Deep Learning in Sentiment Analysis: Recent Architectures

TARIQ ABDULLAH and AHMED AHMET, University of Derby, United Kingdom

Humans are increasingly integrated with devices that enable the collection of vast unstructured opinionated data. Accurately analysing subjective information from this data is the task of sentiment analysis (an actively researched area in NLP). Deep learning provides a diverse selection of architectures to model sentiment analysis tasks and has surpassed other machine learning methods as the foremast approach for performing sentiment analysis tasks. Recent developments in deep learning architectures represent a shift away from Recurrent and Convolutional neural networks and the increasing adoption of Transformer language models. Utilising pre-trained Transformer language models to transfer knowledge to downstream tasks has been a breakthrough in NLP.

This survey applies a task-oriented taxonomy to recent trends in architectures with a focus on the theory, design and implementation. To the best of our knowledge, this is the only survey to cover state-of-the-art Transformer-based language models and their performance on the most widely used benchmark datasets. This survey paper provides a discussion of the open challenges in NLP and sentiment analysis. The survey covers five years from 1st July 2017 to 1st July 2022.

CCS Concepts: • **Computing methodologies** → Natural language processing; **Neural networks;**

Additional Key Words and Phrases: Deep learning, sentiment analysis, cross-lingual sentiment analysis, cross-domain sentiment analysis, transfer learning, multilingual sentiment analysis

ACM Reference format:
Tariq Abdullah and Ahmed Ahmet. 2022. Deep Learning in Sentiment Analysis: Recent Architectures. *ACM Comput. Surv.* 55, 8, Article 159 (December 2022), 37 pages.
https://doi.org/10.1145/3548772

1 INTRODUCTION

An unprecedented volume of unstructured opinionated text is being generated from diverse sources such as social media, search engines, web forums, blogs, and e-commerce [122, 140]. Accurately obtaining value from this wealth of data is an important problem for both industry and academia. Sentiment analysis is an actively researched area in **NLP (Natural Language Processing)** and focuses on categorising and identifying subjective information from structured and unstructured text. It plays an increasingly important role by unlocking valuable business intelligence in areas such as brand monitoring [97, 109], the voice of the customer [127], workforce analytics [137], product analytics [67], and market research [38]. Furthermore, the perils of social media

Tariq Abdullah and Ahmed Ahmet contributed equally to this research.

Authors' address: T. Abdullah and A. Ahmet, University of Derby, Kedleston Rd, Derby, Derbyshire, United Kingdom, DE22 1GB; emails: t.abdullah@derby.ac.uk, ahmedahmetk@hotmail.co.uk.

0360-0300/2022/12-ART159 $15.00
https://doi.org/10.1145/3548772

have come into focus in recent years, motivating a strong demand for content to be labelled using an automated approach to tackle problems such as cyberbullying [103], hate speech [19], bot detection [9], and spread of misinformation [65]. Sentiment analysis tools can also be extended for predicting and recommending products/services for individuals [180]. Studies applying sentiment analysis on social media content (like Twitter) have shown a strong correlation with opinion polls [105] that even outperform public opinion polls in large election cycles [118].

Deep learning became an increasingly dominant branch of machine learning. It is based on **artificial neural networks (ANNs)** and maintains strong growth within both industry and academia [74]. The accelerated adoption of ANNs is influenced by recent advancements in processing power, availability of larger training datasets, and innovations in deep learning architectures. It is one of the hottest technology trends in recent years [46]. The alignment of these factors provided the environment for unparalleled performance in key areas such as computer vision, speech recognition and NLP, among others [59].

Recent surveys into sentiment analysis provide comprehensive overview of different approaches including machine learning, lexicon-based, and hybrid [20, 179]. The rapidly evolving field of deep learning is reshaping NLP and sentiment analysis domains, and there is a need for an up-to-date survey of deep learning and sentiment analysis to focus on the following research questions:

- Recent trends and advances in Transformer language model architectures and their applications in sentiment analysis tasks.
- Theory, design and implementation of deep learning architectures and how they are incorporated in different NLP and sentiment analysis use cases.
- A task-oriented taxonomy of sentiment analysis at different granularities of sentiment analysis use cases.
- Survey of the most widely used benchmark datasets for sentiment analysis taxonomy.
- Open challenges in sentiment analysis and recommendations for meeting these challenges.

In the remainder of this section, a sentiment analysis workflow with a brief survey of word embedding techniques is provided to better understand the rest of the paper. Section 2 summarises deep learning architectures widely used in sentiment analysis to provide concise descriptions of the theory, design, and implementation of deep learning architectural trends. Section 3 details the search methods and inclusion criteria of the survey methodology. A task-oriented sentiment analysis taxonomy with coarse-grain, fine-grain, cross-domain and cross-lingual broader categories is also explained in this section. Sections 4, 5, 6, and 7 cover a survey of studies and benchmark datasets for the aforementioned sentiment analysis categories. Section 8 reviews open challenges, and the paper is concluded in Section 9.

1.1 Deep Learning Sentiment Analysis Workflow

Before textual input data can be processed by ANNs, it must first be converted into real-number input representations (word embeddings). This conversion is completed with the following three transformations: (1) optional text pre-processing, (2) tokenization, and (3) mapping tokens (words/phrases in vocabulary) to numerical word embedding vector representations. A forward pass in a neural network will propagate the word embeddings through the network parameters (refer to **1**). In supervised learning, forward pass predicts output which is used to calculate loss using the predicted and the desired outcome. This is then used in backpropagation to update network parameters. A detailed description of backpropagation is beyond the scope of this paper and interested readers are encouraged to access related literature [130].

Word embeddings have proven to be a significant innovation in NLP with performance gains over non-embedding approaches [157, 184]. They are initially pre-trained on language modelling

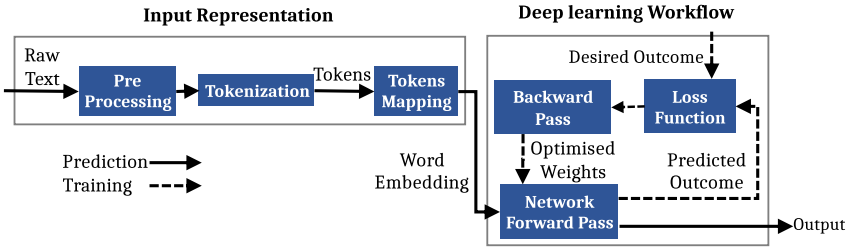

Fig. 1. Supervised sentiment analysis workflow with deep learning.

tasks before they can be applied to sentiment analysis tasks (downstream tasks). Prominent monolingual and cross-lingual embeddings and the techniques applied to transfer linguistic knowledge are explained below.

1.2 Word Embeddings

Low-dimensional vector representations or word embeddings have become the universal approach for textual input representations into a neural network. Embedding techniques are generally grouped into machine learning and statistics based.

1.2.1 Monolingual Word Embeddings. Monolingual word embedding techniques were initially popularised with word2vec [100]. Subsequent approaches include GloVe [112] and FastText [22]. Word2vec captures semantic features with skip-gram and **continuous bag-of-words (CBOW)** models. Both are shallow language models [17] and use a feedforward neural network with an input, context, a hidden layer, and an output layer. The input context is encoded as a one-hot vector with the size of the vocabulary. FastText is an extension of the word2vec model processing each word as an n-gram of characters. GloVe is a hybrid approach exploiting both machine learning and statistic matrix operations. It applies a bilinear regression model with global matrix factorisation and local context window techniques for capturing the co-occurrences of words across a whole corpus.

WordPiece [135] was first used in voice recognition problems. It applies a segmentation algorithm to initialise a vocabulary with individual characters which is then populated using the most frequent combinations of symbols. It can process out-of-vocabulary words by segmenting them into sub-words that may occur in vocabulary. WordPiece has two limitations: 1) tokenizing across distinct languages due to differences in whitespace practice, and 2) detokenization cannot be executed in some instances because of splits in punctuation and white spaces. Sentencepiece [76] is another segmentation algorithm which addresses the limitations of WordPiece. It processes input sentences as a stream of Unicode characters (including white spaces) and a unigram language model or **BPE (Byte Pair Encoding)** to build vocabulary.

1.2.2 Cross-lingual Word Embeddings. Cross-lingual word embeddings are based on parallel [6, 48, 56, 73, 90, 99**?**] or comparable data [40, 147]. VecMap [7] aligned embedding vectors for different languages into a unified embedding vector space. It requires a large corpus of parallel embeddings for both source and target languages. Dual mono-lingual embeddings with bilingual dictionaries [43] are used to perform canonical correlation analysis to produce a shared vector space. Word2vec's skip-gram model [90] captures the context of words in source and target languages requiring parallel sentence corpus with alignment information. Instead of a parallel corpus, a comparable corpus [147] is applied to the skip-gram model without the use of alignment information. The normalisation of word embeddings on a hypersphere and constraining the linear

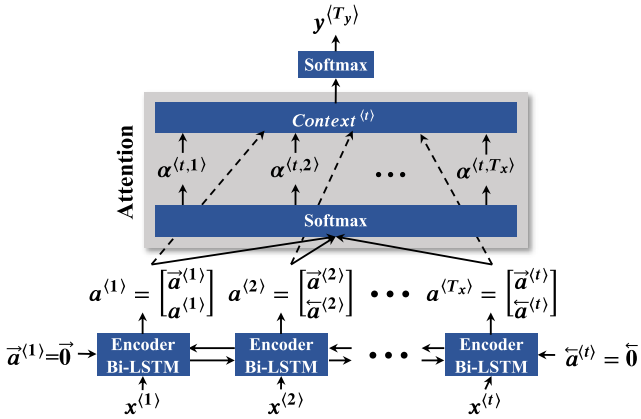

Fig. 2. Attention mechanism with bi-directional LSTM encoder for text classification.

transform as an orthogonal transform was applied to improve VecMap [163]. Additional refinement was made to this approach using mean centering [7]. MUSE [34], an unsupervised method, exploits adversarial training to build a bilingual dictionary and eliminates the need for parallel corpora. Barista [48] is a technique for training bilingual embeddings using pseudo-bilingual corpus generated with a small task-specific bilingual lexicon.

2 DEEP LEARNING ARCHITECTURES IN SENTIMENT ANALYSIS

Neural Networks have undergone a substantial transformation from the multi-layered perceptron [50] and provide a diverse selection of architectures to model sentiment analysis tasks. In this section, the prominent deep learning architectures utilised in sentiment analysis and NLP are introduced. For each architecture, a description of their inner work is provided for a better understanding of the studies surveyed in the later sections of this paper. A detailed examination of all the different architecture variants is outside the scope of this study, interested readers are encouraged to refer to the respective studies for further details.

2.1 Attention Mechanism

Attention mechanism is one of the most widely applied architectures and first grew to prominence with machine translation tasks [10]. It addressed a key limitation of traditional encoder-decoder architecture where the decoder learns representations that are not likely to be relevant to the end task. Attention determines the significance of each information it is exposed to and provides a weighted sum of all the features. Content-based [49], additive [10], location-based [91], general [90], dot-product [91], and scaled dot-product [146] are well known types of Attention. Each Attention mechanism has an alignment score function. Examples of architectures utilising these mechanisms include content-based in Neural Turing Machines [49], additive Attention in Fastformer [159] architecture, and scaled dot-product in multi-headed self-attention which forms the basis for Transformers [146]. Location-based attention mechanism is commonly used for text classification in sentiment analysis tasks. It calculates the context vector as a weighted average of all the provided encoder hidden vectors.

$$\alpha^{(t,t')} = softmax(a^{<t>}) \tag{1}$$

$$context^{<t>} = \sum_{t'=1}^{T_x} \alpha^{<t,t'>} a^{<t'>} \tag{2}$$

Table 1. Survey of Transformer Models

Model	Date	Pre-training Data	Pre-training	No. Parameters
GPT [124]	Jun-18	BooksCorpus (800m words)	Generative pre-training	117m
BERT [37]	Oct-18	BooksCorpus (800m words) and English Wikipedia (2.5B words)	Masked Language Modelling, Next Sentence Prediction	Base = 110m, large = 340m
ERNIE [181]	May-19	English Wikipedia (2.5B words)	Denoising Entity Auto-Encoder	114m
XLNet [169]	May-19	Common crawl (110B), ClueWeb (19B), Giga5 (16B), English Wikipedia (2.5B words)	Permutation language modelling	Base = 117m, large = 360m
RoBERTa [89]	Jul-19	BooksCorpus (800m words) and English Wikipedia (2.5B words), CC News, OpenWebText, Stories (33B)	Masked Language Modelling, Next Sentence Prediction	365m
ALBERT [79]	Sep-19	BooksCorpus (800m words) and English Wikipedia (2.5B words)	Masked Language Modelling, Next Sentence Prediction	Base = 12m, large = 18m, X large = 60m
BART [81]	Oct-19	BooksCorpus (800m words) and English Wikipedia (2.5B words), CC News, OpenWebText, Stories (33B)	Denoising Language Modelling: Token Masking, Token Deletion, Sentence Permutation and Document Rotation	Base = 120m, large = 374m
T5 [125]	Oct-19	Colossal Clean Crawled Corpus (750GB)	Sequence to Sequence Denoising: Corrupting Span Objective	Small = 60m, Base = 220m, Large = 770m, 3B, 11B
ELECTRA [32]	Mar-20	BooksCorpus (800m words) and English Wikipedia (2.5B words)	Masked Language Modelling, Next Sentence Prediction	Base = 110m, Large = 330m
DeBERTa [52]	Jun-2020	Wiki+Book 16GB, OpenWebText 38GB, Stories 31G, BCCNews 76GB	Masked Language Modelling, Virtual Adversarial Training	Base = 110m, large = 340m, 1.5B
PaLM [31]	Apr-2020	filtered webpages, books, Wikipedia, news articles, source code, and social media conversation (780B tokens)	Masked Language Modelling or Corrupting Span Objective	540B

A common implementation of location-based attention mechanism is on top of bi-directional LSTM's encoder (depicted in Figure 2). A sequence of words $s = [x^{<1>}, x^{<2>}, \ldots, x^{<t>}]$ is fed to bi-directional LSTM encoder to produce a single hidden feature vector $a^{<T_x>} = [\overrightarrow{a}^{<t>}; \overleftarrow{a}^{<t>}]$. At time step t, all the hidden states from encoder $[a^{<1>}, a^{<2>}, \ldots, a^{<T_x>}]$ are fed to the attention layer. Attention calculates the attention weights $[\alpha^{<t,1>}, \alpha^{<t,2>}, \ldots, \alpha^{<t,T_x>}]$ using Softmax 1 function. Context vector 2 is calculated using Attention weights $\alpha^{<t,t'>}$ for each encoder hidden state $a^{<t'>}$. Attention weight parameters are tuned during backpropagation.

2.1.1 Transformer Language Models. The Transformer language model is an attention-based architecture applying multi-headed self-attention. Transformers [146] were originally introduced for language modelling and have become the most widely used deep learning architectures in the NLP domain. These models undergo pre-training before being fine-tuned for the downstream task. In the pre-training stage, language models are pre-trained on a large domain-general dataset with unsupervised training (refer to Table 1). This enables language models with millions or billions of parameters to model natural language on an unprecedented scale and granularity. The

pre-trained models are then fine-tuned for downstream tasks using domain-specific data. Unsupervised training regimes can include generative pre-training [124], predicting randomly masked words in masked language modelling [32, 37, 79, 89], next sentence prediction [32, 37, 79, 89], corrupting span objective [125], document rotation [81], and permutation language modelling [169].

Transformer language models consist of stacked encoders/decoders layers (refer to Figure 7). The original Transformer [146] encoders contained a position-wise feed-forward network and multi-headed self-attention. Each encoder applies a residual connection followed by a normalisation layer. The decoder layers consist of a masked multi-headed self-attention layer. Receiving the output of the masked multi-headed self-attention encoder layer as query Q combined with value V and key K. Each sublayer in the decoder has a residual connection around it followed by a normalization layer.

Self-Attention pays attention to feature representations using weight proportional to a similarity score between representation pairs. The sequence of inputs with n tokens, d dimensions and b batch, $X \in \mathbb{R}^{n \times d \times b}$, gets projected with matrices $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$ and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ to compute representations key, value and query K, V and Q. Self-Attention is computed using scaled dot Attention. The output for each of the Attention heads are obtained using Attention operation (4). Multi-headed self-Attention is obtained by concatenating multiple self-Attention mechanisms (3).

$$Multihead(Q, K, V) = Concact_i(head_1, head_2, \ldots, head_h)W^O \tag{3}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{4}$$

$$Attention(Q, K, V) = softmax\left(QK^T/\sqrt{d_k}\right)V \tag{5}$$

Multi-headed self-Attention utilises multiple scaled dot-product 5 in parallel and enables the model to capture relevant information in different representative subspaces while also reducing complexity from $O(d^2 \cdot n)$ per layer (recurrent approaches) to $O(n^2 \cdot d)$, where d is the representation dimension and n is the sequence length. Another key benefit of multi-head self-Attention over RNNs is the ability to capture more distant dependencies. Shorter paths make it easier to learn these long-range dependencies. Multi-headed self-Attention maximum path length requires $O(1)$ sequential operations while recurrent approach requires $O(n)$.

2.2 Recurrent Neural Networks

Engineered for time-series tasks, **Recurrent Neural Networks (RNN)** became the dominant approach for tasks like natural language understanding, machine translation, speech recognition, video recognition, and DNA sequence analysis before the introduction of Transformers. RNN architecture maintains a state and iterates a series of input data while learning from what it has seen from past timesteps. Each input element in the time series is fed to units in RNN and hidden states produced from each RNN unit are also fed to the proceeding RNN unit.

2.2.1 Long Short-Term Memory (LSTM). When exposed to longer sequences of inputs, RNNs are vulnerable to vanishing and exploding gradients [18]. This occurs when gradients less than or more than 1, are multiplied during backpropagation, causing a convergence towards either 0 or infinity. This overarching drawback to RNN architecture created a demand for the LSTM network [58]. LSTMs have a hidden state but with an additional cell state designed to manipulate its memory in a more deliberate approach with a unique gating mechanism.

$$\Gamma_f^{<t>} = \sigma\left(W_f\left[h^{<t-1>}, x^{<t>}\right] + b_f\right) \tag{6}$$

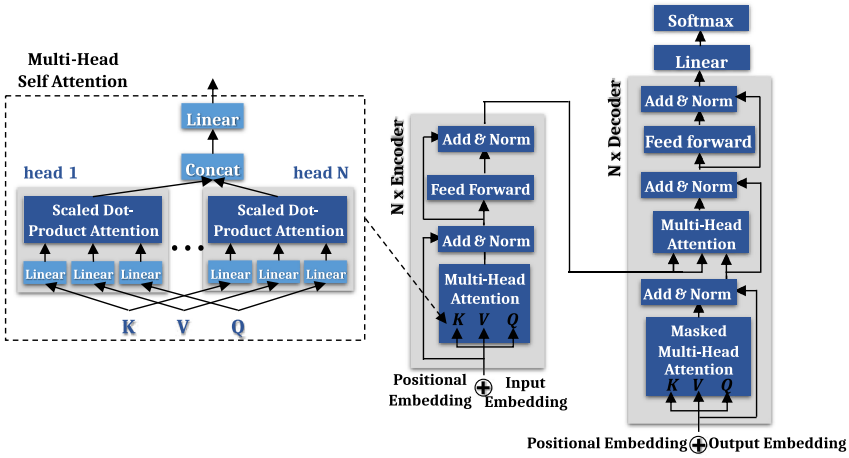

Fig. 3. Transformer architecture with Multi-Headed Self-Attention [146].

$$\Gamma_u^{<t>} = \sigma\left(W_u\left[h^{<t-1>}, x^{<t>}\right] + b_u\right) \tag{7}$$

$$\Gamma_o^{<t>} = \sigma\left(W_o\left[h^{<t-1>}, x^{<t>}\right] + b_o\right) \tag{8}$$

$$\widetilde{c}^{<t>} = tanh\left(W_c\left[h^{<t-1>}, x^{<t>}\right] + b_c\right) \tag{9}$$

$$c^{<t>} = c^{<t-1>}\Gamma_f^{<t>} + \widetilde{c}^{<t>} \cdot \Gamma_u^{<t>} \tag{10}$$

$$h^{<t>} = \Gamma_o^{<t>} \cdot tanh(c^{<t>}) \tag{11}$$

For a sequence of words $s = [x^{<1>}, x^{<2>}, \ldots, x^{<i>}]$, there are h hidden units for each time step t, the input is $X \in \mathbb{R}^{n \times d}$, where n is the number of samples and d is the number of hidden units. The previous hidden state is $h^{<t-1>}\mathbb{R}^{n \times h}$. LSTMs have three gate mechanism interacting in a unique way (refer to Figure 4). Forget gate $\Gamma_f^{<t>} \in \mathbb{R}^{n \times h}$, update gate $\Gamma_u^{<t>} \in \mathbb{R}^{n \times h}$ and output gate $\Gamma_o^{<t>} \in \mathbb{R}^{n \times h}$ are calculated with sigmoid layers (6)–(8) using input $x^{<t>}$, previous hidden state h^{t-1} and bias terms $b_f, b_u, b_o, b_c \in \mathbb{R}^{1 \times h}$. LSTM cell output is the cell state (11) calculated using forget gate $\Gamma_f^{<t>}$, update gate $\Gamma_u^{<t>}$, previous cell state $c^{<t-1>}$ and candidate cell state $\widetilde{c}^{<t>}$. $W_f, W_u, W_o, W_c \in \mathbb{R}^{d \times h}$ are the cell weight parameters.

2.2.2 Gated Recurrent Unit (GRU). A variant of LSTM, GRU uses only two gates and does not maintain a cell state $c^{<t>}$ different from the hidden state $h^{<t>}$. A simple implementation of the GRU classifier will involve a sequence of words $s = [x^{<1>}, x^{<2>}, \ldots, x^{<i>}]$ being fed to a GRU [30] alongside a Softmax classifier for text classification (refer to Figure 5). Reset gate $\Gamma_f^{<r>} \in \mathbb{R}^{n \times h}$ and Update gate $\Gamma_u^{<r>} \in \mathbb{R}^{n \times h}$ are computed using sigmoid layers (12) and (13) using input $x^{<t>}$, previous hidden state $h^{<t-1>}$ and bias terms $b_r, b_u \in \mathbb{R}^{1 \times h}$. $W_r, W_u, W_h \in \mathbb{R}^{d \times h}$ are the cell weight parameters.

$$\Gamma_r^{<t>} = \sigma\left(W_r\left[h^{<t-1>}, x^{<t>}\right] + b_r\right) \tag{12}$$

$$\Gamma_u^{<t>} = \sigma\left(W_u\left[h^{<t-1>}, x^{<t>}\right] + b_u\right) \tag{13}$$

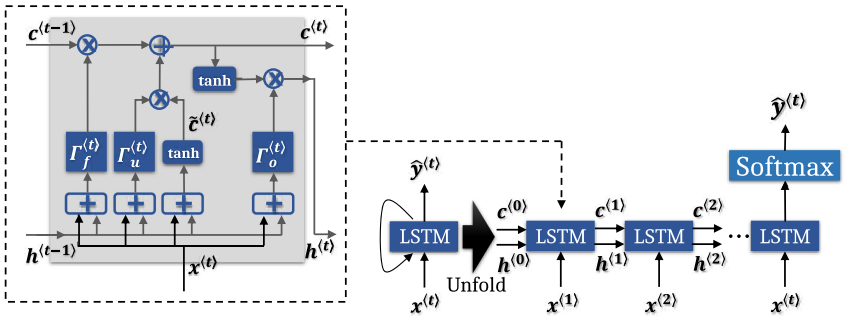

Fig. 4. Long-Short-Term Memory cell architecture for text classification [58].

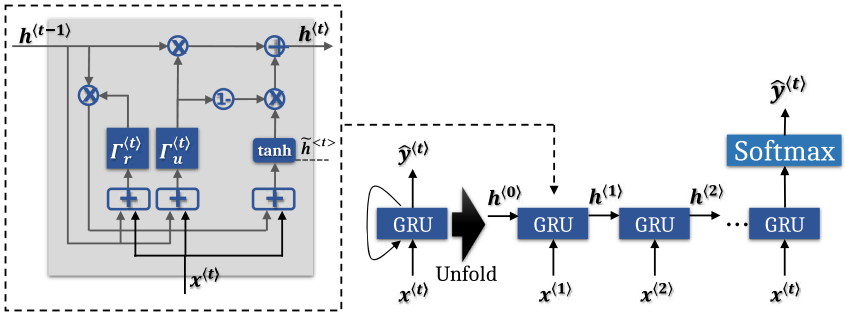

Fig. 5. Gated Recurrent Unit architecture for text classification [30].

$$\widetilde{h}^{<t>} = tanh\left(W_h\left[\Gamma_r^{<t><t>}, x^{<t>}\right] + b_h\right) \tag{14}$$

$$h^{<t>} = \left(1 - \Gamma_u^{<t>}\right) \cdot \widetilde{h}^{<t>} + \Gamma_u^{<t>} \cdot h^{<t-1>} \tag{15}$$

The update gate learns long-term dependencies while the reset gate learns short-term dependencies. A simpler gating mechanism leads to fewer parameters which make GRUs better candidates for deeper architectures while also requiring fewer samples to generalise.

2.2.3 Convolutional Neural Networks (CNN). CNN significantly outperforms traditional ANNs [80] and has demonstrated potential for text classification tasks [72]. Traditional CNN is composed of representational input, feature extraction, fully connected, and output layers. Feature extraction consists of convolutional and pooling layers. They apply filters to feature maps obtained from previous layers. Filters are learnable weights and are optimised during backpropagation.

For sentiment classification (refer to Figure 6), a sequence of words $s = [x^{<1>}, x^{<2>}, \ldots, x^{<i>}]$ where each text input sample s is represented by an embedding matrix $S \in \mathbb{R}^{n \times d}$, where n is number of words and d is embedding dimension. Embedding matrix of size $n \times d$ is fed to convolutional layer. A convolutional filter $k \in \mathbb{R}^{h \times d}$, h being the number of words is slid across the embedding matrix. Element-wise product operation (16) is applied for all the elements in the matrix. The bias term $b \in \mathbb{R}$ is added to the sum of element-wise product operation and a non-linear activation (17) function is applied. Representations produced from feature extraction layers undergo flattening and are fully connected before being fed to a Softmax classifier.

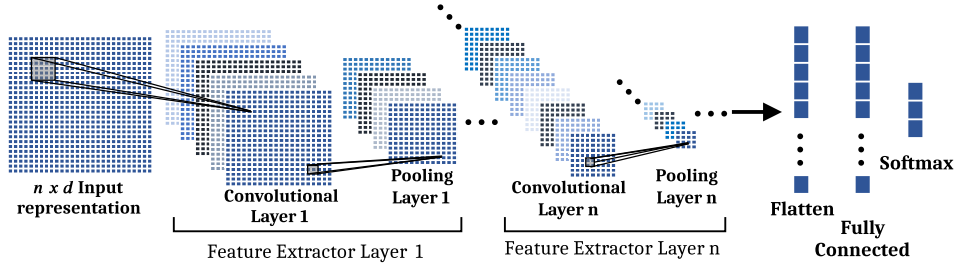

Fig. 6. Text classification flow with Convolutional Neural Network.

$$g = \left(\sum_{i=1}^{m} w \cdot x \right) + b \tag{16}$$

$$z = f(g)b \tag{17}$$

3 SURVEY METHODOLOGY

This section provides an overview of the survey methodology, covering the search methodology, inclusion criteria, and taxonomy of sentiment analysis. The taxonomy categorises and classifies the surveyed sentiment analysis.

3.1 Search Methodology

In compiling this survey, digital libraries and digital search engines were utilised; these included libraries such as ACM digital, IEEE Xplorer, Scopus ScienceDirect, ResearchGate, Hindawi, and arXiv and digital search engines like Google Scholar, Google, and Microsoft Academic. To survey benchmark datasets and top-performing architectures, published results were cross-examined with live benchmark dataset trackers keeping live leaderboards of the different architectures. These included www.paperswithcode.com, www.nlpprogress.com, and www.gluebenchmark.com/leaderboard.

3.2 Inclusion Criteria

A set of inclusion rules were established when considering studies to be part of the survey: (1) studies published between 1st July 2017 and 1st July 2022, (2) studies applying deep learning architectures on tasks involving sentiment analysis, (3) studies that are highly cited, innovative in their approach or performed well on benchmark datasets, and (4) model architectures must also be presented in a published format to provide documentation for technical review. All the architectures surveyed have met these aforementioned rules.

3.3 Taxonomy of Sentiment Analysis

Sentiment analysis tasks are classified into coarse-grain, fine-grain, cross-domain and cross-lingual categories (refer to Figure 7). **Coarse-grain sentiment analysis** tasks may include a single sentence or a very large document. These tasks focus on document and sentence-level and document-level granularity and do not go below this granularity. Document-level sentiment analysis is traditionally used to describe the task of sentiment polarity classification on a document. However, it is not limited to this and can include a whole host of other tasks such as sarcasm analysis, offensive language analysis, hate speech analysis, and other text classification tasks.

Fine-grain sentiment analysis covers polarity classification and is performed with aspect and targeted sentiment analysis. Aspect sentiment analysis involves one or more aspects (or features)

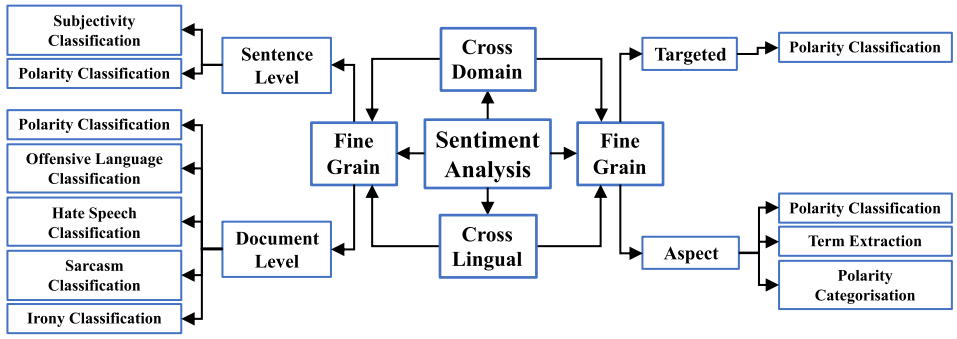

Fig. 7. Taxonomy of sentiment analysis.

in the text while targeted sentiment analysis addresses the analysis of different target entities. Aspect sentiment analysis is also applied for aspect term extraction and aspect polarity categorisation tasks. A detailed explanation of the aspect and target is covered in Section 4.

Cross-domain covers sentiment analysis across distinct domains whereas cross-lingual sentiment analysis involves sentiment analysis across different languages. Cross-domain and cross-lingual categories also cover tasks that fall into coarse and fine-grain categories.

4 FINE-GRAIN SENTIMENT ANALYSIS

Fine-grain sentiment analysis encompasses aspect-level and targeted analysis with a focus on analysing one or more aspects and/or target entities in a sentence. An aspect defines a specific feature of a piece of text. An example of an aspect could be price, software or hardware in a laptop review [116]. Target is an entity and can define a service, product, topic, issue, event, or organisation that has sentiment expressed towards it [85]. Aspect and targeted polarity classification involve determining the sentiment polarity of a sentence of words towards a pre-defined set of aspects or target entities. Aspect term extraction is the task of detecting all subjective and/or non-subjective aspect terms in a given sentence.

Fine-grain analysis focused architectures lag behind coarse-grain architectures on both performance and research activity [104]. A model is required to capture syntax and semantic information on a finer scale and often involves shorter text. This requires modelling semantic relations involving aspect, target, and context words. Fine-grain sentiment analysis remains an active research area in NLP and is considered more challenging than coarse-grain tasks to due the difficulty in agreeing on sample sentiment [45].

The surveyed studies are grouped by their task focus with accompanying details of deep learning architecture applied, published date, input word embeddings, and the main contributions (summarised in Table 2). An analysis of recent trends in fine-grain sentiment analysis is provided in Section 4.1 followed by insights into the top-performing architectures on fine-grain benchmark datasets in Section 4.2.

4.1 Trends in Fine-Grain Sentiment Analysis

The majority of the surveyed studies address aspect-level tasks (aspect term extraction, aspect term categorisation, and aspect term polarity classification, aspect term categorisation, and aspect term polarity classification). Targeted sentiment analysis was also surveyed with target polarity classification. A diverse selection of deep learning architectures is examined with pre-trained word embedding techniques as the default method for input representations. Word2vec [54, 93], GloVe [53, 61, 62, 82, 92, 152, 172, 177], and WordPiece [70, 128, 142, 161, 164, 166, 167] are commonly applied on fine-grain tasks. However, these traditional word embeddings techniques can train models

Table 2. Survey of Fine-Grain Sentiment Analysis Studies

Task	Architecture	Date	Embeddings	Contribution
Aspect Term Extraction	BERT Adversarial Training [70]	Jan-21	Wordpiece	Adversarial training on post-trained BERT. Applied on aspect term polarity classification and aspect term extraction. BERT Adversarial Training (BAT) improves post-trained BERT performance.
	Local Context Focus Aspect Term Extraction [167]	Jan-21	Wordpiece	Local context focus mechanism with multi-task learning model for Chinese-oriented term extraction and polarity classification. Integrating domain-adapted BERT model achieves SOTA on aspect term extraction and aspect polarity classification in SemEval-2014 task4.
Aspect Term Categorisation	Sentic LSTM [93]	Feb-18	Word2vec	Task-related embeddings of common-sense knowledge incorporated into the bi-directional LSTM encoder and hierarchical Attention. Aspect and sentence-level information is used as input.
Aspect Term Polarity Classification	Parameterized CNN [61]	Jul-19	GloVe	Novel CNN incorporates aspect information using parameterised filters and gates. Extraction of feature matrix using convolutional filter on a sentence CNN. Parameterised gates CNN capture aspect term features which are utilised as a gate in sentence CNN. Aspect-specific features concatenated with sentiment feature.
	Gated Alternate Neural Network [87]	May-18	GloVe	Novel gated architecture using gated truncation RNN to capture aspect-dependent information. This encodes information on aspect and context work distance. Convolution and pooling are employed to capture position invariant sentiment features.
	Hierarchical iterative Attention (HIA) [172]	Sep-17	GloVe	Modelling document-level multi-aspect sentiment analysis as a machine comprehension problem. Automatically attends to parts to output aspect ratings using pseudo-question-answer pairs via a small set of aspect-related keywords. Attention captures aspect-related features using recurrent interactions across documents and aspect questions.
	Recurrent Attention Memory [26]	Sep-17	GloVe & Word2vec	Bi-directional LSTM captures sentiment phrases. Position-weighted memory module captures the relevance of each word to the target. Recurrent Attention uses multiple-Attention to amplify information learned from the position-weight memory layer.
	Interactive Attention network [92]	Aug-17	GloVe	Models Attention for target and context words, generating separate representations for target and context. LSTM obtains hidden states from word embeddings for the target and its context words. Hidden states for target and context are fed to each other's Attention layer to generate Attention vectors.
	Attention-over-Attention Networks [62]	Apr-18	GloVe	Explicitly learns the relationship between context and aspect words for both aspect and context words. Aspect and context word embeddings are fed to two bi-directional LSTM networks to capture hidden semantics of words. Hidden vectors are then fed to Attention-over-Attention layer to capture relevance for each word.
	Attention-based LSTM [54]	Jun-18	GloVe & Word2vec	Transferring document-level knowledge onto aspect sentiment, Attention and LSTM utilised on aspect level while a standard LSTM classifier is used for document-level. LSTM produces hidden vectors. Aspect level Attention learns the significance of each word. Two methods to transfer knowledge: pretraining and multi-task learning.
	LSTM+SynATT [53]	Aug-18	GloVe	Attention learns the semantic meaning of the opinion target. Aspect embeddings are jointly trained with the Attention-based LSTM using an autoencoder network. A sentence's syntactic structure is encoded using Attention, obtaining syntactic information from a dependency parser.

(Continued)

Table 2. Continued

Task	Architecture	Date	Embeddings	Contribution
	BiAtt+GCN [182]	Jun-19	GloVe	The bi-directional LSTM layer captures high-level syntactic and semantic information which are concatenated with position encoding. Features fed to bi-directional Attention layer to capture aspect-specific information between aspects and context words. **Graph convolutional network (GCN)** captures interdependency and relationships between entities.
	Hierarchical Attention [82]	Oct-18	GloVe	Position-aware features are obtained using position embeddings. This generates target-specific representations for contextual words. Bi-GRU-based encoder for sentences captures an abstract representation of sentences. Attention captures aspect and context-related information. Features from Attention components are fused.
	BERT-ADA [128]	Aug-19	Wordpiece	Self-supervised domain-specific language models fine-tuning of BERT and task-specific fine-tuning. Laptop and restaurant datasets are used for fine-tuning while BERT language model (next sentence prediction) fine-tuning stage uses Amazon laptop reviews and Yelp restaurant reviews. BERT base uncased and XLNet base utilised.
	BERT-PT [164]	Apr-19	Wordpiece	Post-training on BERT for converting customer reviews into a knowledge base to answer user questions. BERT alongside joint post-training to improve domain and task knowledge. The joint post-training method uses pre-trained BERT weights for basic language understanding while modifying BERT with task and domain knowledge before the model is fine-tuned with downstream task labelled data.
	Local Sentiment Aggregating + DeBERTa large [166]	Oct-21	Wordpiece	Local sentiment aggregating mechanism is used to learn sentiment dependency which is more efficient than existing dependency tree-based models. BERT encoder learns embedding-based local context focus features of aspects. Differential weighting to gauge the significance of sentiment information for aspects.
Targeted Polarity Classification	Segmentation Attention-based LSTM [152]	Apr-18	GloVe	Segmented Attention-based LSTM with a linear-chain conditional random field learns the semantic dependencies between target and context words in a sentence. Bi-directional LSTM is used to capture contextual information and the segmented Attention layer is used to model the semantic dependencies between the target and the context words.
	Attentional Encoder Network [141]	Apr-19	GloVe & BERT	Context and target embeddings fed to Attentional encoder layer consisting of (1) intra-multi-headed self-Attention to obtain introspective context features, (2) inter multi-headed self-Attention to obtain context-sensitive target features. Both are followed by a point-wise convolutional transformation sublayer. A fully connected layer is used to project this final representation onto targeted classes.
	BERT-pair-QA-B, QA-M [142]	Mar-19	Wordpiece	Auxiliary sentence from an aspect in a sentence for QA-M and a sentence for QA-B sentence pair classification tasks. QA-M involves constructing a pseudo-question for the target aspect pair. With QA-B label information is appended while also temporarily converting the task into a binary classification (yes or no).
	Gated Conv& Aspect Embeddings [165]	Jul-18	GloVe	CNN trained on word embeddings, resulting features and aspect embeddings fed to gated Tanh-ReLU Units. The gating mechanism selectively outputs sentiment features. Max pooling follows the gated Tanh-ReLU unit layer.
	Quasi-Attention Context-Guided BERT [161]	Oct-20	Wordpiece	Adding context to self-Attention. BERT made context-aware to create a model which utilises context-guided Softmax-Attention. This is improved using quasi-Attention learning compositional Attention that facilitates subtractive Attention.

with only a few hidden layers [22, 100, 112], producing static context-independent features with low-level syntactic and semantic information, unable to capture higher-level features.

The challenge of training word embeddings which capture higher-level semantic and syntactic information is addressed with language models that produce contextualised word-embedding. These architectures capture the semantics of words in distinct contexts with deeper architectures having subsequent layers capture higher-level features from low-level word embeddings [54, 61, 61, 83, 92, 93, 165, 172, 182]. Recent breakthroughs in language models address this challenge and produce deeper architectures (LSTM-based ELMo [113] and ULMFiT [60]) and even larger bi-directional contextualised Transformer models (BERT [37], OpenAI-Transformer [124] and XLNet [169]). These larger models can capture low-level and high-level features, providing a better foundation to model natural language tasks.

Most of the surveyed studies do not incorporate components modelling aspect or target information and instead rely on deeper architectures to model aspect and/or target information in a sentence [54, 62, 70, 92, 128, 144, 152, 164, 172]. Aspect embeddings [61, 93, 141, 165] are the most common approach to directly incorporate aspect or target information followed by other approaches which use LSTM encoders [183] to learn representations from aspect or target term, parametrised filters/gates in CNNs [61], and the joint training of aspect embeddings using autoencoder [54]. Other architectures also utilise position-aware models to incorporate information on the relative position of aspect and context words using position weights in Attention-LSTM models [26, 182] and position embedding in Attention [82].

Attention is used to model hierarchical dependencies, modelling interactions between documents and aspect questions [84] and capture aspect and context-related information [82]. Attention layers enable a model to determine the importance of each input towards the outcome of a task by its semantic association. Bi-directional encoders enable a model to capture forward and backward information from a sequence of words and modelling dependencies in both directions. Bi-directionality is commonly used with LSTM [86] and Attention-LSTM architectures [26, 54, 62, 92, 152]. The use of Attention on top LSTM [26, 53, 54, 61, 92, 93, 152, 182] and GRU encoders [82] was a common architecture theme before the arrival of Transformer-based language models on both activity and performance [110].

Current trends in fine-grain tasks are to robustly model syntax and semantics of natural language using large pre-trained Transformer language models and provide a powerful approach for knowledge transfer [70, 128, 142, 164]. These architectures utilise pre-trained language models on domain-general data and are then fine-tuned using domain-specific data for aspect or targeted sentiment analysis. Common techniques for task-specific fine-tuning include the use of auxiliary sentences to convert aspect-level task [142], transforming customer reviews into a source of knowledge for answering user questions [165], and single sentence classification to predict the label with output sentence [128, 142, 161, 166].

4.2 Survey of Fine-Grain Datasets

Fine-grain benchmark datasets are scarce and this survey found only SemEval 2014 [116] Task 4 subtask 2 and Sentihood [132] to be widely used for benchmarking. The top-performing architectures ranked by their performance are presented in **3**. SemEval 2014 Task 4 subtask 2 contains two datasets with 3,841 single sentence reviews for laptops and 3,845 single sentence reviews for restaurants. This is an aspect term and aspect polarity classification task on a three-point scale (positive, negative and neutral).

Sentihood dataset is used for targeted-polarity classification and contains sentences with questions and answers including one or two location entities. This dataset is used to predict the sentiment label for a target aspect pair and contains a total of 5,215 sentences with 3,862 sentences for

Table 3. Fine-Grain Benchmark Datasets with Top Performing Architectures

Task	Dataset	Rank	Architecture	Performance
Aspect Term Polarity Classification	SemEval 2014 Task 4 Subtask 2	1	Local Sentiment Aggregating + DeBERTa large [166]	ACC: 0.886
		2	Local Context Focus Aspect Term Extraction [167]	ACC: 0.862
		3	BERT-ADA [128]	ACC: 0.841
Targeted Polarity Classification	Sentihood	1	QACG-BERT [161]	ACC: 0.938
		2	BERT-pair-QA-M & QA-B. [142]	ACC: 0.93.8
		3	Recurrent Entity with Delayed Memory Update [154]	ACC: 0.910

one target and 1,353 sentences for multiple targets. Each target is classified on a three-point scale polarity (positive, negative and neutral) in a sentence.

Both datasets are dominated by recent architectures using pre-trained Transformer language models. Before the arrival of Transformer-based language models, Attention-LSTM [182] and Gated-GRU [86] based approaches dominated both datasets. These have been pushed down in the rankings with the exception of a gated GRU model in 3rd rank on Sentihood. SemEval 2014 Task 4 subtask 2 1st, 2nd and 3rd and Sentihood 1st and 2nd rank are BERT-based models that utilised BERT [37] Transformer as a foundation for architectures.

5 COARSE-GRAIN SENTIMENT ANALYSIS

Sentiment analysis tasks conducted on a document or sentence-level are categorised as coarse-grain. A document is a body of text containing one or more sentences. Examples of documents can include product reviews [94, 98, 108, 171] or social media posts [55, 175]. Document-level sentiment analysis has become synonymous with sentiment polarity classification (positive, negative and/or neutral) of a document but can also include other tasks such as sarcasm analysis, offensive language or hate speech analysis. Sentence-level sentiment analysis involves tasks on a single sentence including sentence-level polarity classification and subjectivity classification (determining whether a sentence is subjective or not). The majority of the coarse-grain studies are focused on document and sentence polarity classification on a binary (positive or negative) or three-point scale (positive, negative and/or neutral). The challenge of document-level sentiment analysis is to capture the syntactic and semantic relations between sentences which determine the overall semantic polarity of a document of text. Sentence-level polarity classification requires modelling the semantic relations of a sentence of words. The remainder of this section covers the survey of relevant studies for the following coarse-train tasks: document-level polarity classification, sentence-level polarity classification, and offensive language and hate language detection. The surveyed studies are grouped by their task focus with the accompanying details of deep learning architecture applied, published date, input word embeddings, and the main contributions (summarised in Table 4). An analysis of recent trends in fine-grain sentiment analysis is provided in Section 5.1 followed by insights into the top-performing architectures on coarse-grain benchmark datasets.

5.1 Trends in Coarse-Grain Sentiment Analysis

Most of the surveyed studies in document-level coarse-grain sentiment analysis are targeted at document-level polarity classification, offensive language detection, hate speech detection, sarcasm detection, and irony detection. Sentence-level tasks, analysing text on a single sentence,

Table 4. Survey of Coarse-Grain Sentiment Analysis Studies

Task	Architecture	Embeddings	Date	Contribution
Sentence-Level Polarity Classification	Attention Capsule model [155]	GloVe	Apr-18	Capsules model sentiment category. LSTM produces hidden feature vectors, for embeddings, which are then passed to capsules (containing Attention). Sigmoid function predicts active state probability and a module that reconstructs representations by multiplying the active state probability by the capsule representation.
Sentence-level Subjectivity Classification	Recurrent Convolutional Neural Network [180]	Word2vec	Dec-18	A convolution neural network learns features for classification and a recurrent neural network models semantic information of the text. Pre-trained word embeddings are used to reduce the dimensions of word vectors and avoid sparsity issues. Bi-directional LSTM is used for recurrent layers to capture forward and backward information.
	Multilayer Perceptron [16]	Word2vec	Sep-20	Word embeddings are used to transfer semantic information to give representations of subjective terms not present in training samples. Word representations are concatenated to create a final vector for a sentence that provides semantic information. The network contains input, hidden, and output layers.
	Multitask Learning & BERT Embedding [134]	BERT & GloVe	Jan-22	Multitasking learning on polarity and subjectivity classification. Pre-trained BERT and GloVe embeddings for subjectivity and polarity dataset fed to bi-directional LSTM layer which extracts hidden features from embeddings, and this is passed to self-Attention. Softmax classifiers are used for different tasks.
Document-level Polarity Classification	Tensor Fusion Network [174]	GloVe	Jul-17	Layers are used to explicitly learn unimodal, bimodal and trimodal dynamics. Layers: (1) embedding subnetworks model feature input and output embeddings, (2) tensor fusion layer captures unimodal, bimodal and trimodal relations, and (3) inference subnetworks predict sentiment from the output of the 2nd layer.
	MSA & TSA model [14]	GloVe	Aug-17	LSTM-Attention is used for message-level and topic-based analysis. The model consists of an input word embedding layer, a bi-directional LSTM layer producing high-level features. The Attention layer captures the contributions of words to document-level polarity.
	ULMFiT [60]	Embedding layer	Jan-18	AWD-LSTM is used to pre-train language models on general-domain corpus and fine-tuned for document-level analysis of differing label types and document size. AWD-LSTM is modified by adding two linear blocks.
	BERT-large [37]	Wordpiece	Jun-19	Language model pre-trains bi-directional representations from the unlabelled text. Composed of multiple layers of Transformer encoders/decoders and is closely based on the original Transformer architecture [146]. The bi-directional self-Attention can attend to tokens on both left and right.
	T5-3B & T5-11B [128]	Wordpiece	Oct-19	Transforms NLP tasks into text-to-text (fed text to generate target text) format. Based on the original Transformer [146]. T5-3B and T5-11B model utilises approximately 3 and 11 billion parameters, respectively. Models are pre-trained on the "colossal Clean Crawled Corpus".

(Continued)

Table 4. Continued

Task	Architecture	Embeddings	Date	Contribution
	XLNet [169]	Wordpiece	Dec-19	Designed to overcome the limitations of BERT. XLNet attempts to integrate bi-directional context while also circumventing mask tokens in pre-training and independence prediction of these tokens. Tokens are predicted using the preceding context but in random order.
	ALBERT [79]	Wordpiece	Feb-20	Using two parameter-reduction techniques on BERT: (1) factorized embedding parameterisation for efficient learning of context-dependent representations, and (2) cross-layer parameter sharing all parameters across all layers leading to best parameter efficiency. Significantly smaller architectures with state-of-the-art performance.
	SMART-RoBERTa Large [69]	Wordpiece	Jul-20	Framework for efficient fine-tuning of pre-trained language models for better generalisation on downstream tasks. Smoothness-inducing regularisation for managing model complexity and Bregman proximal point optimisation to prevent aggressive updating.
	ERNIE 3.0 [143]	Wordpiece	Jul-21	Framework for pre-training large knowledge enhanced language models which fuse auto-regressive network and auto-encoding network to excel in both natural language understanding and text generation tasks in zero-shot learning. Uses Transformer-XL for the backbone.
	METRO-LMXXL [11]	Wordpiece	Apr-22	Large autoencoding language models pre-training through efficient denoising. Parameter efficiency through denoising model generated signals. Uses an auxiliary model to generate signals dynamically to provide efficient pretraining. Other techniques include ZeRO optimizer [126], a set of customised Fused operations in mix-precision training and scaled initialization techniques.
Offensive Language Detection	LSTM Classifiers [114]	Embedding layer	Jan-18	Ensemble RNN classifiers fed tweets with behavioural features. Features crafted using neutral, racist, and sexist tendencies. Mechanisms for accumulating classifications voting and confidence. Majority voting is used when 2+ of the root classifiers agree. When all classifiers conflict, preference is given to the classifier providing the strongest prediction confidence.
	BERT [37]	Wordpiece	Jun-19	Pre-trained BERT language model to transfer knowledge to downstream tasks. Additional pre-processing before BERT tokenizer is available including substituting emoji using open source and hashtag segmentation using open-source Github projects.
	Real-Time ALBERT [1]	Wordpiece	Mar-21	Knowledge transfer using distilled Transformer-based architecture on user-generated data alongside a text normalisation pipeline. Designed to address the challenge of utilising deep learning models on a large volume of Twitter data. This challenge is known as the Achilles heel of deep learning on user-generated data.
Hate Speech Detection	Fermi [64]	Wordpiece	Jun-19	Pretrained Transformer based universal encoder on sentences, phrases or short paragraphs using a variety of data tasks and sources with aim of generalising NLP tasks. Encoder sentence embeddings (512-dimensional vector) are used with SVM and RBF kernel to classify hateful and language detection.

(Continued)

Table 4. Continued

Task	Architecture	Embeddings	Date	Contribution
	Atalaya [121]	Bag-of-words, bag-of-characters	Jun-19	Bi-directional LSTM and dense layer taking as input context-dependent ELMo embeddings. Another approach used is bag-of-words input alongside ELMO embeddings. Bag-of-words encoding to build tweet representations for each token in the training text. The model focuses on embeddings to make performance gains.
	Augment to Prevent [129]	Word2Vec, GloVe, FastText	Nov-19	Tackling hate speech overfitting problem using data augmentation to reduce class imbalance. Augmentation techniques include synonym swapping, word token wrapping with sequence padding, and recurrent language generation to generate correct classes. The proposed framework is tested prominent neural network architectures.
	DeepHate [24]	GloVe, Word2vec	Jul-20	Sentiment and word embeddings are fed to neural network layer to learn sentiment, semantic and topic representations. The network consists of a CNN-LSTM-Attention encoder alongside Dirichlet Allocation model. Representations are fused using the gate Attention network before a Softmax classifier.
Sarcasm Detection	ContextuAl Sarcasm Detector (CASCADE) [51]	Word2vec	Aug-18	Sarcasm is modelled using personality, stylometric, and discourse information from discussion forums. CNN is used to produce feature representations for text, capturing semantic and syntactic features. Contextual features are captured from user and discourse information. These are concatenated and fed to classification layers.
	Soft Attention-Bi-LSTM [77]	GloVe	Feb-19	A hybrid architecture using soft Attention and bi-directional LSTM with a CNN. Input word embeddings are used to transfer semantic information. Attention-Bi-LSTM is used to create a feature vector for the convolutional network and classifier. Punctuation based auxiliary information is incorporated into the convolutional network.
	Recurrent CNN RoBERTa [117]	WordPiece	Jun-2020	Pre-trained RoBERTa produces hidden features which are fed to the bi-directional LSTM layer. The pooling layer incorporates RoBERTa and bi-directional LSTM features before being fully connected and fed to the Softmax layer. RoBERTa is used to map words onto rich embedding space and bi-directional LSTM captures dependencies of RoBERTa output to identify sarcasm and irony features.
Irony Detection	Deep contextualized word representations [63]	ELMo	Oct-18	Character level word embeddings based on ELMo language models are used to capture multifaceted morpho-syntactic information which is used as an indicator for irony or even sarcasm. The output of ELMo model is fed to the bi-directional LSTM layer. This outputs to max-pooling and feed-forward network for binary classification.
	Sentiment based Transfer Learning [178]	FastText	Sep-19	Attention-based bi-directional LSTM model where input word embeddings fed to Bi-directional LSTM and resulting hidden features being fed to Attention layer. Sentiment word corpora is used as a resource for Attention which is incorporated into the main Attention mechanism operating on top of Bi-directional LSTM layer.
	BERT contextualised word embeddings [47]	WordPiece	Jul-20	Contextualised word embeddings obtained from pre-trained BERT language model encoder. The output of the BERT encoder is fed to the feed-forward layer and a Softmax classifier is used to predict the outcome. Authors provide analysis of Attention heads and relationship of word pairs and the effect on task outcome.

were also surveyed with sentence-level polarity classification and sentence-level subjectivity classification.

The use of pre-trained word embeddings for input representations is a ubiquitous feature of the surveyed studies. Popular word embedding approaches like Word2vec [16, 24, 51, 129, 180], GloVe [14, 24, 77, 129, 134, 155, 174], WordPiece [1, 37, 64, 69, 79, 117, 128, 143, 169], and FastText [129, 178] are effective in providing a model with low-level syntactic and semantic features. The coarse-grain survey also found the utilisation of embeddings layers [60, 114, 121] as alternatives to shallow pre-trained embeddings. Pre-trained word embeddings have been shown to outperform embeddings layers [123] on NLP tasks with few training samples. This edge in performance is due to the large pre-training dataset word embeddings are exposed to which transfers knowledge to downstream task as opposed to learning from scratch.

LSTMs are the prevalent approach for using encoders to learn high-level hidden features from input word embeddings [36, 63, 69, 117, 153, 159]. Bi-directional LSTMs [11, 13, 69] are utilised to capture dependencies on both sides of input and to enhance the modelling of structural information.

The application of Attention mechanism on top of RNN encoders was the dominant approach before the adoption of Transformer language models [14, 24, 77, 134, 155, 174, 178]. In Attention-based LSTM models, Attention layers are fed hidden features from RNN encoders (LSTM or GRU), attending to the most important features on a sentence or word level. Attention-LSTM [14, 24, 134, 155, 174, 178] are commonly applied to capture the hierarchical structure of text and to model the hierarchical relationships in the text. Attention can capture the importance of individual words contributing to the polarity of a sentence at the word level, whereas Hierarchical Attention can determine the relevance of different sentences to the polarity of a document at the sentence-level.

Recent trends in coarse-grain sentiment analysis indicate an increasing adaptation of pre-trained language models over CNN, RNN, and Attention architectures. It is a shift toward a more robust knowledge transfer using deeper pre-trained LSTM [60, 113] and Transformer-based language models [79, 102, 125, 169]. Larger language models pre-trained on vast unlabelled data are able to produce context-dependent embeddings (ELMo [113] and BERT [37]). In line with all other natural language understanding tasks (see GLUE [150] and superGLUE [149]), state-of-the-art Transformer-based language models are the focus of the research community for coarse-grained tasks. Transformer architecture can transfer syntactic and semantic knowledge more robustly on coarse-grain tasks and are relatively simple to fine-tune, requiring fewer samples than CNN/RNN/Attention-based architectures to achieve state-of-the-art performance (refer to Section 8.1 for a detailed discussion).

5.2 Survey of Coarse-Grain Datasets

SST-2 [108], Yelp [171], IMDB [94], and Amazon Review [98] are commonly used benchmark datasets for document-level polarity classification tasks on a binary scale. They comprise of 11,855, 500,000, 50,000, and 142.8 million opinionated samples, respectively. SST-2 and IMBD contain movie reviews while Yelp contains reviews for businesses and Amazon Review covers product reviews. SemEval 2020 Task 12 [175] contains 9.1 million samples and is a binary classification task on offensive language detection using English social media data. SemEval 2018 Task 3 [55] contains 3,000 data samples on a binary scale for irony detection classification task. It should be noted the Yelp binary Classification, IMDB, Amazon Review Polarity, SemEval-2020 Task 12, and SemEval-2018 Task 3 are standalone benchmark datasets and updated less frequently than SST-2. SST-2 dataset is part of the **GLUE (General Language Understanding Evaluation)** leader board which is updated regularly with state-of-the-art architectures and provides a better reflection of the latest trends in coarse-grain architectures. All the above-discussed datasets are dominated by Transformer-based language models (refer to Table 5).

Table 5. Datasets for Coarse-Grain Studies with Top Performing Architectures

Task	Dataset	Rank	Architecture	Performance
Document-level Polarity Classification	SST-2 Binary classification	1	ERNIE 3.0 [143]	ACC: 0.978
		2	METRO-LMXXL [11]	ACC: 0.976
		3	SMART-RoBERTa Large [69]	ACC: 0.975
	Yelp Binary Classification	1	XLNet. [169]	ACC: 0.985
		2	BERT Large [37]	ACC: 0.982
		3	BERT large fine-tune UDA. [162]	ACC: 0.981
	IMDB	1	XLNet [169]	ACC: 0.962
		2	BERT large fine-tune UDA [162]	ACC: 0.958
		3	BERT large [37]	ACC: 0.955
	Amazon Review Polarity	1	BERT Large. [37]	ACC: 0.974
		2	DRNN [151]	ACC: 0.965
		3	SRNN [173]	ACC: 0.953
Offensive Language Detection	SemEval-2020 Task 12: English Offensive Language Identification in social media	1	UHH-LT [158]	F1: 0.920
		2	Galileo [153]	F1: 0.919
		3	Rouges [36]	F1: 0.918
Irony Detection	SemEval-2018 Task 3 Task A: Irony Detection in English Tweets	1	Recurrent CNN RoBERTa [117]	ACC: 0.820
		2	THU NGN [159]	ACC: 0.735
		3	NTUA-SLP [13]	ACC: 0.732

ERNIE 3.0 is a framework for pre-training Transformer XL backbone; it is 1st rank on SST-2 dataset. METRO-LMXXL is a large autoencoding language model pre-trained through efficient denoising. SMART-RoBERTa Large is 3rd rank and is a framework for enhancing generalisation of Transformer language model through efficient fine-tuning for natural language understanding tasks. All three architectures are aimed at natural language understanding tasks and benchmarked using GLUE benchmark of which SST-2 is a part.

Yelp, IMDB, and Amazon datasets are less popular compared to SST-2 and are dominated by less recent Transformer language models XLNet and BERT. XLNet outperforms BERT on both Yelp and IMDB but was not benchmarked on Amazon. It outperforms BERT on comparable datasets and is considered an improvement over BERT.

SemEval-2020 Task 12 is dominated by pre-trained Transformer language models UHH-LT, Galileo and Rouges. UHH-LT achieves 1st rank through an ensemble of ALBERT models. Galileo is 2nd rank with RoBERTa model in large configuration with unsupervised fine-tuning. Rouges uses XLM-RoBERTa-base and XLM-RoBERTa-large models to achieve 3rd position.

SemEval-2018 Task 3 Task A provides a mixed picture with Transformer-based Recurrent CNN RoBERTa architecture in 1st place followed by LSTM-based architecture utilising pre-trained word

embeddings. The 3rd rank is an ensemble of deep learning models which includes word and sentence-level LSTM utilising a majority voting classifier.

6 CROSS-DOMAIN SENTIMENT ANALYSIS

Sentiment analysis tasks are domain-dependent [4], with a domain being defined as a semantic concept that may include consumer items such as laptops or books. Domain dependence remains a central challenge in sentiment analysis. Model parameters tuned using a dataset from a source domain are not guaranteed to generalise well with test data from a different target domain [21]. Additionally, variance in the quality of corpora in source and target domain may also negatively contribute to the performance of a model [138].

Supervised deep learning approaches require large, annotated, high-quality datasets to produce desirable performance with high-quality datasets being difficult to obtain. This imbalance in the availability of training data creates a demand for cross-domain models to avoid the costly process of manually labelling new datasets by humans. Addressing this drawback is the main challenge in the deployment of models that are reliable across different domains with few annotated data.

Cross-domain sentiment analysis is the task of training a model on a single domain or set of multiple domains and performing sentiment analysis on the target domain(s). Cross-domain analysis can involve tasks on fine-grain and coarse-grain levels. The foremost challenges in cross-domain sentiment analysis can be grouped into the following:

- Sparsity in words or phrases used in source and target domain.
- Feature divergence–the domain-specific features being learned by a model trained on a different source domain.
- Polysemy–conflicting semantics between source and a target domain and polarity divergence resulting from source and target domain having different feature polarities.

The remainder of this section covers studies on cross-domain sentiment analysis. The surveyed studies are grouped by their task focus with accompanying details of deep learning architecture applied, published date, input word embeddings, and the main contributions (summarised in Table 6). An analysis of recent trends in the cross-domain sentiment analysis is in Section 6.1, followed by insights into the top-performing architectures on prominent cross-domain benchmark datasets in Section 6.2.

6.1 Trends in Cross-Domain Sentiment Analysis

The predominant tasks in cross-domain sentiment analysis are aspect term polarity classification, document-level polarity classification tasks, and offensive language detection. Tackling domain invariance through domain adaptation is considered the central challenge in cross-domain sentiment analysis. Effective domain adoption enables models to robustly learn domain-invariant sentiment features from the source domain (in the training phase) and apply this knowledge to infer sentiment on the target domain.

Input word embeddings are used to feed low-level syntactic and semantic features to an architecture before tackling domain adaptation. Pre-trained word embeddings like Word2vec [27, 68, 83, 88, 96, 107, 181], GloVe [23, 68], WordPiece [25, 71, 102, 107], and FastText [107] are the prevalent approaches in the surveyed studies. The cross-domain survey revealed a diverse array of training regimes (adversarial training [27, 84], jointly training networks with shared layers [83, 84], and multi-task learning [96, 180] compared to coarse-grain and fine-grain surveys. Adversarial training discovers shared parameters, not specific to the source domain, and provides a model with the ability to capture domain-invariant features using jointly trained adversarial networks. Joint training of networks is designed to make a model classifier more domain invariant

Table 6. Survey of Cross-Domain Sentiment Analysis Studies

Task	Architecture	Embeddings	Date	Contribution
Aspect Term Polarity Classification	Neural Attentive model [168]	Word2vec	Feb-19	Domain classifier for knowledge transfer between source and target domain. Aspect classifier uses domain classifier results and aspect aware document representations to analyse aspect sentiment. Latent allocation Dirichlet learns domain-specific sentiment and aspect representations in a supervised and unsupervised manner. Representations are incorporated into bi-directional LSTM using multi-view Attention.
	IATN [176]	Word2vec	Jul-19	Incorporates information on sentence and aspect level. Two Attention networks, S-net and A-net, recognise common features across domains via domain classification and learn information on aspects using common features as a bridge. Both use the bi-directional LSTM layer to learn hidden representations from the input and interactive word Attention.
	DTLM [25]	Wordpiece	May-21	A **deep transfer learning mechanism (DTLM)** provides a mechanism for transferring domain invariant features by incorporating BERT and Kullback-Leibler divergence. BERT encodes features of input text to a shared representation state. Domain adaptive Kullback-Leibler model removes different feature distributions between target and source.
Document-level Polarity Classification	Adversarial Memory Network [84]	Word2vec	Aug-17	Designed to mirror the hierarchical structure of documents, capturing pivots and non-pivots over all the domains using word and sentence Attention layer. Word Attention captures the relevance of each word in a sentence and sentence-level Attention captures the importance of sentences toward the sentiment. Hierarchical positional embeddings encode position into the model.
	HATN [83]	Word2vec	Apr-18	**Hierarchical Attention Transfer Network (HATN)** Identifies pivots across different domains. Attention captures the pivots. Consists of two parameter-sharing networks: (1) MN-sentiment, aimed at sentiment classification, and (2) MN-domain predicts domain labels from samples. Networks contain memory networks consisting of multiple hops containing Attention and linear layer.
	MAN [27]	Word2vec	Jun-18	**Multinomial adversarial network (MAN)** Capture domain-specific features that influence task. Adopts shared-private paradigm. Composed of four modules: a shared feature extractor capturing domain invariant features, a domain feature extractor that learns features specific to a domain, a text classifier which does the job of classification, and a domain discriminator that predicts the shared feature vector coming from each domain.
	Bifurcated-LSTM [68]	Glove & Word2vec	Jun-18	LSTM produce hidden feature vectors and fed to topic and sentiment classifiers which consist of word-level Attention and Softmax layer. Attention attends to important parts of a sentence in classifiers. Topic classification helps the model distinguish between source and target domain. Bifurcated classifiers enable the model to capture domain-independent sentiment and domain-dependent topic features.

(Continued)

Table 6. Continued

Task	Architecture	Embeddings	Date	Contribution
	Bi-LSTM [88]	Word2vec	Jun-18	Domain-general bi-directional LSTM learns cross-domain features. Domain-specific layer captures domain characteristics. Dot product Attention obtains domain-specific input representation. Domain classifier layers determine the likelihood that input comes from a domain. Adversarial training is used to improve the domain-general representations.
	Bi-LSTM Attention [23]	GloVe	Aug-19	Bi-directional LSTM extracts domain-specific features for the word which are combined with word embeddings to create domain-aware embeddings. Another bi-directional LSTM for domain-aware embeddings to produce sentence features. Utilising domain-aware sentence features like query vector, Attention selects features.
	HANP [96]	Word2vec	Mar-19	A **hierarchical Attention network with prior knowledge information (HANP)** uses prior knowledge to attain domain-specific and independent features simultaneously. Prior knowledge is produced from the pre-training of word2vec embeddings producing knowledge for pivots, non-pivots and dis-pivots. Hierarchical Attention is used to obtain sentiment pivots on word and Sentence-level.
	BERT and XLNet [102]	WordPiece	Nov-19	BERT and XLNet fine-tuned on cross domains. Significant improvements over existing latest models. XLNet more effective than BERT learning contextual data more efficiently with only 50 training examples used, 120 times less than the previous model.
	CFd [170]	Embedding layer	Nov-20	**Class-aware feature self-distillation (CFd)** learns discriminative information from pre-trained language models and the features are self-distilled producing a feature adaptation module. Self-training predicts pseudo labels on target domain data for training.
	UDALM [71]	Wordpiece	Apr-21	**Unsupervised domain adaptation through language modelling (UDALM)** uses mixed classification and masked language modelling loss. This approach helps to adapt to target domain distribution more robustly in a sample efficient manner. BERT language model is further pre-trained on unlabelled target data using masked language modelling. The classifier is trained on source domain labelled data.
Offensive Language Detection	mBERT, LSTM, GRU [107]	Wordpiece, FastText, MUSE	Nov-20	Detects misogyny in a cross-domain and cross-lingual setting. Joint learning framework designed to transfer knowledge between languages using multi-lingual embeddings and LSTM layers. Replacing LSTM layers with mBERT was used with joint learning. Framework tested on other cross-domain tasks.

towards the target and source features. Multi-task learning architectures have shared layers to extract desired features and improve learning capacity in various related tasks.

Before the arrival of state-of-the-art Transformer-based language models, there was a preference for Attention-LSTM and Attention models. LSTMs were also a prevalent choice to model high-level syntactic and semantic information from low-level syntactic and semantic word embeddings. Using bi-directional LSTMs [23, 88, 168] leads to more powerful encoders by enabling the LSTM layers to model left-to-right and right-to-left information from the series of input text.

Table 7. Datasets for Cross-Domain Studies with Top Performing Architectures

Task	Dataset	Rank	Architecture	Performance
Document-level Polarity Classification	Amazon Reviews Dataset (20 source-target domain pairs)	1	XLNet-large [102]	ACC: 0.951
		2	BERT-large [102]	ACC: 0.926
		3	HANP [96]	ACC: 0.878
	Amazon Reviews Dataset (12 source-target domain pairs)	1	UDALM [71]	ACC: 0.917
		2	DPT BERT [71]	ACC: 0.908
		3	CFd [170]	ACC: 0.906

Attention mechanism is commonly used with LSTM encoders in cross-domain studies. It is also used in purely Attention architectures where Attention layers can attend to input word embeddings. Attention can be applied on top of low-level representations from the word embedding input layer [83, 84] and higher-level hidden features produced from LSTM encoders [102, 107, 160]. hierarchical Attention architectures aim to capture domain invariant pivots for word-level and sentence-level information [107, 160].

Like fine-grain and coarse-grain, recent trends in cross-domain point towards the adoption of large pre-trained Transformer language models offering more powerful knowledge transfer capabilities [125]. These approaches take advantage of robust language modelling pre-training for domain adaption [12, 15, 21, 28] and have shown promise with the exploitation of unsupervised language modelling. These language models provide a more robust transfer of linguistic and domain invariant features [71] and self-learning with Transformer models to produce pseudo labels for target training [170]. Bi-directional contextualized Transformer language models (like BERT and XLNet) offer the most promising approach to producing robust domain-invariant architectures. These are the state-of-the-art route for transferring linguistic knowledge learned from pre-training on large domain-general data.

6.2 Survey of Cross-Domain Sentiment Datasets

Amazon review dataset [21] contains 20 and 12 pairs of source and target domain in two different configurations (refer to Table 7). This dataset contains a total of 24 product pairs of source and target domains. Each domain has 6,000 reviews split into 3,000 positive and 3,000 negative samples. Reviews are categorised into positive and negative categories using 5-star Amazon reviews with the positive reviews being 3 or more stars and negative being 2 stars or lower.

For the 20 source-target domain pairs, XLNet-large transformer language model is a top-performing model followed by BERT in the 2nd place and hierarchical Attention network in the 3rd place. XLNet outperforms BERT on almost all natural language understanding tasks [169]. The gap between Transformer architectures XLNet and BERT with Attention model HANP is considerable. This gap also demonstrates the robustness with which Transformers can capture syntax and semantic domain invariant features.

For the 12 source-target domain pairs, the top three performing architectures are Transformer-based UDALM, DPT BERT, and CFd, respectively. These architectures utilised the original BERT Transformer [37] as foundations. UDALM uses mixed classification and masked language model loss to robustly capture domain invariant information. DPT BERT pre-trains BERT with unlabelled target domain data and then fine-tune on the labelled source domain, whereas CFd improves self-training in domain adaption using class-aware feature self-distillation.

7 CROSS-LINGUAL SENTIMENT ANALYSIS

Non-English languages are used by an overwhelming percentage of the human population. With the vast majority of recent deep learning approaches addressing a single language, namely English, state-of-the-art approaches cannot be effectively extended to non-English languages. Cross-lingual approaches that can replicate similar performance on fine-grain and coarse-grain tasks for low-resource languages lead to a more democratic environment for the application of cutting-edge deep learning architectures. Current deep learning approaches in sentiment analysis architectures are trained in a supervised manner [2] relying on manually labelled corpora. Superior quality training sets (reliability, size, and source) greatly affect model performance and generalisation capability [44, 145]. Some languages, like English, have an abundant and diverse selection of coarse-grain and fine-grain datasets to model different tasks. Low-resource languages lack the availability of similar training sets [95]. Cross-lingual sentiment analysis can involve coarse-grain and fine-grain tasks and is aimed at addressing the lack of labelled data in low-resource languages [33].

The issues that arise from this gap in the availability of labelled data for low-resource languages become more challenging with tasks that involve manually cumbersome annotations (like fine-grain sentiment analysis tasks). There is a visible gap in the availability of datasets for coarse-grain tasks compared to fine-grain tasks [111] and this is also reflected in fine-grain tasks lagging in performance (refer to Sections 4 and 5). This gap between coarse and fine-grain tasks crosses over into cross-lingual sentiment analysis where the gap in performance between cross-lingual coarse and fine-grain tasks will worsen for low-resource languages as the dual effect of fewer datasets and inferior performance will compound.

The remainder of this section covers important studies on cross-lingual sentiment analysis. The relevant studies are grouped into surveyed tasks (aspect term polarity classification, targeted polarity classification, document-level polarity classification, and offensive language and hate speech detection) and are summarised in Table 8. We provide a brief technical summary for each study survey outlining key contributions. This is followed by a survey and analysis of the top-performing architectures on prominent cross-lingual benchmark datasets.

7.1 Trends in Cross-Lingual Sentiment Analysis

Knowledge transfer is the prevalent approach for tackling cross-lingual sentiment analysis. It involves modelling linguistic knowledge in resource-rich languages and utilising this through the transfer of knowledge over to low-resource languages. The common approaches utilised for cross-lingual knowledge transfer include machine translation [29, 44, 115, 131, 145], bilingual embeddings [41, 42, 78], cross-lingual embeddings [8, 28, 33, 41, 95, 111, 131] and cross-lingual language model pre-training and/or fine-tuning [22, 33, 41, 84, 131].

Cross-lingual and bilingual embeddings are the predominant choice for cross-lingual input features and are less resource-intensive on data and model complexity compared to machine translation systems [131]. Embeddings provide cross-lingual features in a joint distributional space. Techniques used for generating cross-lingual and bilingual fall into the following: (a) word-level alignment originating from parallel word-aligned data derived from cross-lingual and bilingual dictionaries with translated words in different languages [99], (b) automatically aligned words in a parallel corpus and sentence-level alignments using a parallel corpus similar to machine translation [56], (c) using available word-aligned information [184], and (d) document-level alignments (requiring translation of documents in different languages [57] or comparable document-aligned corpora typically topic or task-oriented [148]).

With machine translation systems, a low-resource non-English language is first translated into a resource-rich language (usually English) and is then utilised for downstream task on target language. Despite being well established, machine translation is less utilised than cross-lingual

Table 8. Survey of Cross-Lingual Sentiment Analysis Studies

Task	Architecture	Embeddings	Date	Contribution
Aspect Term Polarity Classification	CNN & Cross-lingual Embeddings [66]	BWE	Jun-19	Cross-lingual embeddings with CNN. A model trained on labelled data from the source language and performs with zero-shot. A multi-layered CNN is used for the sequence tagging model. English language was used source model while Spanish and Dutch were used as the target in zero-shot setting.
	RTCLD [160]	Wordpiece	Jan-21	**Reinforced Transformer with Cross-Lingual Distillation (RTCLD)** uses a Bi-lingual lexicon for domain-specific training to translate aspect category in source-corpus and translate sentences from source to target via machine translation. RTCLD is combined with target-sensitive adversarial learning to minimize undesirable effects of translation ambiguities. Knowledge distillation through language classifier to learn aspect-aware knowledge.
Targeted Polarity Classification	Bi-lingual Sentiment Embeddings [12]	BWE	Jun-19	Sentence and target-level algorithm. For target level, input sequence split to get left and right context of the target. Averaging is utilised for the left to target, right to target, and target separately to obtain compositional vector representations. Features concatenated and fed to Softmax classifier to predict output class.
	Adversarial Deep Averaging Network [28]	BWE	Dec-18	Labelled English data for classification in resource-poor languages with Bi-lingual word embeddings in three stages (with adversarial deep averaging and feed-forward networks): A joint feature extractor maps input to a shared feature space; a classifier predicts label using the output of stage 1 and a language discriminator takes the labels to predict input language.
	Conv-Char-R [15]	Character embedding	May-17	Cost-effective character-based embeddings were obtained using convolutions on multiple separate languages. Several convolutional layers serve as an embedding learning layer, resulting in few alphabets in memory instead of a large vocabulary.
	NNLS [156]	Word2vec	Apr-18	NNLS architecture consists of input, convolutional, pooling, concatenation, dense and output layers. It employs different filter sizes to learn info on low and mid-level text semantics for sentiment analysis and has 2-3 times fewer parameters than word-embedding approaches.
Document-level Polarity Classification	MultiFiT [41]	Subword unigram	Nov-19	Discriminative fine-tuning. QRNN enable faster training and fine-tuning. Subword tokenization is also used over words as input. Label smoothing and a novel cosine variant of one cycle policy were found to outperform slanted triangular learning rate scheduler and gradual unfreezing used in ULMFit.
	Multi-lingual Sen. Embeddings [8]	BPE	Sep-19	Bi-directional LSTM encoder with an auxiliary decoder to capture joint cross-lingual sentence representations. Train classifier on labelled English data before utilising it on 93 languages. Bi-directional LSTM encoder is fed joint byte-pair encoding vocabulary to produce sentence embeddings. This is fed to the auxiliary decoder. A single encoder-decoder architecture is trained on parallel corpora.
	ELSA [29]	skip-gram	Jun-18	Sentence representation model constructed for source and target languages using tweets acquired using emoji prediction and embeddings. Google translation is used to convert English documents into target language and fed to sentence representation models to produce language representations. Sentence representations are aggregated to construct documents for sentiment classification.

(Continued)

Table 8. Continued

Task	Architecture	Embeddings	Date	Contribution
	Encoder-Classifier [42]	BWE	Sep-18	Leveraging cross-lingual neural machine translation. Knowledge from encoder trained on cross-lingual machine translation is integrated with task-specific classifier. The final hidden layer is composed of GRUs and is used by the classifier to predict a label.
	XLM+UDA [78]	BWE	Dec-19	Labelled source data was applied alongside unlabelled target language data. Unsupervised domain adaptation and weak supervision utilised. Unlabelled target corpora are used for masked language model pre-training while unsupervised data augmentation produces synthetic paraphrases from the unlabelled target corpus. This scores the unlabelled data as a teacher model, and then fine-tunes XLM model using pseudo-labelled data.
	Cross-lingual BERT [46]	Wordpiece	Jun-19	Pre-trained cross-lingual BERT. Additional tweet pre-training as languages on Wikipedia. Masked language model task with 20,000 additional steps to predict whether a tweet and author description matched a tweet. User description was used with additional 20,000 steps of pre-training and the model was pre-trained with twitter screen usernames as a secondary prediction task.
	Self-Learning Multi-lingual BERT [39]	Wordpiece	Nov-19	Self-learning framework using unlabelled source (English) samples during fine-tuning of pre-trained cross-lingual models. Cross-lingual model's self-prediction is used on unlabelled non-English samples to produce additional information that is leveraged during fine-tuning.
	CFd [170]	Embedding layer	Nov-20	**Class-aware feature self-distillation (CFd)** learns discriminative information from pre-trained language models and pre-trained language model features are self-distilled producing a feature adaptation module. Self-training predicts pseudo labels on target domain data for training. Pre-trained Transformer language model used.
Hate Speech Detection	STUFIIT [22]	MUSE & ELMo	Jun-19	Cross-lingual MUSE embeddings were used with LSTM and CNN+LSTM networks and mono-lingual ELMo embeddings were used with an adversarial model containing Attention-LSTM or 1d convolutional layer and max pooling. Switching from MUSE to ELMo produced a noticeable gain.
	Multi-lingual BERT (mBERT) [5]	LASER sentence & MUSE word	Sep-20	Combination of cross-lingual sentence and word embeddings used in combination with CNN-GRU, BERT and mBERT models. 9 source languages from 16 datasets were used for experiments. mBERT based models were found to be more effective with resource-rich languages.
Offensive Language Detection	Galileo [153]	Wordpiece	Oct-20	Using cross-lingual language models XLM-R and ERNIE with knowledge distillation technique executed on soft labels generated using numerous supervised models.
	mBERT, LSTM, GRU [107]	Wordpiece, FastText, MUSE	Nov-20	Detecting misogyny in a cross-domain and cross-lingual setting. Joint learning framework designed to transfer knowledge between languages using multi-lingual embeddings and LSTM layers. Replacing LSTM layers with mBERT was used with joint learning. Framework tested on other cross-domain tasks.

and bilingual embeddings. It suffers from several challenges such as bias towards domains [75], difficulty preserving sentiment features [101], inaccurate translation leading to fragmented coherent sentences and introducing noise and above all, a requirement for a large corpus of parallel text [146] across the source and target languages. These drawbacks undercut the key benefits of cross-lingual sentiment analysis.

Large pre-trained language models have revolutionised deep learning in NLP through the robust transfer of knowledge from large domain-general pre-training to downstream mono-lingual tasks. This trend is also observed across cross-lingual studies. The surveyed approaches utilising pre-trained language models exploit both machine translation [78, 160] and cross-lingual embeddings [28, 41, 84, 95, 131] to transfer semantic and syntactic knowledge across languages. Cross-lingual embeddings are the dominant choice for transferring cross-lingual information. MultiFiT [41] does not exploit cross-lingual embeddings or machine translation and initially undergoes pre-training as a mono-lingual language model before being fine-tuned on as many as 100 target language samples. Recent bi-directional contextual Transformer models have dominated NLP tasks with respect to research activity and this is evident in cross-lingual sentiment analysis. **Multi-lingual BERT (mBERT)** [37] and its variant XLM [33] perform particularly well. mBERT, XLM, and MultiFiT have demonstrated state-of-the-art performance on zero-shotting downstream tasks offer the most promise on cross-lingual tasks.

mBERT applies a cross-lingual masked language model pre-training with a cross-lingual vocabulary. mBERT training dataset is a cross-lingual corpus of 104 languages that are mapped to a shared space without knowing the language identities which is encoded during pre-training. XLM enhanced mBERT with unsupervised mono-lingual corpus using different language modelling techniques (causal language modelling, supervised cross-lingual language modelling using parallel data and translation language modelling using a parallel corpus).

The recent growth in applying self-training to bridge the generalisation gap between source and target language has been effective with unsupervised data augmentation [28, 33, 44] to predict pseudo labels on the unlabelled target language. This technique is used in the fine-tuning of a pre-trained cross-lingual language model. Generating pseudo labels can introduce noise into a training regime and techniques have been introduced to tackle this issue like jointly mapping discriminative features of pre-trained language model target language [39].

7.2 Survey of Cross-Lingual Datasets

Amazon reviews dataset [119] and MLDoc [136], SemEval 2016 Task 5 review [115], and SemEval-2020 Task 12 [175] are the four widely used benchmark datasets for cross-lingual sentiment analysis. Amazon product review and MLDoc datasets are annotated for document-level polarity classification tasks. SemEval 2016 Task 5 review dataset is used for aspect term polarity classification task and SemEval-2020 Task 12 dataset is used for offensive language detection task in social media. These benchmark datasets are summarised by their task focus with the top-performing architectures and performance ranking in Table 9.

SemEval 2016 Task 5 [115] provides laptop and restaurant reviews for English, Spanish, Russian, and Chinese languages with 2,406, 2,600, 3,159 and 2,024 samples, respectively. The reviews are produced on a two-point polarity scale (positive and negative). English-Spanish, English-Russian and English-Chinese rankings are dominated by the following Transformer-based architectures (RTCLD followed by self-learning BERT and Dual BERT).

The Amazon product review dataset contains 800,000 Amazon product reviews for English, German, and French. The reviews are on three product categories which include books, DVDs, and music. In both English-German and English-French, XLM+UDA is state of the art followed by

Table 9. Datasets for Cross-Lingual Studies with Top Performing Architectures

Task	Dataset	Rank	Architectures	Performance
Aspect Term Polarity Classification	SemEval 2016 Task 5 EN-SP	1	RTCLD [160]	ACC: 0.878
		2	Dual BERT [35]	ACC: 0.858
		3	Self-Learning Multi-lingual BERT [39]	ACC: 0.842
	SemEval 2016 Task 5 EN-RU	1	RTCLD [12]	ACC: 0.822
		2	Self-Learning Multi-lingual BERT [39]	ACC: 0.818
		3	Dual BERT [35]	ACC: 0.810
	SemEval 2016 Task 5 EN-CH	1	RTCLD [160]	ACC: 0.803
		2	Self-Learning Multi-lingual BERT [39]	ACC: 0.794
		2	Dual BERT [35]	ACC: 0.794
Document-level Polarity Classification	Amazon product review EN-DE	1	XLM+UDA [78]	ACC: 0.942
		2	MultiFiT [41]	ACC: 0.922
	Amazon product review EN-FR	1	XLM+UDA [78]	ACC: 0.933
		2	MultiFiT [41]	ACC: 0.914
	MLDoc EN-DE	1	MultiFiT [41]	ACC: 0.959
		2	XLM+UDA [78]	ACC: 0.954
	MLDoc EN-FR	1	XLM+UDA [78]	ACC: 0.953
		2	MultiFiT [41]	ACC: 0.948
	MLDoc EN-ZH	1	MultiFiT [41]	ACC: 0.925
		2	XLM+UDA [78]	ACC: 0.919
Offensive Language Detection	SemEval-2020 Task 12: Danish	1	LT@Helsinki [120]	F1: 0.811
		2	Galileo [153]	F1: 0.802
		3	NLPDove [3]	F1: 0.792
	SemEval-2020 Task 12: Greek	1	NLPDove [3]	F1: 0.852
		2	Galileo [153]	F1: 0.850
		3	KS@LTH [139]	F1: 0.848
	SemEval-2020 Task 12: Turkish	1	Galileo [153]	F1: 0.825
		2	SU-NLP [106]	F1: 0.816
		3	KUISAIL [133]	F1: 0.814

MultiFiT. This is in line with other tasks in fine and coarse grain with a Transformer-based language model, in this case, BERT, outperforming an RNN based language model.

MLDoc contains the four categories Economics, Markets, Government/Social, and Corporate/Industrial. Each category contains 250 samples for training. It covers English, German, French, and Simplified Chinese languages. MultiFiT can generalise on fewer training samples. It is top-ranked for England-German and English-Chinese cross-lingual tasks. XLM+UDA is top-ranked for English-French.

SemEval-2020 Task 12 contains social media samples on Danish, Greek, and Turkish languages with 2,961, 8,743, and 31,756 samples, respectively. The types of social media content include Twitter, Facebook, Reddit, and local newspapers. Transformer-based architecture utilising a BERT variant for Danish is LT@Helsinki, NLPDove for Greek, and Galileo for Turkish.

8 OPEN RESEARCH CHALLENGES

The survey of fine-grain, coarse-grain, cross-domain and cross-lingual sentiment analysis provided a picture of incremental performance improvements in recent years with noticeable jumps with the introduction and adoption of Transformer-based language models. Simultaneously, several important and unresolved challenges were observed. The most pressing challenges in sentiment analysis with deep learning are summarised below:

- Ambiguity, compositionality, long-term dependencies, anaphora, sarcasm and negation in long and short text remain challenges in NLP and sentiment analysis. In particular, ambiguity and sarcasm have been stubborn challenges to overcome, and this is evident in the majority of incorrect classifications in document, sentence, aspect and target level tasks originating from the ground-truth neutral class.
- The lack of domain invariance among the surveyed architectures. Recent trends in pre-trained language models have resulted in impressive gains in cross-domain sentiment analysis. Despite this, performance for cross-domain tasks still lags architectures trained and tested on the same domain and further work is required to produce models with better generalisation capabilities across different domains.
- Cross-lingual performance, especially for resource-poor, non-English languages, still trails mono-lingual tasks. First, only a limited number of models are massively cross-lingual. Most approaches are trained and evaluated in a few languages. Massively cross-lingual models will fulfil a vital role in making utilisation feasible on more than just a half dozen widely resourced languages.
- Zero-shot or few-shot learning remain a challenge in cross-lingual and cross-domain sentiment analysis with the upside of producing massively cross-domain and cross-lingual approaches.
- A frequently overlooked challenge in cross-lingual sentiment analysis is the lack of domain invariance leading to cross-lingual domain mismatch. This problem requires bridging the language and domain gap and can be considered a hybrid problem involving both cross-domain and cross-lingual sentiment analysis.
- Pre-trained language models have pushed the boundaries of many NLP tasks. This often comes at the price of extremely large architectures with hundreds of millions or billions of parameters that require substantial resources to train and deploy. This is a barrier for institutions, small and medium enterprises, and independent researchers without access to a large pool of resources, effectively preventing the democratisation of cutting-edge deep learning research. Research in producing more efficient architectures will be crucial to tackling this emerging problem in the fields of machine learning and artificial intelligence.

9 CONCLUSION

This paper provided a detailed contemporary survey of sentiment analysis tasks at fine-grain, coarse-grain, cross-domain and cross-lingual granularities. It covers recent trends in deep learning architectures, covering published works between a five-year period (1st July 2017 to 1st July 2022). A survey of prominent benchmark datasets and top performing architectures is also provided for the above-mentioned granularities and their respective tasks.

Prior to the ascendance of Transformer language models which now dominate deep learning, NLP and sentiment analysis, CNN, RNN and Attention architectures were extensively utilised to capture high-level semantic and syntax features from low-level word embedding features. Traditional word embeddings techniques use shallow models to produce word embeddings with low-level syntactic and semantic information. Recent trends indicate a trend away from CNN, RNN, and Attention architectures and a widespread adoption of bi-directional contextual Transformer language models. Transformer language models require fewer additional fine-tuning and can model natural language tasks more effectively through pre-training on large corpus of domain-general data.

Cross-domain sentiment analysis tasks have had strong gains in performance from the application of large pre-trained Transformer language models such as BERT and XLNet. These language models have pushed forward domain adaption through robust pre-training (language modelling) than the combination of shallow word embeddings and CNN, RNN, and Attention architectures.

There is ongoing research activity to bridge the gap between low resource and high resource languages. Large pre-trained language models like mBERT, XLM, and MultiFiT have had a noticeable leap in progress for cross-lingual sentiment analysis and have greatly improved performance compared to CNN, RNN, and Attention architectures. Further work is required to make massively multi-lingual architectures that perform competitively across monolingual tasks.

REFERENCES

[1] A. Ahmet and T. Abdullah. 2020. Real-time social media analytics with deep transformer language models: A big data approach. In *IEEE 14th International Conference on Big Data Science and Engineering (BigDataSE)*. Guangzhou.

[2] A. Ahmet and T. Abdullah. 2020. Recent trends and advances in deep learning-based sentiment analysis. In *Deep Learning-based Approaches for Sentiment Analysis*. Springer. 33–56 pages.

[3] H. Ahn, J. Sun, C. Y. Park, and J. Seo. 2020. NLPDove at SemEval-2020 task 12: Improving offensive language detection with cross-lingual transfer. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona (online)*.

[4] T. Al-Moslmi, N. Omar, S. Abdullah, and M. Albared. 2017. Approaches to cross-domain sentiment analysis: A systematic literature review. *IEEE Access* 5 (2017), 16173–16192.

[5] S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee. 2021. A deep dive into multilingual hate speech classification. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD* 2020 (2021), 2020.

[6] S. Chandar AP, S. Lauly, H. Larochelle, M. Khapra, B. Ravindran, V. C. Raykar, and A. Saha. 2014. An autoencoder approach to learning bilingual word representations. *Advances in Neural Information Processing Systems* 27 (2014).

[7] M. Artetxe, G. Labaka, and E. Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia.

[8] Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics* 7 (2019), 597–610.

[9] A. Badawy, E. Ferrara, and K. Lerman. 2018. Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

[10] D. Bahdanau, K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. (2014).

[11] P. Bajaj, C. Xiong, G. Ke, X. Liu, D. He, S. Tiwary, T.-Y. Liu, P. Bennett, X. Song, and J. Gao. 2022. METRO: Efficient Denoising Pretraining of Large Scale Autoencoding Language Models with Model Generated Signals. (2022).

[12] J. Barnes and R. Klinger. 2019. Embedding projection for targeted cross-lingual sentiment: Model comparisons and a real-world study. *Journal of Artificial Intelligence Research* 66 (2019), 691–742.

[13] C. Baziotis, A. Nikolaos, P. Papalampidi, A. Kolovou, G. Paraskevopoulos, N. Ellinas, and A. Potamianos. 2018. NTUA-SLP at SemEval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive RNNs. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. New.

[14] C. Baziotis, N. Pelekis, and C. Doulkeridis. 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)*.

[15] W. Becker, J. Wehrmann, H. E. L. Cagnini, and R. C. Barros. 2017. An efficient deep neural architecture for multilingual sentiment analysis in Twitter. In *The Thirtieth International FLAIRS Conference*.
[16] L. B. Belisário, L. G. Ferreira, and T. A. S. Pardo. 2020. Evaluating richer features and varied machine. *Information* 11, 9 (2020), 437.
[17] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research* 3 (2003), 1137–1155.
[18] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks* 5, 2 (1994), 157–166.
[19] S. Biere and S. Bhulai. 2018. Hate Speech Detection Using Natural Language Processing Techniques. (2018).
[20] M. Birjali, M. Kasri, and A. Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems* 226 (2021), 107134.
[21] J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.
[22] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.
[23] Y. Cai and X. Wan. 2019. Multi-domain sentiment classification based on domain-aware embedding and attention. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*.
[24] R. Cao, R. K.-W. Lee, and T.-A. Hoang. 2020. DeepHate: Hate speech detection via multi-faceted text representations. In *12th ACM Conference on Web Science*.
[25] Z. Cao, Y. Zhou, A. Yang, and S. Peng. 2021. Deep transfer learning mechanism for fine-grained cross-domain sentiment classification. *Connection Science* (2021), 1–18.
[26] P. Chen, Z. Sun, L. Bing, and W. Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.
[27] X. Chen and C. Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. (Long Papers). New Orleans, Louisiana.
[28] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics* 6 (2018), 557–570.
[29] Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. 2019. Emoji-powered representation learning for cross-lingual sentiment classification. In *The World Wide Web Conference*. 251–262.
[30] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Association for Computational Linguistics*. Doha, Qatar.
[31] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).
[32] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. 2003. ELECTRA: Pre-training text encoders as discriminators rather than generators. (2003).
[33] A. Conneau and G. Lample. 2019. Cross-lingual language model pretraining. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. Vancouver, Canada.
[34] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations*. Vancouver, Canada.
[35] Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang, S. Wang, and G. Hu. 2019. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2019), 3504–3514.
[36] T. Dadu and K. Pant. 2020. Team rouges at SemEval-2020 task 12: Cross-lingual inductive transfer to detect offensive language. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona (online)*.
[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*. 2019, Minneapolis, Minnesota.
[38] X. Ding, T. Liu, J. Duan, and J.-Y. Nie. 2015. Mining user consumption intention from social media using domain adaptive convolutional neural network. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
[39] Xin Luna Dong and Gerard de Melo. 2019. A robust self-learning framework for cross-lingual text classification. 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)*. 6306–6310.
[40] L. Duong, T. Cohn, S. Bird, and P. Cook. 2015. Cross-lingual transfer for unsupervised dependency parsing without parallel data. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*.
[41] J. Eisenschlos, S. Ruder, P. Czapla, M. Kardas, S. Gugger, and J. Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. In *Association for Computational Linguistics, Hong Kong*.

[42] A. Eriguchi, M. Johnson, O. Firat, H. Kazawa, and W. Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *ArXiv* 1809, 4686 (2018).
[43] M. Faruqui and C. Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*.
[44] G. Fenza, M. Gallo, V. Loia, F. Orciuoli, and E. Herrera-Viedma. 2021. Data set quality in machine learning: Consistency measure based on group decision making. *Applied Soft Computing* 106 (2021), 107366.
[45] Clayton R. Fink, Danielle S. Chou, Jonathon J. Kopecky, and Ashley J. Llorens. 2011. Coarse- and fine-grained sentiment analysis of social media text. *Johns Hopkins Apl. Technical Digest* 30, 1 (2011), 22–30.
[46] Gartner. 2019. Gartner Identifies the Top 10 Strategic Technology Trends for 2018. (2019). https://www.gartner.com/en/newsroom/press-releases/2017-10-04-gartner-identifies-the-top-10-strategic-technology-trends-for-2018.
[47] J. Á. González, L. F. Hurtado, and Ferran Pla. 2020. Transformer based contextualization of pre-trained word embeddings for irony detection in Twitter. *Information Processing & Management* 57, 4 (2020), 102262.
[48] S. Gouws, Y. Bengio, and G. Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *International Conference on Machine Learning*.
[49] A. Graves, G. Wayne, and I. Danihelka. 2014. *Neural Turing Machines*. (2014).
[50] J. C. Hay, B. E. Lynch, and D. R. Smith. 1960. *Mark I Perceptron Operators' Manual*. (1960).
[51] D. Hazarika, S. Poria, S. Gorantla, E. Cambria, R. Zimmermann, and R. Mihalcea. 2018. CASCADE: Contextual sarcasm detection in online discussion forums. In *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA.
[52] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654* (2020).
[53] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier. 2018. Effective attention modeling for aspect-level sentiment classification. In *Proceedings of the 27th International Conference on Computational Linguistics*.
[54] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier. 2018. Exploiting document knowledge for aspect-level sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia.
[55] C. Van Hee, E. Lefever, and V. Hoste. 2018. Semeval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*.
[56] K. M. Hermann and P. Blunsom. 2013. Multilingual distributed representations without word alignment. *ArXiv* 1312, 6173 (2013).
[57] K. M. Hermann and P. Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland.
[58] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9 (1997), 1735–1780.
[59] J. Howard. 2013. The business impact of deep learning. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
[60] J. Howard and S. Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. (Long Papers). Melbourne.
[61] B. Huang and K. Carley. 2018. Parameterized convolutional neural networks for aspect level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.
[62] B. Huang, Y. Ou, and K. M. Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*.
[63] S. Ilić, E. M. Taylor2, J. A. Balazs, and Y. Matsuo. 2018. Deep contextualized word representations for detecting sarcasm and irony. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Brussels, Belgium.
[64] V. Indurthi, B. Syed, M. Shrivastava, N. Chakravartula, M. Gupta, and V. Varma. 2019. Fermi at semeval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*.
[65] M. R. Islam, S. Liu, X. Wang, and G. Xu. 2020. Deep learning for misinformation detection on online social networks: A survey and new perspectives. *Social Network Analysis and Mining* 10 (2020), 1–20.
[66] S. Jebbara and P. Cimiano. 2019. Zero-shot cross-lingual opinion target extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota.
[67] B. Jeong, J. Yoon, and J.-M. Lee. 2019. Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management* 48 (2019), 280–290.

[68] J. Ji, C. Luo, X. Chen, L. Yu, and P. Li. 2018. Cross-domain sentiment classification via a bifurcated-LSTM. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*.
[69] H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao. 2019. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
[70] A. Karimi, L. Rossi, and A. Prati. 2020. Adversarial training for aspect-based sentiment analysis with BERT. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 2020.
[71] C. Karouzos, G. Paraskevopoulos, and A. Potamianos. 2021. UDALM: Unsupervised Domain Adaptation through Language Modeling. (2021).
[72] Y. Kim. 2014. Convolutional neural networks for sentence classification. In *Association for Computational Linguistics*. (2014).
[73] A. Klementiev, I. Titov, and B. Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING*.
[74] J. Klinger, J. C. Mateos-Garcia, and K. Stathoulopoulos. 2018. Deep learning, deep change? Mapping the development of the artificial intelligence general purpose technology. In *Mapping the Development of the Artificial Intelligence General Purpose Technology*.
[75] P. Koehn and R. Knowles. 2017. Six challenges for neural machine translation. (2017).
[76] T. Kudo and J. Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural*. Brussels, Belgium.
[77] A. Kumar, S. R. Sangwan, A. Arora, A. Nayyar, and M. Abdel-Basset. 2019. Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network. *IEEE Access* 7 (2019), 23319–23328.
[78] Guokun Lai, Barlas Oguz, Yiming Yang, and Veselin Stoyanov. 2019. Bridging the domain gap in cross-lingual document classification. *arXiv preprint arXiv:1909.07009* (2019).
[79] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. In *Eighth International Conference on Learning Representations, Addis Ababa*. (2019).
[80] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86 (1998), 2278–2324.
[81] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. (2019).
[82] L. Li, Y. Liu, and A. Zhou. 2018. Hierarchical attention based position-aware network for aspect-level sentiment analysis. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*.
[83] Z. Li, Y. Wei, Y. Zhang, and Q. Yang. 2018. Hierarchical attention transfer network for cross-domain sentiment classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
[84] Z. Li, Y. Zhang, Y. Wei, Y. Wu, and Q. Yang. 2017. End-to-end adversarial memory network for cross-domain sentiment classification. In *IJCAI*. (2017).
[85] B. Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.
[86] F. Liu, T. Cohn, and T. Baldwin. 2018. Recurrent entity networks with delayed memory update for targeted aspect-based sentiment analysis. In *Proceedings of NAACL-HLT*. 2018, New Orleans, Louisiana.
[87] Ning Liu and Bo Shen. 2020. Aspect-based sentiment analysis with gated alternate neural network. *Knowledge-Based Systems* 188 (2020), 105010.
[88] Q. Liu, Y. Zhang, and J. Liu. 2018. Learning domain representation for multi-domain sentiment classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume*. 1 (Long Papers).
[89] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. (2019).
[90] M.-T. Luong, H. Pham, and C. D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*.
[91] M.-T. Luong, H. Pham, and C. D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal.
[92] D. Ma, S. Li, X. Zhang, and H. Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*. Melbourne, Australia.
[93] Y. Ma, H. Peng, and E. Cambria. 2018. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[94] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

[95] A. Magueresse, V. Carles, and E. Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *CoRR* 2006, 7264 (2020).

[96] T. Manshu and W. Bing. 2019. Adding prior knowledge in hierarchical attention neural network for cross domain sentiment classification. *IEEE Access* 7 (2019), 32578–32588.

[97] M. Mazloom, R. Rietveld, S. Rudinac, M. Worring, and W. Van Dolen. 2016. Multimodal popularity prediction of brand-related social media posts. In *Proceedings of the 24th ACM International Conference on Multimedia* (2016).

[98] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[99] T. Mikolov, Q. V. Le, and I. Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR* 1309, 4168 (2013).

[100] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. (2013).

[101] S. M. Mohammad, M. Salameh, and S. Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research* 55 (2016), 95–130.

[102] B. Myagmar, J. Li, and S. Kimura. 2019. Cross-domain sentiment classification with bidirectional contextualized transformer language models. *IEEE Access* 7 (2019), 163219–163230.

[103] V. Nahar, S. Unankard, X. Li, and C. Pang. 2012. Sentiment analysis for effective detection of cyber bullying. In *Asia-Pacific Web Conference*. (2012).

[104] nlpprogress. 2022. NLP-progress. (2022). http://nlpprogress.com/english/sentiment_analysis.html.

[105] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth International AAAI Conference on Weblogs and Social Media*.

[106] A. Ozdemir and R. Yeniterzi. 2020. SU-NLP at SemEval-2020 task 12: Offensive language identification in Turkish tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona (online)*.

[107] E. W. Pamungkas, V. Basile, and V. Patti. 2020. Misogyny detection in Twitter: A multilingual and cross-domain study. *Information Processing & Management* 57 (2020), 102360.

[108] Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075* (2005).

[109] M. Paolanti, C. Kaiser, R. Schallner, E. Frontoni, and P. Zingaretti. 2017. Visual and textual sentiment analysis of brand-related social media pictures using deep convolutional neural networks. In *International Conference on Image Analysis and Processing*.

[110] paperswithcode. 2020. Aspect-Based Sentiment Analysis on SemEval 2014 Task 4 Sub Task 2. (2020). https://paperswithcode.com/sota/aspect-based-sentiment-analysis-on-semeval.

[111] paperswithcode. 2022. Natural Language Processing. (2022). https://paperswithcode.com/area/natural-language-processing.

[112] J. Pennington, R. Socher, and C. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[113] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*. 2018, New Orleans, Louisiana.

[114] G. K. Pitsilis, H. Ramampiaro, and H. Langseth. 2018. Detecting offensive language in tweets using deep learning. (2018).

[115] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, and O. De Clercq. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International Workshop on Semantic Evaluation*.

[116] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin.

[117] R. A. Potamias, G. Siolas, and A.-G. Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications* 32, 23 (2020), 17309–17320.

[118] N. Dwi Prasetyo and C. Hauff. 2015. Twitter-based election prediction in the developing world. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*.

[119] Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. 1118–1127.

[120] M. Pámies, E. Öhman, K. Kajava, and J. Tiedemann. 2020. LT@Helsinki at SemEval-2020 task 12: Multilingual or language-specific BERT?. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona (Online)*.

[121] J. M. Pérez and F. M. Luque. 2019. Atalaya at SemEval 2019 task 5: Robust embeddings for tweet classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation.*
[122] V.-D. Păvăloaia, E.-M. Teodor, D. Fotache, and M. Danileț. 2019. Opinion mining on social media data: Sentiment analysis of user preferences. *Sustainability* 11 (2019), 4459.
[123] Ye Qi, Devendra Singh Sachan, Matthieu Felix, Sarguna Janani Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? *arXiv preprint arXiv:1804.06323* (2018).
[124] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. 2018. Improving language understanding by generative pre-training. (2018). https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language.
[125] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21 (2019), 1–67.
[126] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis.*
[127] S. Ramaswamy and N. DeClerck. 2018. Customer perception analysis using deep learning and NLP. *Procedia Computer Science* 140 (2018), 170–178.
[128] A. Rietzler, S. Stabinger, P. Opitz, and S. Engl. 2019. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. (2019).
[129] G. Rizos, K. Hemker, and B. Schuller. 2019. Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management.*
[130] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1985. Learning internal representations by error propagation. (1985).
[131] I. Vulić S. Ruder and A. Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research* 65 (2019), 569–631.
[132] M. Saeidi, G. Bouchard, M. Liakata, and S. Riedel. 2016. SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka.*
[133] A. Safaya, M. Abdullatif, and D. Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech. In *Proceedings of the 14th International Workshop on Semantic Evaluation.* Barcelona, Spain.
[134] R. Satapathy, S. Pardeshi, and E. Cambria. 2022. Polarity and subjectivity detection with multitask learning and BERT embedding.
[135] M. Schuster and K. Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*
[136] H. Schwenk and X. Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki.
[137] S. A. Ali Shah, I. Uddin, F. Aziz, S. Ahmad, M. A. Al-Khasawneh, and M. Sharaf. 2020. An enhanced deep neural network for predicting workplace absenteeism. *Complexity* 2020, (2020).
[138] J. Shen, P.-J. Chen, M. Le, J. He, J. Gu, M. Ott, M. Auli, and M. Ranzato. 2019. The source-target domain mismatch problem in machine translation. (2019).
[139] K. Socha. 2020. KS@LTH at SemEval-2020 Task 12: Fine-tuning multi- and monolingual transformer models for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona (online).*
[140] We Are Social. 2022. Digital 2022: Another Year of Bumper Growth. *We Are Social* 10, 05 2022 (2022). https://wearesocial.com/uk/blog/2022/01/digital-2022-another-year-of-bumper-growth-2/.
[141] Y. Song, J. Wang, T. Jiang, Z. Liu, and Y. Rao. 2019. Attentional encoder network for targeted sentiment classification. (2019).
[142] C. Sun, L. Huang, and X. Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of NAACL-HLT.* 2019, Minneapolis, Minnesota.
[143] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, and Y. Lu. 2021. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. (2021).
[144] D. Tang, B. Qin, and T. Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Austin, Texas.
[145] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox. 2003. Natural language processing advancements by deep learning: A survey. (2003).
[146] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ÅĄ. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems.*
[147] I. Vulic and M.-F. Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*

[148] I. Vulić and M. F. Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

[149] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems* 32 (2019).

[150] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018). https://gluebenchmark.com/.

[151] B. Wang. 2018. Disconnected recurrent neural networks for text categorization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[152] B. Wang and W. Lu. 2018. Learning latent opinions for aspect-level sentiment classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[153] S. Wang, J. Liu, X. Ouyang, and Y. Sun. 2020. Galileo at SemEval-2020 Task 12: Multi-lingual learning for offensive language identification using pre-trained language models. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona (online)*.

[154] Y. Wang, M. Huang, and L. Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

[155] Y. Wang, A. Sun, J. Han, Y. Liu, and X. Zhu. 2018. Sentiment analysis by capsules. In *Proceedings of the 2018 World Wide Web Conference*.

[156] Jônatas Wehrmann, Willian E. Becker, and Rodrigo C. Barros. 2018. A multi-task neural network for multilingual sentiment classification and language detection on Twitter. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. 1805–1812.

[157] D. Weiss, C. Alberti, M. Collins, and S. Petrov. 2015. Structured training for neural network transition-based parsing. (2015).

[158] G. Wiedemann, E. Ruppert, and C. Biemann. 2019. UHH-LT at SemEval-2019 task 6: Supervised vs. unsupervised transfer learning for offensive language detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

[159] C. Wu, F. Wu, S. Wu, J. Liu, Z. Yuan, and Y. Huang. 2018. THU_NGN at SemEval-2018 task 3: Tweet irony detection with densely connected LSTM and multi-task learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. New.

[160] H. Wu, Z. Wang, F. Qing, and S. Li. 2021. Reinforced transformer with cross-lingual distillation for cross-lingual aspect sentiment classification. *Electronics* 10 (2021), 270.

[161] Z. Wu and D.C. Ong. 2020. Context-guided BERT for targeted aspect-based sentiment analysis. In *Association for the Advancement of Artificial Intelligence*. 1–9.

[162] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q.V. Le. 2019. Unsupervised data augmentation. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*. Vancouver, Canada.

[163] C. Xing, D. Wang, C. Liu, and Y. Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[164] H. Xu, B. Liu, L. Shu, and P.S. Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. (2019).

[165] W. Xue and T. Li. 2018. Aspect based sentiment analysis with gated convolutional networks. (2018).

[166] H. Yang, B. Zeng, M. Xu, and T. Wang. 2021. Back to Reality: Leveraging Pattern-driven Modeling to Enable Affordable Sentiment Dependency Learning. (2021).

[167] H. Yang, B. Zeng, J. Yang, Y. Song, and R. Xu. 2021. A multi-task learning model for Chinese-oriented aspect polarity classification and aspect term extraction. *Neurocomputing* 419 (2021), 344–356.

[168] M. Yang, W. Yin, Q. Qu, W. Tu, Y. Shen, and X. Chen. 2019. Neural attentive network for cross-domain aspect-level sentiment classification. *IEEE Transactions on Affective Computing* (2019).

[169] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. Vancouver, Canada.

[170] H. Ye, Q. Tan, R. He, J. Li, H. T. Ng, and L. Bing. 2020. Feature adaptation of pre-trained language models across languages and domains with robust self-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online*.

[171] Inc. Yelp. 2022. Yelp dataset. (Mar. 2022). https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset.

[172] Y. Yin, Y. Song, and M. Zhang. 2017. Document-level multi-aspect sentiment classification as machine comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

[173] Z. Yu and G. Liu. 2018. Sliced recurrent neural networks. (2018).
[174] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark.
[175] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and ÃĞ. ÃĞÃűltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona (online)*.
[176] K. Zhang, H. Zhang, Q. Liu, H. Zhao, H. Zhu, and E. Chen. 2019. Interactive Attention Transfer Network for Cross-domain Sentiment Classification. (2019).
[177] S. Zhang, L. Yao, A. Sun, and Y. Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)* 52 (2019), 1–38.
[178] S. Zhang, X. Zhang, C. Jeffrey, and P. Rosso. 2019. Irony detection via sentiment-based transfer learning. *Information Processing & Management* 56, 5 (2019), 1633–1644.
[179] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam. 2022. A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. (2022).
[180] X. Zhang, S. Huang, J. Zhao, X. Du, and F. He. 2018. Exploring deep recurrent convolution neural networks for subjectivity classification. *IEEE Access* 7 (2018), 347–357.
[181] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. 2019. ERNIE: Enhanced language representation with informative entities. (2019).
[182] P. Zhao, L. Houb, and O. Wua. 2019. Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification. *Knowledge-Based Systems* 193 (2019).
[183] P. Zhu and T. Qian. 2018. Enhanced aspect level sentiment classification with auxiliary memory. In *Proceedings of the 27th International Conference on Computational Linguistics*.
[184] W. Y. Zou, R. Socher, D. Cer, and C. D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Received 21 April 2021; revised 18 May 2022; accepted 27 June 2022