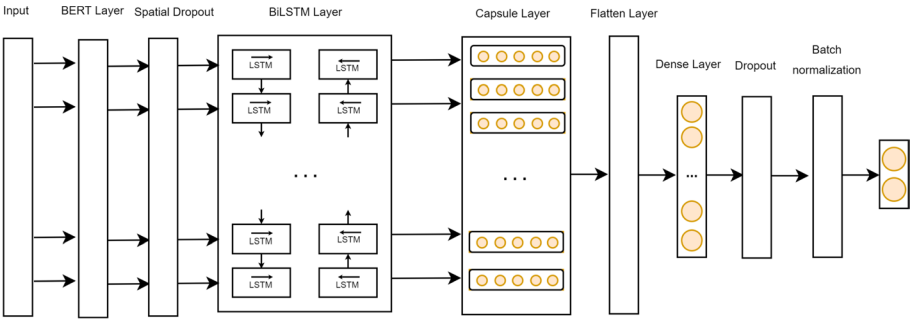


Fig. 14 The architecture of the BERT-Bi-LSTM-Capsule model (Vlad et al. 2019)

the last four layers was used as the input to the prediction head, and the WCE used to handle the data imbalances. They evaluated the performance of the datasets on their model with decomposing paragraphs and without decomposing paragraphs and obtained results indicated in Table 12

Vlad et al. (2019) proposed a model that classified emotions and detected propaganda and non-propaganda news articles in the NLP4IF-2019 shared Sentence Level Classification (SLC) task. Their architecture, as shown in Fig. 14, ensembled BERT, BiLSTM, and the Capsule Model. They used the BERT-large with pre-trained weights for their work. A single sentence was used as input to their BERT model. In the last encoder layer of the BERT, a spatial dropout was applied to avoid overfitting. The sequence $V \in \mathbb{R}^{t \times d}$, where t= number of encoded tokens, and d= dimension of the token, was extracted and fed as input to the Bi-LSTM layer. The Bi-LSTM (Schuster and Paliwal 1997), utilizing a sequence of hidden states, was used to process and extract the knowledge obtained from processing previous tokens and process current tokens. The resulting hidden states produced by each LSTM cell were concatenated together bi-directionally and passed to the capsule layer. The capsule layer's essence was to extract relevant features passed on to it from the previous networks. Their capsule network involved two layers, primary and convolutional. The primary capsule captured important features from previous layers, and the convolutional capsule checked the path between capsules. Dropout and batch normalization layers were applied to avoid overfitting. The output was passed to a dense layer for binary classification into propaganda or non-propaganda news. For the emotion detection phase, the BERT transformer layers were frozen to conserve the pre-trained weights, and the Bi-LSTM and capsule layers were fine-tuned. Their model was then trained on the SLC task dataset for the binary classification and the Daily dialogue unified with the Semeval-2019 task 3 dataset was used for emotion detection. The SLC task dataset contains 350 news articles labeled as propaganda and non-propaganda. The Daily dialogue dataset (Li et al. 2017) was built by crawling dialogues from regular human conversations. It contains 13, 118

sentences annotated for neutral, anger, disgust, fear, happiness, sadness, surprise discrete emotion labels. The SemEval-2019 Task 3 dataset (Chatterjee et al. 2019) contains 15, 000 emotion labeled texts for happy, sad, and angry, and an additional 15, 000 unannotated. Their model attained an F1 score of 0.5868 exceeding the SOTA by 0.1521. However, the model's performance could have been improved by introducing additional contextual embeddings such as ELMo and FLAIR (Akbik et al. 2018) to test the performance of the BERT-Emotion detection.

During their participation in the NLP4IF-2019 shared task SLC, Gupta et al. (2019) proposed the ensemble of neural architectures and BERT to detect propaganda in the data provided by the competition organizers. Their work could be separated into two viz., the SLC and Fragment Level Classification (FLC) (Da San Martino et al. 2019). The SLC focused on classifying sentences into propaganda and non-propaganda while the FLC was for token-level classification. In the SLC, because existing techniques had highlighted the harmful effects exaggerations, doubts, and other emotionally related words or phrases in sentences had on their performance, linguistic, layout, and topical features were extracted from the data. Then pre-trained vectors from FastText (Bojanowski et al. 2017) and BERT were used for word and sentence representations. The features were then passed to the logistic regression and CNN classifiers separately. In the CNN, the extracted features were concatenated in the last layer before classification was done. Then BERT fine-tuning for binary classification based on a given threshold value was applied. When the classifier's prediction probability exceeded the threshold value, it predicted propaganda; otherwise, it predicted non-propaganda. In the final component, all the logistic regression predictions, CNN, and BERT classifiers were collated using majority voting and open voting. One prediction was selected as the predicted class for a sentence. In their FLC task, sequence taggers (Vu et al. 2016; Gupta et al. 2016) in three modes, i.e., LSTM-CRF, LSTM-CRF+Multi-grain, and LSTM-CRF+Multitask, were designed. The LSTM-CRF handled the word and character level embedding and token-level feature extractions like the NER, POS, etc. The LSTM-CRF+Multi-grain collectively performed SLC and FLC using FastTextWordEmb and BERTSentEmb, respectively. LSTM-CRF+Multitask performed propagandistic fragment detection (PFD) and FLC. These sequence taggers were ensembled with BERT, considering the propagandistic fragments from each of the sequence taggers. Their model attained some gains in their features, ensemble schemes, multi-tasking, and multi-granularity architectures. However, they highlighted a better neural architecture capable of extracting salient fragments/keywords could improve performance (Gupta and Schütze 2018). Figure 15 depicts the proposed model by Gupta et al.

With the identification of speaker sensitive utterance being a significant challenge in Conversational Emotion Recognition (CER), Li et al. in Li et al. (2020) detected emotions in conversations using Multi-task Learning and BERT with emphasis on speaker-aware contextual representations. They used an attention-based hierarchical network encoder for two tasks viz. To predict the emotion label of utterances and classify utterances in conversations as speakers' utterances, i.e., Speaker Identification (SI). Their CER model was subdivided into two utterance encoder levels, i.e., the individual utterance encoder and the contextual utterance encoder. The individual utterance encoder generates a feature vector for each utterance achieved by inputting a sequence of utterances into their model and using BERT as their primary encoder. A single-layered bi-directional gated recurrent unit (Bi-GRU) was then integrated to extract word-level outputs better. An attention-based aggregation was then used to combine the word-level outputs into a feature vector. Before the individual utterance encoding, a sequence of utterance-level representations was obtained from which contextualized information was encoded using another Bi-GRU. It

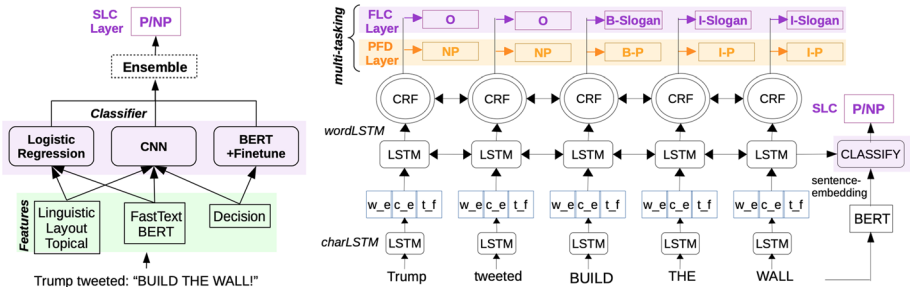


Fig. 15 Overview of the MIC-CIS architecture (Gupta et al. 2019)

produced the final contextualized feature presentation. The output probabilities of each of the emotion labels were calculated using a linear transformation and a softmax operation, as shown in Eq. (6). The model was then optimized by minimizing the predicted emotion distribution's cross-entropy and the actual emotion distribution. For the SI task, the same hierarchical network as CER was used to encode an utterance. Then the individual and contextual utterance encoders were used to obtain specific feature vectors and contextualized feature representations, respectively. With the primary goal of SI being a binary classification of determining which utterances were made by a speaker or not, pairs of utterances were randomly selected for classification. During the classification process, given a pair of utterances, $\boldsymbol{f}'_i$, and $\boldsymbol{f}'_j$, they selected four sources of features i.e., $\boldsymbol{f}'_i$, $\boldsymbol{f}'_j$, $\left\|\boldsymbol{f}'_i - \boldsymbol{f}'_j\right\|$, and $\boldsymbol{f}'_i \odot \boldsymbol{f}'_j$ where $\odot$ represented an element-wise multiplication, concatenated them, and applied a nonlinear Multi-layer Perceptron (MLP) layer to obtain a final feature vector, $\boldsymbol{f}^{SI}_{i,j}$. A binary classification was further performed on the derived final feature vector. The process was mathematically expressed as Eq. (7).

Their SI model was trained using the cross-entropy. Since the two tasks shared the same encoder network structure, multi-task learning was used to bridge the two networks for mutual interactions. Figure 16 shows the graphical view of their model. They evaluated their model on the MELD and EmoryNLP datasets whose distributions are as tabled out in Table 13. The Multimodal Multi-Party Dataset for Emotion Recognition in Conversation (MELD) (Poria et al. 2019) is a multi-modal dataset circling modality such as audio, video, and text. It contains over 1400 dialogues and 13, 000 utterances from the Friends Television Show with utterances in dialogues labeled categorically as anger, disgust, sadness, joy,

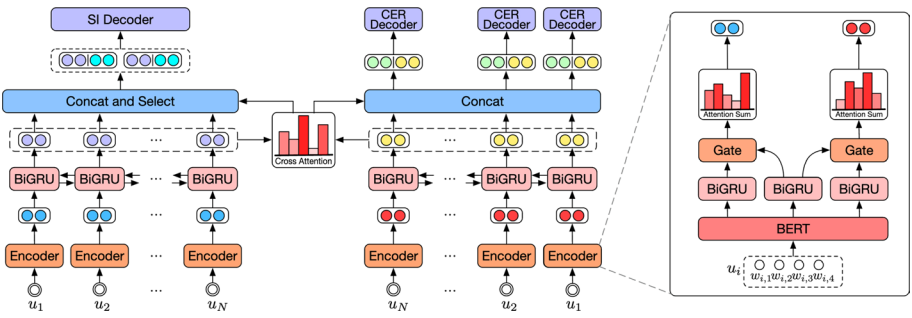


Fig. 16 The overview of the CER and SI architecture (Li et al. 2020)

Table 13 Features of the EmoryNLP, MELD and friends datasets

Dataset	#Conversation Train/val/test	#Utterances Train/val/test	#Avg. length
EmoryNLP	659/89/79	7551/954/984	11.5
MELD	1028/114/280	9989/1109/2610	9.6
Friends	3329	61,038	18.3

surprise, fear, and neutral. The EmoryNLP datasets (Zahiri and Choi 2018) contains a total of 12, 606 utterances from the Friends TV show and annotated for sad, mad, scared, powerful, peaceful, joyful, and neutral emotion classes. The authors compared their model with other SOTA models and attained results as shown in Table 14

$$\boldsymbol{p}_i^{CER} = softmax(\boldsymbol{W}_{CER}\boldsymbol{f}_i), \ \mathbf{s}.t.\ i \in [1, N] \tag{6}$$

$$\begin{aligned} \boldsymbol{\delta}_{i,j} &= \boldsymbol{f}'_i \oplus \boldsymbol{f}'_j \oplus (\left\| \boldsymbol{f}'_i - \boldsymbol{f}'_j \right\|) \oplus (\boldsymbol{f}'_i \odot \boldsymbol{f}'_j) \\ \boldsymbol{f}_{i,j}^{SI} &= RELU(\boldsymbol{W}_f \boldsymbol{\delta}_{i,j} + \boldsymbol{b}_f) \\ \boldsymbol{p}_{i,j}^{SI} &= softmax(\boldsymbol{W}_{SI}\boldsymbol{f}_{i,j}^{SI}) \end{aligned} \tag{7}$$

Luo and Wang (2019) adopted the BERT pre-trained model to classify emotions in the Friends and EmotionPush datasets during the EmotionX-2019, shared task. They utilized a two-step approach of Encoding utterances into vectors and then classifying them into the various emotion classes, i.e., joy, sadness, anger, and neutral using the softmax classifier. Their architecture is shown in Fig. 17. The pre-processed data samples solve the data's imbalances and then put them in the right format for further processing. To solve the imbalance problem, they identified that most utterances were categorized in the neutral class. Therefore, they selected utterances labeled with the emotions of interest, while extra neutral-labeled utterances were discarded. They also set a fixed sentence length of 128, and sentences more extended than the assigned length were truncated. Input sentences were

Table 14 Comparison of their results with others

Method	EmoryNLP	MELD
Our (GloVe)	32.57	59.67
+ MTL	**34.54**	**60.69**
Our (ELMo)	33.55	61.10
+ MTL	**34.85**	**61.86**
Our (BERT)	34.76	61.31
+ MTL	35.92	**61.90**
CNN	32.59	55.02
cLSTM	32.89	56.44
DialogueRNN	31.70	57.03
DialogueGCN	–	58.10
KET	34.39	58.18
ConGCN	–	**59.40**

The bold text shows the best results for various works

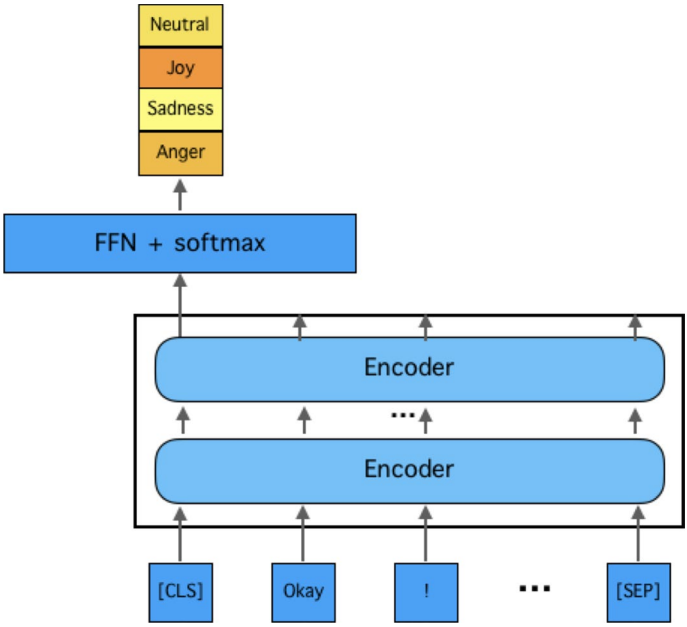


Fig. 17 The EmotionX-HSU architecture (Li et al. 2020)

then converted to lower case and tokenized using the WordPiece tokenizer. After preparing the data, it was fine-tuned using BERT. Having a batch size of 24 with a learning rate of $2e^{-5}$ was adopted, and the model was trained using an epoch size of 4. Their results (Table 15 and Table 16) were competitive in comparison with other models. The drawback identified with their model was the out of vocabulary (OOV) problem associated with the WordPiece tokenizer. Also, relative low training parameters meant that their model could suffer from overfitting if the parameters were slightly raised hence not robust.

Cambria et al. (2020) designed a commonsense knowledge-based architecture made up of an ensemble of Bi-LSTM and BERT so that deep learning architectures may possess some logical inference ability. They deconstructed words and multi-word expressions into primitives and super primitives. Primitives were automatically unearthed using the LSTM, and super primitives were obtained manually. Multi words containing prefixes and suffixes were represented in the form $[w1, \ldots w_{i-1}]$ and $[w_{i-1}, \ldots w_n]$, respectively, and $c = w_i$ as the

Table 15 Performance of the EmotionX-HSU model–friends dataset

	Precision	Recall	F1-score	Support
Neutral	0.801	0.914	0.854	1035
Joy	0.854	0.648	0.736	505
Sadness	0.608	0.512	0.556	121
Anger	0.662	0.638	0.650	141
micro avg	0.791	0.791	0.791	1802
macro avg	0.731	0.678	0.699	1802
weighted avg	0.792	0.791	0.785	1802

Table 16 Performance of the EmotionX-HSU model–EmotionPush dataset

	precision	recall	F1-score	Support
Neutral	0.908	0.917	0.913	2146
Joy	0.747	0.730	0.738	601
Sadness	0.627	0.627	0.627	110
Anger	0.474	0.333	0.391	27
micro avg	0.862	0.862	0.862	2884
macro avg	0.689	0.652	0.667	2884
weighted avg	0.860	0.862	0.861	2884

target word. These words were represented in vectors using the word2vec embedding, and a Bi-LSTM model was used to extract contextual information. An attention module followed by a softmax layer, was placed on top of the Bi-LSTM architecture to capture sub-phrasal contexts, as shown in Fig. 18. The BERT architecture was then used to obtain the context embedding of words in a sentence by initially fine-tuning BERT on the ukWaC corpus (Baroni et al. 2009)and then computing the context embedding with the target words first removed and the rest of the sentences fed into the BERT architecture. The cosine distance was used to calculate how similar other words in the sentence were to both the contextual embedding and the target word. AffectiveSpace (Cambria et al. 2015) was then used to locate words and multi-word expressions automatically. The regularized *k*-means (RKM) was used to obtain the discrete path between a key polar state and its opposite. The model was evaluated on six benchmark datasets, i.e., the semantic text similarity (STS) dataset consisting of 632 positive and 1402 negative tweets (Saif et al. 2013), the Stanford Sentiment Analysis Treebank (SST) dataset containing 4871 positive and 4650 negative movie reviews (Socher et al. 2013), the SemEval-2013 dataset made up of 5349 positive and 2186 negative tweets. The SemEval-2015 Task 10 dataset was built consisting of 5809 positive

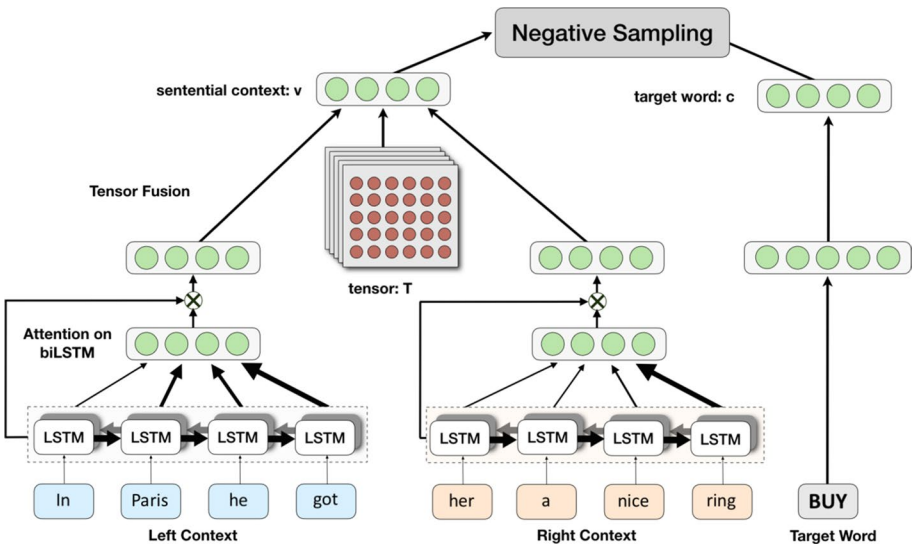


Fig. 18 context and word embedding generation framework (Cambria et al. 2020)

and 2407 negative tweets making a total of 15,195 tweets, the SemEval-2016 task 4 dataset containing 13,942 positive and 3697 negative tweets, respectively, and the Sanders dataset consisting of 5512 tweets on four distinct topics out of which 570 are positive, and 654 are negative. These datasets were used to compare SenticNet 6 with 15 sentiment lexica, and results, as indicated in Table 17, were obtained.

Kazameini et al. (2020) extracted contextualized word embeddings from text data using BERT and the bagged-SVM classifier to predict the personality traits of authors automatically. The input to their model was essays with an average word size of 650. Since BERT can only process 512 input tokens, the essays were divided into sub-documents. The sub-documents were pre-processed and fed into the BERT base model. Feature vectors for the document were obtained by computing the mean of each of the 12 BERT layers' contextual token representations. The last four-layer representations were then concatenated with the corresponding 84 Mairesse features for the essay. The feature vector was then fed into ten SVM classifiers, producing a prediction and the final prediction obtained by majority voting. Figure 19 graphically depicts their model. They obtained an increased performance of 1.04% in comparison with baseline methods.

Using the Big Five Essay dataset and the Myers-Briggs' type indicator Kaggle Dataset, Mehta et al. (2020) proposed integration of deep learning models and traditional psycholinguistic features with language model embeddings for the detection of personality traits. The Essay dataset consisted of 2468 essays written by students, while the Kaggle dataset contained 8675 records of the last 50 things people posted on a PersonalityCafe forum website. Psycholinguistic and language model features were then extracted from the acquired datasets.They extracted psycholinguistic features using the Mairesse (Mairesse et al. 2007), SenticNet (Cambria et al. 2018), NRC Emotion Lexicon (Mohammad and Turney 2013), the VAD lexicon (Mohammad 2018) and the Readability. Language model features were then extracted using the BERT. Their experiments compared the performance of BERT-base and BERT-large used in synergy with the SVM or Multi-layer Perceptron (MLP). BERT-base + MLP yielded an average score of 60.6 on the Essay dataset, while BERT-large + MLP yielded an average score of 77.1 on the Kaggle dataset. However, they reported that using a dataset that provided a continuous personality trait score could have improved model performance. Also, their model suffered an overestimation of the performance problem.

Fadel et al.'s JUSTers model evaluated the performance of five pre-trained transformer models against commonsense validation and explanation at the SemEval 2020 Task 4 competition (Fadel et al. 2020). Using the dataset provided for the competition, they performed the first task of commonsense validation by feeding two similar statements to their model and letting the model predict the statement that makes sense. They evaluated and fine-tuned the BERT, RoBERTa, and ALBERT (Lan et al. 2019) language models on this task by considering the task as a binary classification problem and assigning labels 1 and 0 to all statements that had commonsense conformities and statements that did not, respectively. Using these models independently, they obtained the probabilities of correct and wrong statements. The statement with the least probability was the one at odds with commonsense. Task 2 involved feeding statements and concatenated reasons into pre-trained XLNET, BERT, and RoBERTa models for the models to choose the right reason why a statement is against commonsense or not. This task was inferred by obtaining the probabilities using a softmax layer. In the last task, statements were fed into a GPT-2 model, and after fine-tuning, the model was expected to explain why a particular statement contradicted commonsense or otherwise. Their models obtained an accuracy of 92.90% and 92.30% for tasks 1 and 2, respectively, and a BLEU score of 16.10 for task 3.

Table 17 Model performance of senticnet 6 in comparison with other models

Model	Year	SST Dataset (%)	STS Dataset (%)	SemEval-2013 (%)	SemEval-2015 (%)	SemEval-2016 (%)	Sanders (%)
ANEW (Bradley and Lang 1999)	1999	31.21	37.05	42.83	33.13	42.27	27.61
WordNet-Affect (Strapparava et al. 2004)	2004	04.51	11.98	03.83	03.27	03.55	05.64
Opinion Lexicon (Ferrarotti et al. 2018)	2004	**54.21**	61.00	41.08	43.15	37.95	54.58
Opinion Finder (Wilson et al. 2005)	2005	53.60	55.71	47.61	43.97	46.90	46.98
Micro WNOp (Cerini et al. 2007)	2007	15.45	18.94	19.18	16.97	17.92	15.69
Sentiment140 (Go et al. 2009)	2009	**57.74**	**67.97**	45.69	**50.92**	41.75	**64.30**
SentiStrength (Thelwall et al. 2010)	2010	36.76	51.81	37.37	41.51	34.08	45.34
SentiWordNet (Baccianella et al. 2010)	2010	50.19	49.30	50.11	50.51	**49.62**	43.95
General Inquirer (Taboada et al. 2011)	2011	25.91	11.14	16.12	12.47	16.83	10.62
AFINN (Nielsen 2011)	2011	44.81	58.77	43.88	44.99	40.22	53.68
EmoLex (Mohammad and Turney 2013)	2013	46.95	47.63	45.22	42.33	42.47	44.53
NRC HS Lexicon (Zhu et al. 2014)	2014	47.09	49.86	28.56	42.54	25.27	54.25
VADER (Gilbert and Hutto 2014)	2014	50.74	**65.18**	**50.42**	49.08	45.95	**57.76**
MPQA (Deng and Wiebe 2015)	2015	53.71	55.43	46.86	43.97	45.52	46.57
SenticNet 5 (Cambria et al. 2018)	2018	53.61	55.99	**68.17**	**55.83**	**70.75**	48.04
SenticNet 6 (Riedel et al. 2017; Hermans and Schrauwen 2013)	2020	**75.43**	**83.82**	**81.79**	**80.19**	**82.23**	**77.62**

The bold text shows the best results for various works

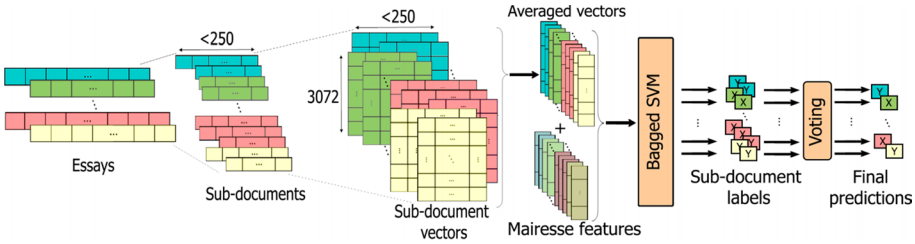


Fig. 19 The proposed architecture for the automatic personality detection (Kazameini et al. 2020)

A summary of the discussed state of the art is presented in Table 18 at the end of the section.

6 Conclusion and future work

The paper elucidated transformer-based models currently producing the SOTA in NLP applications. The reviewed transformer-based models comprised the Generative Pre-training (GPT) models, Transformer-XL model, Cross-lingual models (XLM), and the Bi-directional Encoder Representations from Transformers (BERT). The discourse of these models showed the workings, strengths, and weaknesses of the various transformer models. The paper further identified current applications of BERT-based models for the detection of emotions in text. The discussion of the identified BERT-based models highlighted contributions, datasets used, results, and limitations of the various models.

Some observations were made from the recent works discussed to encourage future work in text-based emotion detection. As at the time of this paper, BERT is the most explored transformer-based model for text-based emotion detection. However, considering the solutions the various BERT variants present to resolve BERT's associated weaknesses, the paper recommends exploring these BERT variants in detecting emotions in textual data. An ensemble of these variants using various ensemble techniques is strongly recommended to provide a generalized and diversified output. The XLNet significantly outperforms other BERT variants in related NLP tasks, but it is computationally expensive to use. It is recommended that distillation techniques (Tang et al. 2019) be applied to the original model. This paper also recommends exploring other transformer-based models to detect emotions in text-based data to determine their effectiveness in the field. In view that the field of text-based emotion detection has an enormous amount of unlabeled data in comparison to labeled datasets, a robust semi-supervised learning architecture could be efficient for an in-depth exploration of this available data (Hussain and Cambria 2018). The use of the GPT model for this purpose is highly encouraged; this is because GPT possesses lexical robustness and is being used in related tasks to produce current state of the art.

Table 18 Summary of current advances in text-based emotion detection

Proposed Work	Dataset(s) Used	Contributions	Results	Limitations
Yang et al. (2019)	EmotionLines	Showed the performance of the uncased BERT-base model	Observed that post-training the uncased BERT-base model with specific domain corpus predicts contextual emotions better as shown in Tables 5 and 6	Fixed length input size limitations
		Derived a method to deal with context understanding, domain adaptation and class imbalances in the dataset used		
Huang et al. (2019a)	Friends	Investigate the emotion recognition ability of BERT	Obtained an F1 score of 81.5 and 88.5 for Friends and EmotionPush datasets, respectively	Low level of inference
	EmotionLines			Fixed input size limitation
				Disregard for personality embedding
Huang et al. (2019b)	Tweets	Classified emotions using HRLCE and BERT	Obtained an F1 score of 0.7709	High misclassification rate for the '*Others* class
Malte and Ratadiya (2019)	TRAC-1	Detected cyber abuse in English and Hindi texts using BERT	Obtained an F1 score of 0.6244 and 0.6596 on English and Hindi texts, respectively	English slang were loosely handled
				Suffered from the OOV problem
Park et al. (2019)	ISEAR	Classified emotions categorically and predicted the VAD scores of dimensional emotions using BERT-large	Obtained a micro F1 score of 0.688 and 0.695 for categorically classifying texts in the ISEAR and SemEval data and obtained VAD scores of 0.659, 0.327, and 0.287, respectively	Low VAD inference without VAD annotations.
	SemEval			
	Emobank			
Matero et al. (2019)	Reddit	Assessed suicide risks using multi-level dual-context language and BERT	Obtained results as shown in Table 9	Suffered from the fixed input length limitation

Table 18 (continued)

Proposed Work	Dataset(s) Used	Contributions	Results	Limitations
Jwa et al. (2019)	News Headlines	Used BERT for fake news detection	Obtained a macro averaged F1 score of 0.746	Fixed input length size limitation
Yang et al. (2019)	XIAONIU	Detected humor in paragraph texts using BERT	Obtained results as shown in Table 12	Better decomposition strategies will improve performance
	FUN			
	HAHA			
Vlad et al. (2019)	SLC task	Presented an ensemble of BERT, bi-LSTM and Capsule model to detect emotions and propaganda.	Obtained a micro F1 score of 0.5868	Additional contextualized embeddingss could improve results
	Daily dialogue			
	and SemEval-2019 task 3			
Li et al. (2020)	MELD	Detected emotions using multi-task learning and BERT	Obtained results as indicated in Table 14	Limited input fixed size
	EmoryNLP			
Luo and Wang (2019)	Friends	Fine-tuned BERT to detect emotions	Obtained a micro F1 score of 0.79 and 0.86 for Friends and EmotionPush, respectively, as shown in Table 14	imbalances in the data affected model's performance
	EmotionPush			Fixed input length size limitation
Cambria et al. (2020)	STS	Infused logical reasoning into an ensemble of Bi-LSTM and BERT architecture	Obtained results as indicated in Table 17	Obtaining super primitives automatically could improve on the model results
	SST			
	SemEval-2013			
	SemEval-2015 Task 10			
	SemEval-2016 task 4			
	Sanders			

Table 18 (continued)

Proposed Work	Dataset(s) Used	Contributions	Results	Limitations
Kazameini et al. (2020)	Essays	Predicted the personality traits of authors using BERT word embeddings and bagged SVM	Obtained an average accuracy of 59.03%	– The limited quantum of data used affected the model's performance
Mehta et al. (2020)	Essay and Kaggle Datasets	Detected personality traits using deep learning models and psycholinguistic features	Obtained an average score of 60.6 and 71.1 on Essays and Kaggle datasets, respectively	– Negation biases affected the performance of the model – Model may have suffered an inadequacy of data
Fadel et al. (2020)	SemEval 2020 Task 4 Dataset	Evaluated the performance of five pre-trained transformer models against commonsense validation and explanation tasks	Obtained an accuracy of 92.90% and 92.30% for tasks 1 and 2, respectively, and a BLEU score of 16.10 for task 3	Model may have suffered an overestimation of performance problem

Furthermore, contributions towards resolving polarity ambiguation in emotion-conveying words remain limited and are greatly encouraged.

References

Acheampong FA, Wenyu C, Nunoo-Mensah H (2020) Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports e12189

Akbik A, Blythe D, Vollgraf R (2018) Contextual string embeddings for sequence labeling. In: Proceedings of the 27th international conference on computational linguistics, pp 1638–1649

Akhtar MS, Ekbal A, Cambria E (2020) How intense are you? predicting intensities of emotions and sentiments using stacked ensemble. IEEE Comput Intell Mag 15(1):64–75

Al-Rfou R, Choe D, Constant N, Guo M, Jones L (2019) Character-level language modeling with deeper self-attention. Proc AAAI Conf Artif Intell 33:3159–3166

Baccianella S, Esuli A, Sebastiani F (2010) Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. Lrec 10:2200–2204

Baird S, Doug S, Pan Y (2017) Talos targets disinformation with fake news challenge victory. (2017), URL https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html

Baroni M, Bernardini S, Ferraresi A, Zanchetta E (2009) The wacky wide web: a collection of very large linguistically processed web-crawled corpora. Lang Resour Eval 43(3):209–226

Baziotis C, Pelekis N, Doulkeridis C (2017) Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017), pp 747–754

Blinov V, Bolotova-Baranova V, Braslavski P (2019) Large dataset and language model fun-tuning for humor recognition. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 4027–4032

Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146

Bradley MM, Lang PJ (1999) Affective norms for english words (anew): Instruction manual and affective ratings. Tech Report C-1, Center Res Psychophysiol 30(1):25–36

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. arXiv 75

Buechel S, Hahn U (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: vol 2, Short Papers, pp 578–585

Cambria E, Livingstone A, Hussain A (2012) The hourglass of emotions. In: Cognitive behavioural systems. Springer, pp 144–157

Cambria E, Fu J, Bisio F, Poria S (2015) Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis. In: AAAI, pp 508–514

Cambria E, Poria S, Hazarika D, Kwok K (2018) Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In: Thirty-second AAAI conference on artificial intelligence, pp 1795–1802

Cambria E, Li Y, Xing FZ, Poria S, Kwok K (2020) Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In: Proceedings of the 29th ACM international conference on information & knowledge management, pp 105–114

Cerini S, Compagnoni V, Demontis A, Formentelli M, Gandini G (2007) Language resources and linguistic theory: typology, second language acquisition, english linguistics, chapter micro-wnop: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. Franco Angeli Editore, Milano, IT, pp 200–210

Chatterjee A, Narahari KN, Joshi M, Agrawal P (2019) Semeval-2019 task 3: Emocontext contextual emotion detection in text. In: Proceedings of the 13th international workshop on semantic evaluation, pp 39–48

Chen SY, Hsu CC, Kuo CC, Huang K, Ku LW (2019) Emotionlines: An emotion corpus of multi-party conversations. In: 11th international conference on language resources and evaluation, LREC 2018. European language resources association (ELRA), pp 1597–1601

Chiruzzo L, Castro S, Etcheverry M, Garat D, Prada JJ, Rosá A (2019) Overview of haha at iberlef 2019: Humor analysis based on human annotation. In: Proceedings of the Iberian languages

evaluation forum (IberLEF 2019). CEUR workshop proceedings, CEUR-WS, Bilbao, Spain (9 2019), pp 132–144
Conneau A, Lample G (2019) Cross-lingual language model pretraining. In: Advances in Neural Information Processing Systems, pp 7057–7067
Dai Z, Yang Z, Yang Y, Carbonell JG, Le Q, Salakhutdinov R (2019) Transformer-xl: Attentive language models beyond a fixed-length context. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 2978–2988
Da San Martino G, Yu S, Barrón-Cedeno A, Petrov R, Nakov P (2019) Fine-grained analysis of propaganda in news article. 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), pp 5640–5650
Davis R, Proctor C (2017) Fake news, real consequences: Recruiting neural networks for the fight against fake news. Stanford CS224d Deep Learning for NLP final project, p 8
Deng L, Wiebe J (2015) Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 179–189
Devlin J, Chang M-W, Lee K, Toutanova K (June 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, vol 1 (Long and Short Papers), (Minneapolis, Minnesota). Association for computational linguistics, pp 4171–4186
Du K-L, Swamy MN (2013) Neural networks and statistical learning. Springer Science & Business Media, Berlin
Ekman P (1999) Basic emotions. Handbook Cogn Emotion 98(45–60):16
Fadel A, Al-Ayyoub M, Cambria E (2020) Justers at semeval-2020 task 4: Evaluating transformer models against commonsense validation and explanation. In: SemEval-2020, p 9
Felbo B, Mislove A, Søgaard A, Rahwan I, Lehmann S (2017) Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 1615–1625
Ferrarotti MJ, Rocchia W, Decherchi S (2018) Finding principal paths in data space. IEEE Trans Neural Netw Learn Syst 30(8):2449–2462
Gilbert C, Hutto E (2014) Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth international conference on weblogs and social media (ICWSM-14). Available at (20/04/16) http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf, vol. 81, p 82
Gobinda G (2003) Natural language processing. Ann Rev Inf Sci Technol 37:1
Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224N Project Rep Stanford 1(12):2009
Gupta P, Schütze H (2018) Lisa: Explaining recurrent neural network judgments via layer-wise semantic accumulation and example to pattern transformation. In: Proceedings of the 2018 EMNLP workshop BlackboxNLP: analyzing and interpreting neural networks for NLP, pp 154–164
Gupta P, Schütze H, Andrassy B (2016) Table filling multi-task recurrent neural network for joint entity and relation extraction. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers, pp 2537–2547
Gupta P, Saxena K, Yaseen U, Runkler T, Schütze H (2019) Neural architectures for fine-grained propaganda detection in news. In: Proceedings of the second workshop on natural language processing for internet freedom: Censorship, Disinformation, and Propaganda, pp 92–97
Hanselowski A, Avinesh P, Schiller B, Caspelherr F, Chaudhuri D, Meyer CM, Gurevych I (2018) A retrospective analysis of the fake news challenge stance-detection task. In: Proceedings of the 27th international conference on computational linguistics, pp 1859–1874
Hermans M, Schrauwen B (2013) Training and analysing deep recurrent neural networks. In: Advances in neural information processing systems, pp 190–198
Hou L, Yu C-P, Samaras D (2016) Squared earth mover's distance-based loss for training deep neural networks. arXiv preprint arXiv:1611.05916, p 9
Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. In: Proceedings of the 56th annual meeting of the association for computational linguistics, vol 1: Long Papers, pp 328–339
Huang Y-H, Lee S-R, Ma M-Y, Chen Y-H, Yu Y-W, Chen Y-S (2019) Emotionx-idea: Emotion bert–an affectional model for conversation, arXiv preprint arXiv:1908.06264, p 6
Huang C, Trabelsi A, Zaiane OR (2019) Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert. In: Proceedings of the 13th international workshop on semantic evaluation, pp 49–53

Hussain A, Cambria E (2018) Semi-supervised learning for big social data analysis. Neurocomputing 275:1662–1673

Jwa H, Oh D, Park K, Kang JM, Lim H (2019) exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). Appl Sci 9(19):4062

Kao EC-C, Liu C-C, Yang T-H, Hsieh C-T, Soo V-W (2009) Towards text-based emotion detection a survey and possible improvements. In: 2009 International conference on information management and engineering. IEEE, pp 70–74

Kazameini A, Fatehi S, Mehta Y, Eetemadi S, Cambria E (2020) Personality trait detection using bagged svm over bert word embedding ensembles, arXiv preprint arXiv:2010.01309, p 4

Khosla S (2018) Emotionx-ar: Cnn-dcnn autoencoder based emotion classifier. In: Proceedings of the sixth international workshop on natural language processing for social media, pp 37–44

Kumar R, Ojha AK, Malmasi S, Zampieri M (2018) Benchmarking aggression identification in social media. In: Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), pp 1–11

Lample G, Ott M, Conneau A, Denoyer L, Ranzato M (2018) Phrase-based & neural unsupervised machine translation. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 5039–5049

Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) Albert: A lite bert for self-supervised learning of language representations. In: International conference on learning representations, p 17

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach, arXiv:abs/1907.11692, p 13

Li Y, Su H, Shen X, Li W, Cao Z, Niu S (2017) Dailydialog: A manually labelled multi-turn dialogue dataset. In: Proceedings of the eighth international joint conference on natural language processing, vol 1: Long Papers, pp 986–995

Li J, Zhang M, Ji D, Liu Y (2020) Multi-task learning network for emotion recognition in conversation. arXiv preprint arXiv:2003.01478, p 7

Luo L, Wang Y (2019) Emotionx-hsu: Adopting pre-trained bert for emotion classification, arXiv preprint arXiv:1907.09669, p 4

Mairesse F, Walker MA, Mehl MR, Moore RK (2007) Using linguistic cues for the automatic recognition of personality in conversation and text. J Artif Intell Res 30:457–500

Malte A, Ratadiya P (2019) Multilingual cyber abuse detection using advanced transformer architecture. In: TENCON 2019-2019 IEEE region 10 conference (TENCON). IEEE, pp 784–789

Matero M, Idnani A, Son Y, Giorgi S, Vu H, Zamani M, Limbachiya P, Guntuku SC, Schwartz HA (2019) Suicide risk assessment with multi-level dual-context language and bert. In: Proceedings of the sixth workshop on computational linguistics and clinical psychology, pp 39–44

Mehta Y, Fatehi S, Kazameini A, Stachl C, Cambria E, Eetemadi S (2020) Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In: 20th IEEE international conference on data mining (ICDM), p 6

Mohammad S (2018) Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In: Proceedings of the 56th annual meeting of the association for computational linguistics, vol 1: Long Papers, pp 174–184

Mohammad SM, Turney PD (2013) Crowdsourcing a word-emotion association lexicon. Comput Intell 29(3):436–465

Mohammad S, Bravo-Marquez F, Salameh M, Kiritchenko S (2018) Semeval-2018 task 1: Affect in tweets. In: Proceedings of the 12th international workshop on semantic evaluation, pp 1–17

Murugesan S (2007) Understanding web 2.0. IT Prof 9(4):34–41

Nielsen FÅ (2011) A new anew: Evaluation of a word list for sentiment analysis in microblogs. In: 1st Workshop on making sense of Microposts, pp 93–98

Ortony A, Clore GL, Collins A (1990) The cognitive structure of emotions. Cambridge University Press, Cambridge

Park S, Kim J, Jeon J, Park H, Oh A (2019) Toward dimensional emotion detection from categorical emotion annotations, arXiv preprint arXiv:1911.02499, p 11

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

Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Proceedings of NAACL-HLT, pp 2227–2237

Plutchik R (1980) A general psychoevolutionary theory of emotion. In: Theories of emotion. Elsevier, pp 3–33

Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R (2019) Meld: A multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 527–536
Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training, URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf, p 12
Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8):9
Riedel B, Augenstein I, Spithourakis G, Riedel S (2017) A simple but tough-to-beat baseline for the fake news challenge stance detection task. corr arXiv:abs/1707.03264
Ruder S, Peters ME, Swayamdipta S, Wolf T, (2019) Transfer learning in natural language processing. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, pp 15–18
Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39(6):1161
Russell JA, Mehrabian A (1977) Evidence for a three-factor theory of emotions. J Res Pers 11(3):273–294
Saif H, Fernandez M, He Y, Alani H (2013) Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold. In: Proceedings of the 1st interantional workshop on emotion and sentiment in social and expressive media: approaches and perspectives from AI (ESSEM 2013), p 9
Scherer KR, Wallbott HG (1994) Evidence for universality and cultural variation of differential emotion response patterning. J Pers Soc Psychol 66(2):310
Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11):2673–2681
Schwartz HA, Giorgi S, Sap M, Crutchley P, Ungar L, Eichstaedt J (2017) Dlatk: Differential language analysis toolkit. In: Proceedings of the 2017 conference on empirical methods in natural language processing: System demonstrations, pp 55–60
Shing H-C, Nair S, Zirikly A, Friedenberg M, Daumé III H, Resnik P (2018) Expert, crowdsourced, and machine assessment of suicide risk via online postings. In: Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic, pp 25–36
Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng AY, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 1631–1642
Sordoni A, Bengio Y, Vahabi H, Lioma C, Grue Simonsen J, Nie J-Y (2015) A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In: Proceedings of the 24th ACM international on conference on information and knowledge management, pp 553–562
Strapparava C, Valitutti A, et al. (2004) "Wordnet affect: an affective extension of wordnet. In: Lrec, vol. 4. Citeseer, p 40
Sundermeyer M, Schlüter R, Ney H (2012) Lstm neural networks for language modeling. In: Thirteenth annual conference of the international speech communication association, p 4
Susanto Y, Livingstone AG, Ng BC, Cambria E (2020) The hourglass model revisited. IEEE Intell Syst 35(5):96–102
Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. Comput Linguis 37(2):267–307
Tang R, Lu Y, Liu L, Mou L, Vechtomova O, Lin J (2019) Distilling task-specific knowledge from bert into simple neural networks. arXiv 8
Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A (2010) Sentiment strength detection in short informal text. J Am Soc Inform Sci Technol 61(12):2544–2558
Trinh TH, Le QV (2018) A simple method for commonsense reasoning. arXiv 12
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
Vlad G-A, Tanase M-A, Onose C, Cercel D-C (2019) Sentence-level propaganda detection in news articles with transfer learning and bert-bilstm-capsule model. In: Proceedings of the second workshop on natural language processing for internet freedom: Censorship, Disinformation, and Propaganda, pp 148–154
Vu NT, Adel H, Gupta P, et al. (2016) Combining recurrent and convolutional neural networks for relation classification. In: Proceedings of NAACL-HLT, pp 534–539
Wang S, Peng G, Zheng Z, Xu Z (2019) Capturing emotion distribution for multimedia emotion tagging. IEEE Trans Affect Comput p 11

Wilson T, Hoffmann P, Somasundaran S, Kessler J, Wiebe J, Choi Y, Cardie C, Riloff E, Patwardhan S (2005) Opinionfinder: A system for subjectivity analysis. In: Proceedings of HLT/EMNLP 2005 interactive demonstrations, pp 34–35

Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser Łukasz, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M, Dean J (2016) Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR, abs/1609.08144:23. http://arxiv.org/abs/1609.08144

Xu H, Liu B, Shu L, Yu P (2019) Bert post-training for review reading comprehension and aspect-based sentiment analysis. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, vol. 1, p 12

Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 1480–1489

Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV (2019) Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems, pp 5753–5763

Yang K, Lee D, Whang T, Lee S, Lim H (2019) Emotionx-ku: Bert-max based contextual emotion classifier. CoRR, arXiv:abs/1906.11565, p 6

Yang H, Deng Y, Wang M, Qin Y, Sun S (2019) Humor detection based on paragraph decomposition and bert fine-tuning. In: Reasoning for complex QA workshop 2020, p 4

Yue L, Chen W, Li X, Zuo W, Yin M (2019) A survey of sentiment analysis in social media. Knowl Inf Sys 60(2):617–663

Zahiri SM, Choi JD (2018) Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In: Workshops at the thirty-second aaai conference on artificial intelligence, p 10

Zhu X, Kiritchenko S, Mohammad S (2014) Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), pp 443–447

Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, Fidler S (2015) Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE international conference on computer vision, pp 19–27

Zirikly A, Resnik P, Uzuner O, Hollingshead K (2019) "Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In: Proceedings of the sixth workshop on computational linguistics and clinical psychology, pp 24–33