Chapter 5
Deep Learning for Sentiment Analysis:
An Overview and Perspectives

Vincent Karas
University of Augsburg, Germany

Björn W. Schuller
University of Augsburg, Germany

ABSTRACT

Sentiment analysis is an important area of natural language processing that can help inform business decisions by extracting sentiment information from documents. The purpose of this chapter is to introduce the reader to selected concepts and methods of deep learning and show how deep models can be used to increase performance in sentiment analysis. It discusses the latest advances in the field and covers topics including traditional sentiment analysis approaches, the fundamentals of sentence modelling, popular neural network architectures, autoencoders, attention modelling, transformers, data augmentation methods, the benefits of transfer learning, the potential of adversarial networks, and perspectives on explainable AI. The authors' intent is that through this chapter, the reader can gain an understanding of recent developments in this area as well as current trends and potentials for future research.

INTRODUCTION

In recent years, the amount of information available on the Internet has grown rapidly. At the beginning of 2019, Twitter had 326 million monthly active users, and 500 million tweets were sent per day Cooper (2019). Facebook, the largest social media platform, reported 2.41 billion monthly active users for the second quarter of 2019 Facebook (2019). Every minute, 4.5 million YouTube videos and 1 million Twitch videos are viewed, and the Google search engine processes 3.8 million queries (Desjardins, 2019). This trove of online content constitutes a valuable resource for business applications, e.g. for providing the users with personalised search recommendations and tailored advertisements. If the data is harnessed

DOI: 10.4018/978-1-7998-4240-8.ch005

properly, it may deliver new insights that can help improve existing products and services and inspire future business models. Among the available content, text, in particular, is rich in information, as it can contain nuanced emotions, multiple layers of meaning and ambiguities. However, this complexity also results in it being challenging to analyse. Natural Language Processing (NLP), which addresses this challenge, has become a popular field of research.

Sentiment Analysis (SA), which is often also referred to as opinion mining or comment mining in the literature, is a discipline of NLP-based text analysis whose goal is to determine the writer's feelings about a particular topic. Emotions have been shown to play an essential role in human decision making (Bechara, Damasio, & Damasio, 2000) and behaviour in general. Consequentially, SA has many conceivable applications in business and academia. Examples include companies looking to improve their services by automatically assessing customer reviews (Hu & Liu, 2004), (Zvarevashe & Olugbara, 2018), comparing products online, or analysing newspaper headlines (Rameshbhai & Paulose, 2019).

Sentiment also plays an important role in the financial market. Ranjit, Shrestha, Subedi, and Shakya (2018) used SA to predict the exchange rates of foreign currencies. Shah, Isah, and Zulkernine (2018) predicted stock prices in the pharmaceutical industry based on the sentiment in news coverage. C. Du, Tsai, and Wang (2019) classified financial reports in terms of expected financial risk using SA.

In addition, there are medical applications for SA. Müller and Salathé (2019) introduced an open platform for tracking health trends on social media. Luo, Zimet, and Shah (2019) created an NLP framework to investigate sentiment fluctuation on the subject of HPV vaccination, expressed by Twitter users between 2008 and 2017.

Furthermore, political analysts and campaigns can benefit from mining the opinions and emotions expressed towards candidates, issues and parties on social media. Jose and Chooralil (2016) used an ensemble classifier approach to predict results of the 2015 election in Delhi. Joyce and Deng (2017) applied SA to tweets collected in the run-up to the 2016 US presidential election and compared them to polling data. They found that automatic labelling of tweets outperformed manual labelling.

Many tools used in sentiment analysis are designed for a specific application, which negatively impacts their diffusion. Joshi and Simon (2018) introduced a cloud-based open-source tool which provides various APIs in order to perform SA on data from arbitrary sources.

While SA has attracted considerable attention, the field still faces challenges. These include domain dependence, negations, handling fake reviews (Hussein, Doaa Mohey El-Din Mohamed, 2018), as well as incorporating context, dealing with data imbalance and ensuring high-quality annotations (Boaz Shmueli & Lun-Wei Ku, 2019).

This chapter introduces the reader to selected methods used for sentiment analysis, with a focus on techniques based on deep learning. Its contribution consists of a discussion of the latest advances in the state of the art, as well as an outlook concerning ongoing trends in the field and recommendations on future research directions.

The rest of the chapter is structured as follows. In the next section, the fundamentals of SA and select machine learning concepts are presented. Topics covered include a categorisation of analysis approaches by level of granularity, how to measure sentiment, traditional sentiment analysis methods employing lexica and machine learning, as well as tools for word embedding and sentence modelling such as autoencoders, GloVe, fastText and Word2vec. The chapter will then continue with its main section, focusing on current developments in deep learning-based SA. Topics include popular neural network architectures and their combination into hybrid models, capturing contextual information by adding attention, Transformer networks and the challenges and benefits of transfer learning. In the following section, solutions and recommendations for readers seeking to apply state-of-the-art models to SA are presented. The subsequent section involves an overview of promising research opportunities in the field. Recent data augmentation techniques, zero-shot learning and the potential of generative adversarial networks are covered. In addition, the need for developing explainable AI systems is discussed as well as improving generalisation across topics and languages and defending against adversarial attacks. Finally, a conclusion sums up this chapter.

BACKGROUND

This section presents a taxonomy of sentiment analysis and key methods and frameworks used for sentence modelling and generating word embeddings.

Levels of Sentiment Analysis

A text can be analysed for its sentiment content at different levels. These are document, sentence and phrase levels (P. Balaji, O. Nagaraju, & D. Haritha, 2017). Sentiment analysis at phrase level is also commonly referred to as aspect level analysis, a name that will be adopted for this chapter. The level of analysis informs the choice of deep learning models.

As a motivational example, consider an automotive company wanting to classify product reviews of their cars. A review might read as follows:

"This is a great car. It handles well in corners and has superb acceleration. Like its predecessor, it has a V6 engine. However, I do not like what they did with the new voice-controlled infotainment system. It gets confused too easily to be useful."

The following subsections illustrate the application of SA at different levels based on the example review:

Document Level SA

The task at this level is to classify the entire document as having a positive or negative sentiment (Pang, Lee, & Vaithyanathan, 2002). Such an analysis can serve to determine a general verdict, e.g. to find out whether a reviewer likes or dislikes a product. Therefore, this approach can work only if the document describes a single issue.

For the example review, it appears that the customer has an overall positive opinion. However, there is also criticism. In order to understand the positive and negative feelings expressed by the customer, the document needs to be examined in greater detail.

Sentence Level SA

At this level, individual sentences are examined for their sentiment content. This approach requires splitting the document into objective sentences, which contain factual information, and subjective sentences that reflect opinions and feelings. The classification of subjectivity was investigated by Wiebe, Bruce, and O'Hara (1999). Subjective sentences are then subjected to SA and rated accordingly. Performing

SA at sentence level makes a similar assumption to document level SA in that individual sentences are referring to only one entity, which will often not be true (Christy Daniel & Shyamala, 2019).

Considering our sample review at the sentence level, a more detailed picture emerges: The customer expresses positive sentiment in the first two sentences. The third sentence is a factual statement. The last two sentences show negative sentiment.

Aspect Level SA

Aspect-level analysis examines individual entities within sentences, making it more fine-grained than the previous approaches. It can discover in detail which elements of a topic are liked or disliked, which is useful since the author's opinion on a subject will rarely be entirely positive or negative. Thus, the objective of an aspect-level analysis is to discover the slant of the text (P. Balaji et al., 2017). Multiple sub-tasks can be defined at this level:

1. **Target extraction:** This identifies the entities that sentiments refer to.
2. **Sentiment classification:** The rating of the sentiment.
3. **Temporal opinion mining:** This task is concerned with discovering the temporal relationships in the text and how those affect the evolution of sentiment.
4. **Opinion holder identification:** A text may reference different persons, each having individual opinions.

For the example review, a targeted aspect-level analysis can reveal that the customer approves of the car's driving characteristics, as they commend the acceleration and handling in corners. At the same time, the customer disapproves of a new feature in the infotainment system. For the manufacturer, this is valuable information for identifying the strengths and weaknesses of the product. Opinion holder identification is also quite useful for this task. The example review features a single customer who emphasises a good driving experience, but there could also be references to, e.g. family members having different priorities.

Now that the basic approaches for extracting sentiment from documents have been identified, the next subsection will address the question of how sentiment can be quantified.

Measuring Sentiment

Just as the analysis of a document may be performed at different levels, the discovered sentiment may also be measured at different levels of granularity. One possibility is a binary approach based on polarity, i.e. the text is positive or negative. A neutral state may be added as a third class. Alternatively, categorical emotions may be used. Ekman (1999) identified six basic emotions, namely, happiness, anger, sadness, disgust, surprise and fear. A more fine-grained description is provided by continuous affect dimensions such as valence, arousal, dominance, or novelty. Plutchik (1980) introduced a model which combines elements of the categorical and continuous approaches. It encompasses eight types of emotions, namely joy, anticipation, trust, surprise, fear, anger, disgust and sadness. The emotions are arranged as opposing pairs in a wheel. In addition, each emotion can appear at different levels of intensity, e.g. trust ranges from acceptance to admiration.

Consider the sentences “This car is all right.” and “This car is great.” Both express a positive sentiment, but it is much stronger in the second sentence, which should result in a higher level of valence being detected.

Having introduced levels of granularity and ways to measure sentiment, the next section will explore algorithms traditionally used in SA:

Traditional Approaches for Sentiment Analysis

The methods used for SA can be placed into two broad categories: lexica-based approaches and machine learning approaches. This chapter considers deep learning-based algorithms separately in the following section; therefore, they are not discussed among the machine learning algorithms in this section.

Lexicon-Based Approach

The lexicon-based approach aggregates the polarity and strength of individual words in the document to calculate the overall sentiment. (Turney, 2002). It requires a dictionary of words with associated semantic orientation. The research into lexica-based SA has largely focused on adjectives, cf. Hatzivassiloglou and McKeown (1997), Hu and Liu (2004), Wiebe (2000) and Taboada, Anthony, and Voll (2006).

The dictionary or lexicon can be compiled manually (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011) or automatically starting from a seed list of opinion words. Automatic lexicon compilation is accomplished with thesaurus-based and corpus-based methods. Thesaurus-based methods expand the seed list by parsing existing dictionaries for synonyms or antonyms, while corpus-based methods exploit statistical co-occurrence of words with similar polarity in a corpus, or calculate similarity measures between words (Kaur, Mangat, & Nidhi, 2017).

Machine Learning-Based Approach

Machine learning algorithms perform sentiment classification or regression according to features contained in the text. They can – among many possible discriminations – be divided into linear classifiers and probabilistic classifiers.

Support Vector Machine (SVM) is an example of an (in principle) linear classifier, i.e. it attempts to predict a label y (+1 or -1) from features x based on the function:

$$y = f(x) = w^T x + b \qquad (1)$$

It was developed within the statistical learning theory (Vapnik, 2000). The algorithm searches a hypothesis space of functions in order to find a hyperplane that separates classes. In the simple case, considered up to now, of a linear SVM, the hyperplane lies in the input space. In the generalised form of SVM, a dot product called a kernel is used to define a Reproducing Kernel Hilbert Space as the feature space (Evgeniou & Pontil, 2001). SVM attempts to maximise the distance between the named hyperplane and the instances of each of the two classes (extensions for more than two classes exist, such as one vs one, or one vs all).

Naïve Bayes (NB) is an example of a probabilistic classifier, which predicts a conditional probability $p(y|x)$. Naïve Bayes uses Bayes' rule to determine the probability of a class c belonging to a vector of BoW features x:

$$p(c \mid x) = \frac{p(c)\,p(x \mid c)}{p(x)} \tag{2}$$

This simple algorithm assumes that the features are conditionally independent, which allows it to decompose the numerator (Pang et al., 2002). The classifier then takes the form:

$$p_{NB}(c \mid x_1, \cdots, x_n) = \frac{p(c)\prod_{i=1}^{n} p(x_i \mid c)}{p(x)} \tag{3}$$

Maximum Entropy (ME) is another probabilistic algorithm. It has been used for SA of tweets (Neethu & Rajasree, 2013), (Gautam & Yadav, 2014). The intuitive assumption of this classifier is that the underlying probability distribution should have maximum entropy, i.e. be as uniform as possible within the constraints imposed by the training data (Nigam, Lafferty, & Mccallum, 1999). Those constraints apply to the feature functions $f_i\,(d,\ c)$, whose expected value within the model and the training data are demanded to be equal. The probability distribution takes an exponential form (Della Pietra, Della Pietra, & Lafferty, 1997):

$$p_{ME}(c \mid d) = \frac{1}{Z_d} e^{\sum_{i=1}^{n} \lambda_i f_i(d,c)} \tag{4}$$

Here Z_d is a normalisation factor, and λ_i is a parameter to be estimated. Unlike NB, ME does not assume independence of features, and therefore, it can outperform NB on tasks where that assumption does not hold (Pang et al., 2002).

Sentence Modelling and Word Embeddings

In order to perform SA on a document, the text first has to be converted into a form that the SA algorithm can process. This is done by assigning a vector to each word in the document. A simple solution would be to use an approach known as Bag-of-Words (BoW). The number of occurrences of each unique word within the corpus is determined and used to sort the words in descending order. Then, a one-hot encoding can be applied to that list of words. The same approach can be used with n-grams (word sequences of length n).

A naïve BoW, as described above, is easy to implement but has several disadvantages. First, it can result in very high-dimensional representations, up to the number of entries in the vocabulary. Second, such an encoding does not capture the linguistic relationships between words. However, the goal of sentence modelling should be to obtain feature representations which guarantee that the similarity between two vectors reflects the semantic and syntactic relationship between the corresponding words.

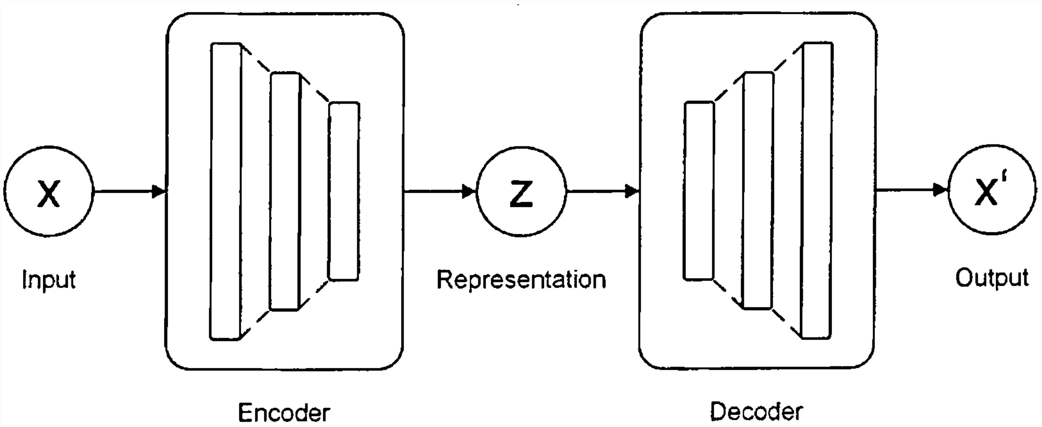
The following subsections introduce a selection of established methods and tools that can be used to discover useful representations for NLP tasks, including but not limited to SA.

Autoencoders

A useful representation should capture the relevant information contained in the raw data, allow for clustering into categories, and reduce the number of features sufficiently to avoid the curse of dimensionality. An example of a deep architecture designed to learn such representations in an unsupervised manner is the autoencoder.

An autoencoder (AE) consists of two networks connected in sequence: The encoder processes the input data and generates a feature vector at its output layer. That vector is usually of lower dimensionality than the input; however, a variant called sparse autoencoder may increase the dimensionality of the encoder output but regularise it to produce sparse activations. The features generated by the encoder are used as the input to the decoder, which produces an output of the same shape as the input data. The autoencoder is trained by setting the target of the decoder to be the same as the input data. Since the layer in the middle of the network has fewer parameters, it acts as a bottleneck, forcing the network to learn how to compress the input into a compact representation. This process can be called self-supervised, as the autoencoder learns by optimising the reconstructing error of the data without a need for labels. Figure 1 illustrates the basic structure of an autoencoder.

Figure 1. Autoencoder architecture. Input is compressed by the encoder, then reconstructed by the decoder. This forces the network to learn an efficient representation of the data.



AEs and their variants are popular tools for sentiment analysis. They are frequently employed in semi-supervised strategies when only part of the data is labelled. The variational autoencoder (VAE) learns latent representations in a probabilistic manner (Kingma & Welling, 2013). Examples from the literature that utilise VAEs for SA include aspect-level classification of user reviews (Fu et al., 2019), multi-task learning for improved generalisation (Lu, Zhao, Yin, Yang, & Li, 2018) and a semi-supervised variant that makes use of the labels in the decoder to boost accuracy (W. Xu, Sun, Deng, & Tan, 2017). Winner-take-all autoencoders (Makhzani & Frey, 2015) enforce sparsity by having the neurons in the embedding layer compete for contributing to the output. Maitra and Sarkhel (2018) used a shallow winner-take-all autoencoder to classify social media texts in multiple languages as overtly, covertly or non-aggressive.

Denoising autoencoders (DA) make the representation more robust by corrupting the input with noise and learning to reconstruct a clean version. A stacked denoising autoencoder (SDA) combines multiple denoising autoencoders, with the latent representation of one AE acting as input to the next one (Vincent, Larochelle, Bengio, & Manzagol, 2008). This allows for learning potent representations while keeping the number of parameters small, saving computational resources and reducing the amount of training data needed to prevent overfitting. During training, the layers are tuned one by one. (Sagha, Cummins, & Schuller, 2017).

Conventional autoencoders have recently become less relevant for generating word embeddings, as NLP researchers increasingly favour the new Transformer networks, which are discussed separately in this chapter's main section on deep learning. The following subsections present several popular open-source frameworks that provide pre-trained word embeddings. For tasks that involve small datasets, pre-trained embeddings learned on large corpora can help mitigate the problem of encountering unseen words at test time (Hsu & Ku, 2018).

Word2vec

Word2vec[1] was introduced by Mikolov, Chen, Corrado, and Dean (2013). It is an extension of the continuous Skip-gram model developed by Mikolov, Sutskever, Chen, Corrado, and Dean (2013). Skip-gram is a log-linear model; the choice of linearity is motivated by training efficiency being valued over additional complexity in the representations. It analyses a sequence of training words T and for each word w, attempts to predict both previous and subsequent words within a context c. The objective that Skip-Gram attempts to maximise is an average log probability given by:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,j\neq 0}\log p\left(w_{t+j}\mid w_t\right) \tag{5}$$

The conditional probability is formulated as a softmax function, which makes the computation of the gradient inefficient for large vocabularies. Word2vec extends Skip-gram by optimising the algorithm, which allows training on larger corpora. This is done by simplifying the softmax function, as well as discarding frequently occurring words that carry little information, e.g. function and conjunction words such as "the" and "and". The authors of Word2vec also demonstrated that phrases can be encoded by the model and that the linear properties of the learned word vectors allow reasoning based on simple arithmetic. For example, the representation of the word "queen" could be found by the following expression:

$$v_{queen} = v_{king} - v_{man} + v_{woman} \tag{6}$$

GloVe

GloVe[2] (Global Vectors) was derived by Pennington, Socher, and Manning (2014). The name reflects that global statistics of a corpus are captured. It is a log-bilinear model with a weighted least-squares objective function for unsupervised learning of word representations. The objective function that GloVe attempts to minimise is given by:

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + b_j - \log X_{ij}\right) \tag{7}$$

GloVe operates on word co-occurrence counts, i.e. on a matrix X whose entries show how many times a word appears in the context of other words. The set of all words together forms the vocabulary V. The word learning of GloVe is based on the ratios of word co-occurrence probabilities, which compared to the raw probabilities are better at distinguishing relevant words (Pennington et al., 2014).

fastText

fastText[3] is an open-source library for text representation learning and text classifier learning provided by Facebook AI Research. It is based on the works of Bojanowski, Grave, Joulin, and Mikolov (2017) and Joulin, Grave, Bojanowski, and Mikolov (2017). In fastText, instead of assigning a fixed vector to each word, words are modelled as bags of character n-grams. For text classification, simple linear models are used, whose performance on SA tasks has been shown to be comparable with deep architectures while being lightweight and faster to train.

THE CURRENT STATE OF DEEP LEARNING-BASED SA

In this section, a number of key concepts and methods for deep learning are presented.

Advantages and Applications of Deep Learning

Deep learning is a popular form of machine learning that has allowed researchers to achieve breakthroughs in many fields, including computer vision (Krizhevsky, Sutskever, & Hinton, 2012) and speech recognition (Hinton et al., 2012). This part of the chapter will introduce key concepts of deep learning.

Deep learning is based on deep neural networks, i.e. models which contain hidden layers. This multilayered architecture allows deep models to overcome a shortcoming of conventional machine learning algorithms such as SVM, which is the requirement of feature engineering. Those algorithms needed a suitable feature extractor to turn raw data into representations they could learn from, which required considerable expertise and effort from the researcher (LeCun, Bengio, & Hinton, 2015).

On the other hand, deep models can adjust their internal states to find appropriate representations without the need for extensive preprocessing of the data. They are capable of learning advanced concepts through a stack of modules connected by nonlinear functions. Each module processes the features extracted by the previous ones, which leads to the development of increasingly complex representations. Bengio, Courville, and Vincent (2013) provide an in-depth discussion of desirable properties of representations and how various deep learning methods can be leveraged for representation learning.

Common Network Architectures

Models based on deep learning have the capability of detecting intricate patterns in data and continue to produce state of the art results in many fields. The following subsections introduce important architectures and techniques and examples of their application to sentiment analysis.

Recurrent Neural Network (RNN)

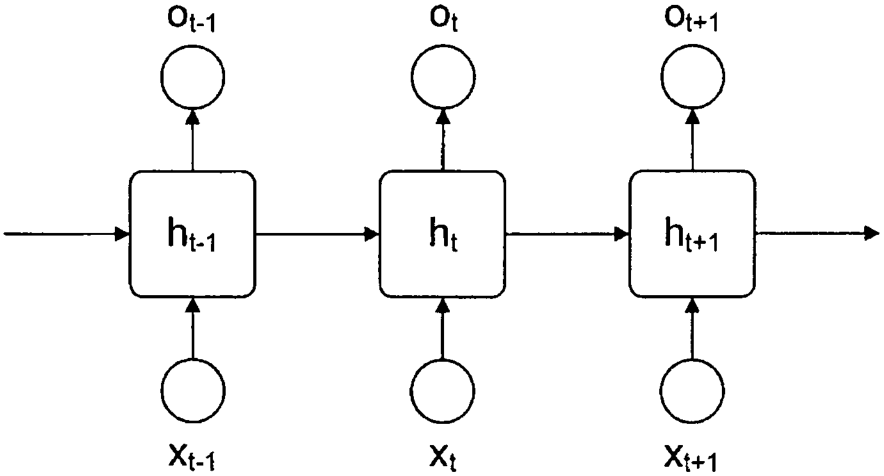
Recurrent Neural Networks are capable of processing sequential inputs, which makes them attractive for handling data of varying length, e.g. speech or text. An RNN makes use of its hidden units to maintain a state vector, which stores information on the previous elements in the input sequence (LeCun et al., 2015). Thus, the RNN can remember the inputs it has seen. The network can be unfolded along the temporal dimension, effectively making it a deep feedforward architecture, with each unit processing one element in the input sequence and generating an output and a state, which feeds into the next unit. The equations for an RNN are as follows:

$$h_t = \sigma\left(U^h x_t + W^h h_{t-1} + b^h\right), \quad (8)$$

$$o_t = softmax(W^o h_t + b^o) \quad (9)$$

With *h, x, o* being the hidden state, input and output respectively and subscripts denoting the time step. *U* and *W* are parameter matrices, and *b* are bias vectors. An illustration of an unfolded RNN can be seen in Figure 2.

Figure 2. Unfolded RNN architecture. Data is processed sequentially, with the hidden state being propagated through time.



In some cases, it can be desirable to use both past and future information contained in a sequence. Bidirectional RNNs achieve this by combining two RNNs, with each net reading the sequence in a different direction. They have been extensively used in NLP, including in sentiment analysis (Tian, Rong, Shi, Liu, & Xiong, 2018).

Plain RNNs suffer from a common problem in training deep architectures with backpropagation, which is that gradients either tend to zero or become very large across many layers. These effects are known as the vanishing gradient problem and exploding gradient problem, respectively. They make it difficult to learn relationships across large time intervals.

To address this problem, Hochreiter and Schmidhuber (1997) proposed an RNN variant called Long Short-Term Memory (LSTM). In this architecture, the standard recurrent cells in the hidden layers are replaced with memory blocks designed to maintain information (Graves, 2012). The original LSTM is built around a self-recurrent internal structure called a constant error carousel (CEC), which prevents the error from vanishing. Furthermore, it uses two multiplicative gates to regulate its connections: the input gate restricts information entering the cell, and the output gate controls information leaving the cell. Gers, Schmidhuber, and Cummins (2000) improved the LSTM by adding a third gate, named forget gate, in place of the fixed CEC connection. This allows the network to reset its previously learned state, which solves the problem of internal states growing too large over long sequences. The LSTM cell can now be described by the following equations:

$$i_t = \sigma\left(W^i x_t + U^i h_{t-1} + b^i\right), \tag{10}$$

$$f_t = \sigma\left(W^f x_t + U^f h_{t-1} + b^f\right), \tag{11}$$

$$o_t = \sigma\left(W^o x_t + U^o h_{t-1} + b^o\right), \tag{12}$$

$$g_t = \tanh\left(W^g x_t + U^g h_{t-1} + b^g\right), \tag{13}$$

$$c_t = f_t \bullet c_{t-1} + i_t \bullet g_t, \tag{14}$$

$$h_t = o_t \bullet \tan h(c_t) \tag{15}$$

Here c_t is the cell state at time t.

Cho, van Merriënboer, Bahdanau, and Bengio (2014) introduced the Gated Recurrent Network (GRU), which simplifies the LSTM cells. The hidden cells contain two gates: a reset gate which makes the cell forget its hidden state and replace it with the current input, and an update gate which controls the contribution of the previous hidden state to the next time step.

An example of the application of RNNs to sentiment analysis is the work of D. Tang, Qin, and Liu (2015). They performed document-level SA on four large datasets containing IMDB and Yelp reviews, using two gated RNN models with adaptive sentence modelling.

Convolutional Neural Network (CNN)

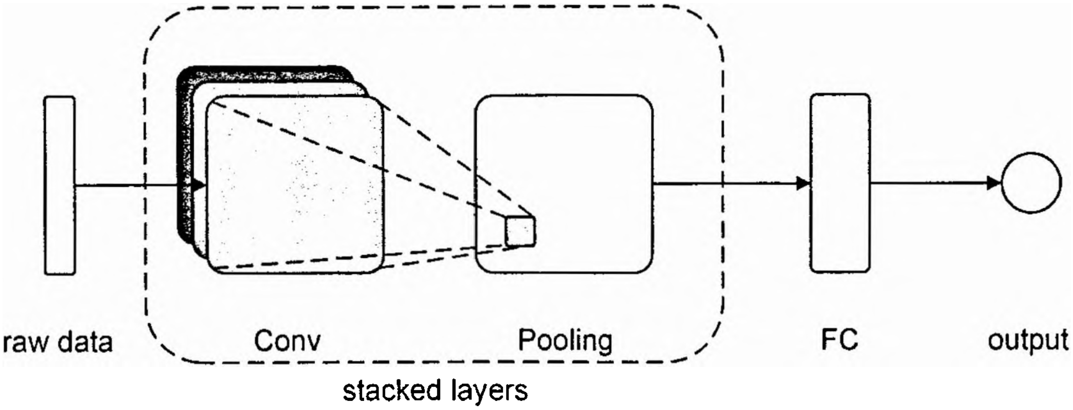
Convolutional Neural Networks process data in the form of arrays (e.g. videos, images, audio spectrograms and word embeddings) through multiple layers that extract hierarchical features. This is achieved by a combination of convolutional layers and pooling layers.

A convolutional layer makes use of arrays of weights called filter banks. A filter slides across the input data, computing a weighted sum at each position. This results in a new array called a feature map, whose size can be adjusted by zero-padding the input data or changing the filter dimensions and stride. A convolutional layer can construct multiple feature maps by applying different filters. The results are passed through a nonlinear activation, e.g. a (potentially "leaky") rectified linear unit (ReLU). The idea behind the use of these filters is to detect certain features in the input data by matching it to the pattern specified by the filter. For a visual recognition system, those features could be simple lines or edges in the first layers, which are then combined to form objects of increasing complexity. The name convolutional layer is due to the fact that the sliding filter effectively performs a discrete convolution of the input.

Pooling layers merge the information contained in neighbouring cells of a feature map. Implementations of CNNs commonly use max-pooling layers, which will retain only the maximum value of the features in a patch, resulting in a smaller map. Pooling has the advantage of reducing the dimension of the internal representations, as well as introducing an invariance to small shifts and distortions (LeCun et al., 2015).

In addition to sequences of convolutional layers, nonlinearities and pooling for feature extraction, CNNs also incorporate fully connected layers to combine the features for classification. The complete network can be trained through backpropagation. Figure 3 illustrates an example of a CNN architecture.

Figure 3. CNN architecture. A stack of convolutional and pooling layers is used to extract features, which are combined by fully connected layers for classification.



The breakthrough of CNNs came in the field of computer vision in 2012, when a model by Krizhevsky et al. (2012) won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) with a top-5 test error rate of 15.3%, which was more than 10% ahead of the second-best entry. CNNs can also be applied to sentiment analysis of text. Kim (2014) showed that a simple CNN which processed embeddings generated by word2vec could perform very well in sentence classification, even improving upon the state of the art at the time.

Hybrid Network

A hybrid network includes components from multiple basic neural network architectures. An example of this is combinations of convolutional and recurrent nets (C-RNNs). As shown previously, CNNs are useful for feature extraction in a hierarchical manner, while RNNs are well suited for processing sequential data and capturing important aspects in memory. A C-RNN allows for the combination of these advantages by processing word embeddings through convolutions and feeding the resulting features to a recurrent network.

Hybrid models are a widely used technique in SA. X. Wang, Jiang, and Luo (2016) performed SA on short texts using combinations of word2vec and randomly initialised word vectors and CNN-GRU/CNN-LSTM models, finding that the joint architecture outperformed CNN and RNN alone. More recently, Hassan and Mahmood (2018) proposed a C-RNN architecture that uses recurrent layers instead of pooling layers in order to overcome the problem of CNNs extracting features locally at each stage and thus needing to be very deep to capture long-term dependencies.

The previously discussed methods can be enhanced through a concept called attention, which will be introduced next.

Capturing Context Through Attention

When sentiment analysis is performed on a text, some words will matter more than others. To determine the sentiment towards a certain target requires knowing the context, i.e., relevant words in the rest of the sequence. When an encoder attempts to model those relationships implicitly, as, e.g., RNNs do when compressing the entire input sequence into a fixed-length representation vector, this can lead to problems with long-term dependencies in very long texts. What is needed is a way for the network to learn how to focus on specific elements of the input, as a human reader would do. This is achieved through the attention mechanism.

Attention was first proposed by Bahdanau, Cho, and Bengio (2014), who used it for the purpose of neural machine translation. A common approach to that task is to use an encoder-decoder structure, with the encoder creating a high-level representation of the input sentence and the decoder turning it into an output sentence in a different language. This model was expanded by an attention component which taught it how to align certain words in the input and output sequences, leading to improved performance in English-French translation.

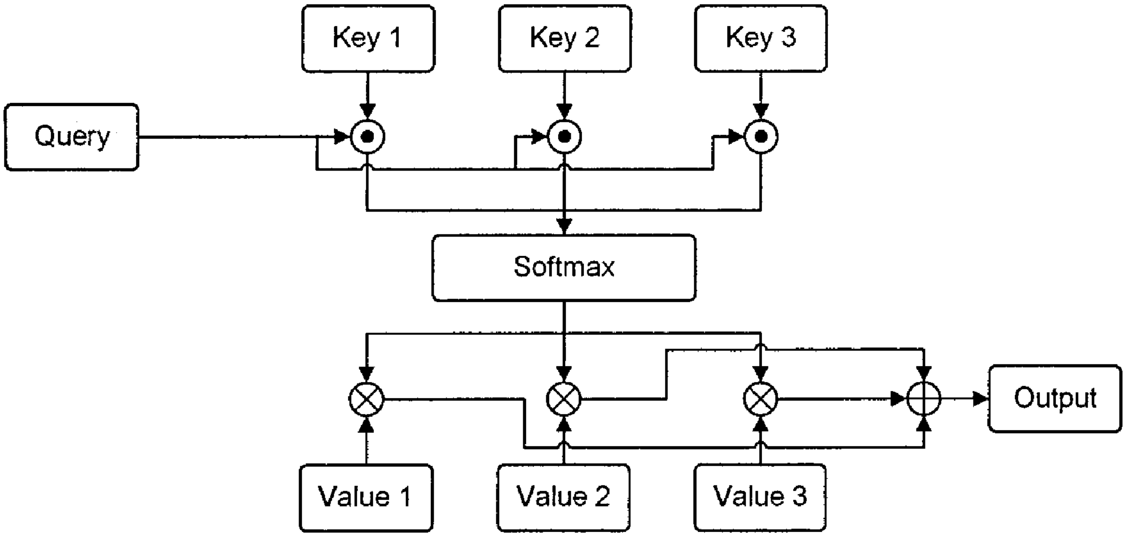
A general way of describing attention is as a function that takes a query Q and a set of key-value pairs (K_i, V_i) and computes a weighted sum of the values based on a comparison between the query and the keys (Vaswani et al., 2017). Thus, assuming an input sequence of hidden states (h_1,...,h_T) as the keys, a context vector c_i is computed by:

$$\alpha_{ij} = \frac{\quad\quad)}{\sum_{k=1}^{T}\exp\left(e_{ik}\right)}, \tag{16}$$

$$c_i = \sum_{j=1}^{T} \alpha_{ij} h_i \tag{17}$$

Here, e_{ij} is an alignment model that functions as a measure of similarity between the query and a key. It is used to compute the weight α_{ij} of each value through a softmax function, and the context vector is the sum of those contributions (Bahdanau et al., 2014). Figure 4 illustrates the concept of attention.

Figure 4. Dot product attention. The dot product is used as a similarity measure between query and keys. A softmax function computes the attention weights of the values, which are then summed into the output.



Attention has become a popular method in sentiment analysis. Works that use attention for aspect-level SA include Q. Liu, Zhang, Zeng, Huang, and Wu (2018), Chen, Sun, Bing, and Yang (2017), and D. Tang, Qin, and Liu (2016). It has been combined with RNNs (Ran, 2019), (G. Liu & Guo, 2019), CNNs (J. Du, Gui, Xu, & He, 2018), (Wu, Cai, Li, Xu, & Leung, 2018) and employed in hybrid networks (Zhu, Gao, Zhang, Liu, & Zhang, 2018). Deng, Jing, Yu, and Sun (2019) used an LSTM with sparse self-attention to construct a sentiment lexicon.

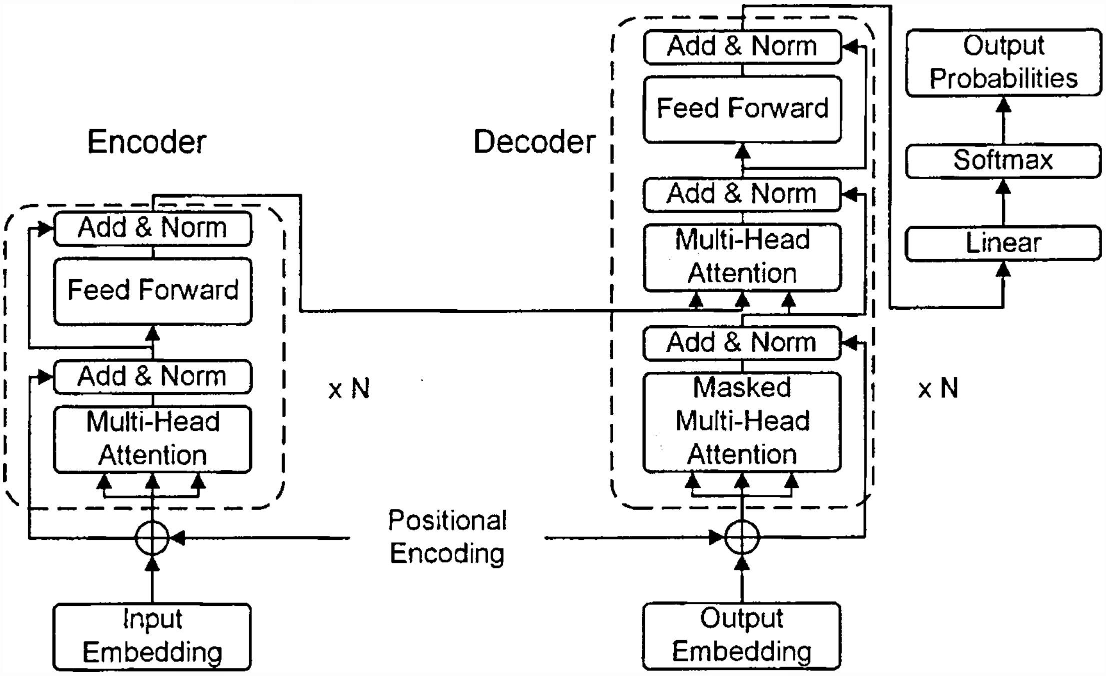
Zichao Yang et al. (2016) proposed a hierarchical attention network (HAN) for document classification that applied attention at word and sentence level. Z. Liu et al. (2019) used HAN for sentence representation learning. N. Xu (2017) combined a text HAN with an image HAN for public sentiment classification. Another work by Niu and Hou (2017) used hierarchical attention with bidirectional LSTM for text modelling. Stappen et al. (2019) employed HAN for detecting sentiment change in transcripts of interviews.

A significant development in the fields of SA and NLP in general that has been enabled by attention was the invention and subsequent popularisation of Transformer networks.

Transformer Networks

Vaswani et al. (2017) introduced a novel type of networks known as Transformers, which do not require recurrent or convolutional layers. Instead, those networks rely on self-attention, i.e. computing attention between all the elements in the input sequence, and make use of multiple structures called attention heads for fine-grained analysis (G. Tang, Müller, Rios, & Sennrich, 2018). The architecture of a Transformer is illustrated in Figure 5, based on Vaswani et al. (2017). The Transformer consists of an encoder and a decoder block, followed by a linear layer and a softmax layer. The encoder and decoder are composed of N blocks, with each block containing multi-head attention and a feedforward network, as well as residual connections and layer normalisation. Positional encoding is added to the input and output embeddings to allow the model to understand word order.

Figure 5. Transformer architecture



Recently, the work by Devlin, Chang, Lee, and Toutanova (2018) on transformers has led to a major breakthrough in NLP. They introduced a framework named Bidirectional Encoder Representations from Transformers (BERT). BERT involves two steps:

1. **Pre-training:** BERT is pre-trained on a document-level corpus using unsupervised learning on two tasks: A Masked Language Model (MLM) randomly masks input tokens in order to teach BERT to predict words based on their context. In addition, Next Sentence Prediction (NSP) is used to learn the relationships between sentences.

2. **Fine-tuning**: The pre-training is used to initialise models for the downstream tasks that BERT should solve. Each model is then fine-tuned separately through end-to-end learning with task-specific data.

The main contribution of BERT is its improvement upon previous unsupervised representation learning methods by using a bidirectional architecture to generate more powerful representations. Models created with BERT showed excellent performance, surpassing the state of the art in eleven NLP tasks by wide margins, including a 7.7% improvement on the GLUE, a benchmark task for natural language understanding (A. Wang et al., 2018). This has led to great popularity of this type of models in the NLP research community.

Among many other applications, Transformers have also been used for sentiment analysis. Q. Zhang, Lu, Wang, Zhu, and Liu (2019) introduced interactive multi-head attention (IMAN) pre-trained on BERT to achieve new state of the art results in aspect-level SA. Jiang, Wu, Shi, and Zhang (2019) proposed a Transformer-based memory network (TF-MN) for sentiment-based Q&A. Cheng et al. (2019) introduced a VAE framework which uses Transformers as encoder and decoder. Gao, Feng, Song, and Wu (2019) used BERT for targeted sentiment classification.

Adapting With Transfer Learning

As shown in the introduction to this chapter, sentiment analysis has many academic and business applications, but still faces challenges, including domain dependence. While deep learning-based methods have been shown to achieve state-of-the-art results, they require a considerable amount of data for training. A common scenario is that one wants to apply a deep learning approach to a specific setting, but it is not feasible to collect and label enough data to train a model. However, there is a large, labelled dataset from a different setting available. As an example of this problem consider the classification of product reviews depending on sentiment. Given the wide variety of products available, it would probably be prohibitively expensive to gather and label a sufficiently large amount of data to separately train a classifier for each product. Instead, it would be preferable to make use of existing reviews for other products. Simply applying a model trained on that data to the new problem will likely yield worse performance, since the same words may have different meaning or polarity depending on the subject of the text (Save & Shekokar, 2017). Because of these issues, a new research field has emerged that combines SA methods with transfer learning (R. Liu, Shi, Ji, & Jia, 2019).

Definitions of Transfer Learning

In their survey paper, Pan and Yang (2010) present a useful categorisation of transfer learning and its relation to other fields. They define machine learning problems in terms of domains D and tasks T. A domain D consists of a feature space spanning all possible features X and a marginal probability distribution $P(X)$. A task T encompasses a label space Y and a function $f(\bullet)$:

$$D = \{X, P(X)\} \tag{18}$$

$$T = \{Y, f(\bullet)\} \tag{19}$$

In transfer learning, as opposed to traditional machine learning, the domains and/or tasks of the source and target settings are different. The survey paper distinguishes the following variants: Inductive transfer learning (the domains are identical, and the tasks are different), Transductive transfer learning (the domains are different, and the tasks are identical) and Unsupervised transfer learning (domains and tasks may be different and labels are not available in each case). In addition, four categories are presented based on what is being transferred: instances, feature representations, model parameters and relational knowledge. Weiss, Khoshgoftaar, and Wang (2016) follow this categorisation in their survey on recent transfer learning methods, while also distinguishing between homogeneous (same feature space in source and target) and heterogeneous (different feature spaces) transfer learning approaches.

For the purpose of this chapter, the focus is placed on transductive transfer learning. This problem is closely related to domain adaptation, and the terms transfer learning and domain adaptation are used somewhat interchangeably in NLP (Pan & Yang, 2010). Within the context of sentiment analysis, the term cross-domain sentiment classification is also commonly used in the literature. Its definition is equivalent to that of transductive transfer learning. A recent survey on the topic of cross-domain transfer learning can be found in R. Liu et al. (2019). Next, several transfer methods are presented along with examples of their applications to SA.

Methods of Transfer

Structural Correspondence Learning (SCL) was introduced by Blitzer, McDonald, and Pereira (2006). It is a feature transfer algorithm that relies on domain-independent features called pivots to learn correspondences between features in the source and target domains. Those pivots are then used to map source and target features into a common latent space, making SCL an example of a symmetric feature transfer algorithm (Weiss et al., 2016). SCL only considers one-to-one mappings between features. N. Li, Zhai, Zhang, and Liu (2017) extended SCL to include one-to-many mappings and used it for cross-lingual SA, with English as the source and Chinese as the target. Spectral Feature Alignment (SFA) was proposed by Pan, Ni, Sun, Yang, and Chen (2010). This algorithm creates clusters of source and target features in a common latent space. It constructs a bipartite graph, using domain-independent features as a bridge to bring corresponding domain-specific features closer together. The pivots are selected by computing the mutual information between features and domains. SFA does not require labelled data in the target domain. Recently, Hao et al. (2019) introduced CrossWord, which makes use of stochastic word embedding to learn an alignment between domains.

Autoencoders have been successfully applied to transfer learning as well. Glorot, Bordes, and Bengio (2011) extracted a high-level shared representation across multiple domains (Amazon product reviews) in an unsupervised manner with SDAs. The benefit of this approach is that is scales well with larger amounts of data. Zhou, Zhu, He, and Hu (2016) used SDAs to learn language-independent features and perform cross-lingual SA from English to Chinese. Long, Wang, Cao, Sun, and Yu (2016) proposed a framework combining unsupervised pre-training with denoising autoencoders and supervised fine-tuning with deep neural nets to improve transferability.

Ganin et al. (2016) introduced Domain-Adversarial Neural Network (DANN) to improve upon existing autoencoder-based methods. DANN is an augmentation technique for feedforward networks, allowing them to learn features that are both discriminative and invariant to domain shift while being trainable with backpropagation.

Yu and Jiang (2016) apply the pivot prediction concept of SCL to neural networks. They introduce two auxiliary binary tasks to detect the presence of positive and negative domain-independent words in a sentence. The network is then jointly trained to learn both the feature embedding and the classifier at the same time, outperforming several state-of-the-art methods.

Attention models can also be applied to cross-domain SA. Z. Li, Zhang, Wei, Wu, and Yang (2017) introduced the Adversarial Memory Network (AMN) as an improvement over previous deep learning-based methods in terms of interpretability of the pivots. Z. Li, Wei, Zhang, and Yang (2018) developed the Hierarchical Attention Transfer Network (HATN). HATN consists of two subsets named P-Net and NP-Net. The P-Net discovers pivots, and the NP-Net performs feature alignment using the pivots as a bridge. The advantage of this method over algorithms like SCL and SFA is that the pivots are selected automatically. CCHAN (Manshu & Xuemin, 2019) is another combined attention model, consisting of a cloze task network (CTN) performing the word embedding task and a convolutional HAN (CHAN) for sentiment classification. The two networks are jointly trained in an end-to-end fashion. The Hierarchical Attention Network with Prior knowledge information (HANP) was further recently proposed by Manshu and Bing (2019). It adds prior knowledge of the contextual meaning of sentiment words via a sentiment dictionary match (SDM) layer to identify domain-dependent and domain-independent features simultaneously.

Yin, Liu, Zhu, Li, and Wang (2019) introduced Capsule Net with Identifying Transferable Knowledge (CITK). This method includes domain-invariant knowledge extracted with a lexicon-based method in the network to help with pivot identification and generalisation.

Transformers have also shown promising results for cross-domain applications due to their capability of learning high-level feature representations. A recent example is the work by Myagmar, Li, and Kimura (2019), applying transformers to Amazon product reviews.

SOLUTIONS AND RECOMMENDATIONS

This section presents solutions and makes recommendations for readers interesting in applying state-of-the-art models to SA problems. First, a number of popular datasets and challenges are described.

Datasets and Tasks

IMDB Dataset

The IMDB dataset[4] (Maas et al., 2011) contains 50000 movie reviews that are annotated as positive or negative. The reviews are highly polarised, and the data is split evenly between positive and negative reviews.

Yelp Dataset

The Yelp review dataset[5] (X. Zhang, Zhao, & LeCun, 2015) was created from the ongoing Yelp Dataset Challenge. It encompasses two tasks: predicting the review polarity and predicting the number of stars given by the user. The dataset is evenly split between classes, with 280000 training and 19000 test samples for each polarity and 130000 training and 10000 test samples for each star rating.

Stanford Sentiment Treebank

The Stanford Sentiment Treebank (SST) dataset[6] (Socher et al., 2013) contains 215154 phrases parsed from 11855 sentences that were extracted from movie reviews. It provides both coarse-grained (binary) and fine-grained (five points) annotations.

SemEval-2017 Task 4

Task 4 of the International Workshop on Semantic Evaluation (Rosenthal, Farra, & Nakov, 2017) is concerned with SA on Twitter. The task was held yearly since 2013 and continuously expanded. The 2017 task added Arabic as a second language to English. There were five subtasks: polarity classification of single tweets, targeted polarity classification of single tweets in two and five classes, estimating the distribution of a set of tweets across two and five classes.

Applying State of the Art Models

The current state of the art in SA, as well as NLP in general, is based on Transformer networks. This means that pre-trained word embeddings generated by GloVe, Word2vec and fastText are no longer recommended. In 2018, all competitors in the SocialNLP EmotionX Challenge (Hsu & Ku, 2018) used one of those toolkits. By 2019, all the best contributions were utilizing pre-trained embeddings generated with BERT.

As discussed in the previous section, BERT provides powerful text representations through pre-training on a large document corpus. Versions of BERT trained for various languages and of different sizes (named BERT-Base and BERT-Large) have been made publicly available[7]. Thus, the recommended workflow for readers interested in using BERT for SA is to obtain a suitable pre-trained model, e.g. BERT-Large in English, and then further adapt it to their specific task.

An instructive example of how this tuning can be achieved is given in the work of C. Sun, Qiu, Xu, and Huang (2019). They outline three steps for improving the performance of BERT-Base and BERT-Large:

1. **Further Pre-training:** BERT is pre-trained on a large collection of documents. In a subsequent step, additional pre-training on within-task or in-domain data is performed.
2. **Multi-Task Learning:** The model is trained on multiple tasks simultaneously, with the tasks sharing layers except for the final classification layer. This allows knowledge from different tasks to be shared.
3. **Fine-Tuning on the target task:** The model is further trained to adapt it to a specific task.

Following this approach and testing a number of fine-tuning strategies, including the layer-wise optimisation approach from Howard and Ruder (2018), (C. Sun et al.) developed BERT_large+ITPT, which achieved new state-of-the-art results on a number of text processing tasks, including SA. Specifically, the model obtained test error rates of 4,21% on the IMDB dataset and 1.81% and 28.62% on the coarse-grained and fine-grained tasks of the Yelp dataset, respectively.

Transformer-based methods are continuing to evolve. Many researchers develop variants of BERT, such as RoBERTa (Y. Liu et al., 2019), which further optimises the training process. Recently, Zhilin Yang et al. (2019) introduced XLNet, which replaces the autoencoding paradigm of BERT with generalised autoregression. XLNet incorporates ideas from the Transformer-XL (Dai et al., 2019), an autoregressive model which improves upon the standard Transformer by better handling long-term dependencies. The advantages of XLNet over BERT are that it predicts permutations of a sequence, allowing it to learn bidirectional context more effectively and that it does not rely on masking, which solves several inherent problems of BERT, such as the assumption that masked tokens are independent.

XLNet further improved upon the state of the art in a number of language understanding tasks including SA, yielding test error rates of 3.20% on IMDB, 1.37% on coarse-grained Yelp and 27.05% on fine-grained Yelp, as well as 3.2% on SST.

To conclude this section, readers are recommended to use the latest developments in Transformer models for SA. While XLNet has outperformed BERT in a number of popular SA tasks and may become the new standard due to its powerful permutation-based language modelling, BERT variants like RoBERTa could still be useful depending on the problem to be solved. Thus, the readers are encouraged to experiment with these models while observing further developments in the field.

FUTURE RESEARCH DIRECTIONS

NLP in general and sentiment analysis in particular are already being used in many business applications, as discussed in the introduction to this chapter. The amount and diversity of available data continue to grow, which motivates the use of deep learning techniques due to their potent feature extraction capabilities. This section outlines a number of trends and promising research opportunities.

Data Augmentation

One open issue is the need for compensating class imbalance, i.e. the number of instances of each class not being evenly distributed in a labelled dataset. Class imbalance affects many datasets collected in realistic settings, and often a minority class will be of great interest. This is problematic since many classifiers, including deep learning methods, will exhibit a bias towards the majority class (Johnson & Khoshgoftaar, 2019).

Data augmentation is a data-based solution to this problem. It enriches the dataset with additional examples of minority instances. While such augmentation can be easily applied to image data, e.g. by adding noise, rotating or mirroring, it is less straightforward for NLP, as the resulting text sample still needs to make sense. Consequentially, this technique has received comparatively little attention in textual SA. Recently, however, a promising approach for applying data augmentation to SA has been presented by Rizos, Hemker, and Schuller (2019), who use it for improving online hate speech classification. The strategies employed in the paper include: replacing words with synonyms which are discovered through similarities of their embeddings, shifting the positions of words within the sentence, and generating new text through sequential prediction with RNNs or transformers.

Zero-Shot Learning

Aside from improving training through data augmentation, an interesting strategy for dealing with missing data is to apply zero-shot learning techniques. The goal of zero-shot learning, also referred to as zero-data learning, is to recognise classes at test time that were not seen during training (Larochelle, Erhan, & Bengio, 2008), i.e. there were no instances of those classes for the model to learn from. In the related case where only a few instances are present in the training data, methods are commonly referred to as one-shot or few-shot learning.

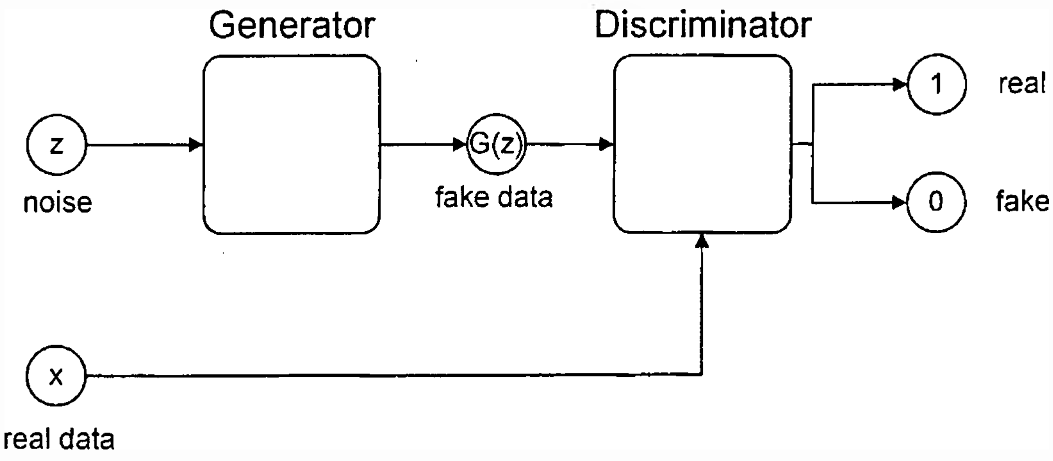
Zero-shot learning is increasingly used for large-scale classification problems where annotating all classes extensively is not possible. For the field of visual object detection, there already exist numerous benchmark datasets such as Animals with Attributes (AWA) (Lampert, Nickisch, & Harmeling, 2014). Recently, Xian, Lampert, Schiele, and Akata (2019) published an overview of the state of the art in zero-shot learning, finding a proliferation of approaches but a lack of comparability and flaws of methodology, and introduced a novel dataset called Animals with Attributes 2 (AWA2), along with proposing a standardised evaluation procedure.

Zero-shot learning techniques frequently rely on knowledge in a semantic embedding space (Norouzi et al., 2014), (Z. Zhang & Saligrama, 2015). Applying such techniques to NLP and SA tasks in particular is a promising research direction.

Adversarial Learning

The concept of adversarial networks was introduced by Goodfellow et al. (2014). In a generative adversarial network (GAN), two networks, named generator and discriminator, compete with each other, with the generator attempting to produce samples resembling that of a target distribution and the discriminator attempting to differentiate between real and artificial samples. A basic GAN architecture is depicted in Figure 6.

Figure 6. GAN architecture. The generator creates a fake sample mimicking the training data. The discriminator attempts to tell real from fake samples. Both networks are trained against each other until an equilibrium is reached.



The concept of adversarial training has been applied to many disciplines, including sentiment analysis. Numerous works make use of adversarial networks for cross-domain sentiment classification (Y. Zhang, Barzilay, & Jaakkola, 2017), (Duan, Zhou, Jing, Zhang, & Chen, 2018), (W. Liu & Fu, 2018). In addition, adversarial networks can be used in a generative way to change the style of sentences, outperforming previous approaches based on encoder-decoder architectures (Choi, Choi, Park, & Lee, 2019), (John, Mou, Bahuleyan, & Vechtomova, 2019). While these results are promising, adversarial networks applied to text and speech have yet to reach the same levels of performance as in image generation (Han, Zhang, Cummins, & Schuller, 2019).

Transfer Learning

An emerging trend that is certain to play a major role in the future is the proliferation and improvement of transfer learning methods. This will allow businesses to leverage existing knowledge in the form of models and datasets for new applications, which could significantly speed up time to market and reduce development costs. In terms of research opportunities, cross-lingual transfer is attractive, since most studies on sentiment analysis focus on English documents.

Explainable AI

While deep learning-based models have achieved impressive results, they are frequently applied in a black-box manner, i.e., no information is given about how those systems reach a conclusion. This is a consequence of the massive datasets processed and the highly complex features derived from them by the deep learning algorithms, which may be difficult or impossible for humans to understand. This lack of transparency limits the effectiveness of such systems and is the motivation for the development of explainable AI (XAI). XAI aims to create models that can maintain high levels of performance while allowing humans to understand and trust their decisions (Mathews, 2019).

XAI strategies can be classified into two broad categories: model-based (intrinsic) and post-hoc explainability (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019). Intrinsic approaches aim to make the model itself more explainable, e.g. by reducing its complexity. Post-hoc methods are designed to analyse an existing model. An example of a popular post-hoc framework is LIME (Marco Tulio Ribeiro, Singh, & Guestrin, 2016). Murdoch et al. (2019) formulate three criteria for grading an interpretation: predictive accuracy, descriptive accuracy, and relevancy.

A possible solution for interpretability is the use of attention. (Letarte, Paradis, Giguère, & Laviolette, 2018) introduced a self-attention network based on the Transformer. They found that visualising the relationships between words found by attention helped explain differences in the model's behaviour between topic classification and sentiment analysis. (Peters, Niculae, & Martins, 2018) demonstrated how regularised attention can be used to create sparse, ordered structures in the layers of deep neural networks, which benefits interpretability.

As automated solutions spread and become increasingly complex, explainable AI will continue to become more relevant, both as a means for building trust with the customers employing a system and as a way for the business offering that system to improve performance.

Defending Against Adversarial Attacks

On a related note, an important area of research that is starting to be explored is the robustness of NLP algorithms. Complex classifiers, while being powerful pattern detectors, are also prone to changing their predictions based on small perturbations in the input data. This weakness has been shown to be exploitable through so-called adversarial attacks. The attacker designs manipulated instances of input data (adversarial samples), which are misclassified by the targeted model. Recently, M. T. Ribeiro, Singh, and Guestrin (2018) demonstrated how to apply this concept to NLP, using semantically equivalent adversarial rules (SEARs) to construct adversarial examples from text while maintaining the semantic content. Given these vulnerabilities, further investigation into adversarial attacks in order to improve models and make them safer to use is a promising line of research.

Multimodal Sentiment Analysis

Another interesting research direction is to perform SA based on multiple modalities, e.g. text, audio and visual data from videos. This will allow for a more robust sentiment detection, as the model can combine information across modalities for decision making. A recent work on cross-domain sentiment analysis that makes use of Bag-of-Words features derived from text, speech and facial expressions is (Cummins et al., 2018).

CONCLUSION

This chapter has introduced sentiment analysis as an important topic in natural language processing. It has highlighted numerous business and academic applications, including customer analytics, financial market predictions and estimating public sentiment from social media posts, and provided a categorisation of sentiment analysis approaches. Deep learning was presented as a useful collection of methods to extract information from increasingly large amounts of unstructured data. The basic architectures of CNNs and RNNs were introduced, as well as their combination into hybrid networks. Current trends and state-of-the-art methods were explored, covering attention, transfer learning and Transformer networks. The challenges of explainable AI, data augmentation, zero-shot learning, adversarial learning. the threat of adversarial attacks and the potential of multimodal analysis were explained and highlighted as opportunities for future research.

ACKNOWLEDGMENT

This research was supported by the BMW Group.

REFERENCES

Bahdanau, D., Cho, K., & Bengio, Y. (2014, September 1). *Neural Machine Translation by Jointly Learning to Align and Translate.* Retrieved from https://arxiv.org/pdf/1409.0473v7

Balaji, P., Nagaraju, O., & Haritha, D. (2017). Levels of sentiment analysis and its challenges: A literature review. In *Proceedings of the 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)* (pp. 436–439). IEEE. 10.1109/ICBDACI.2017.8070879

Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex (New York, N.Y.)*, *10*(3), 295–307. doi:10.1093/cercor/10.3.295 PMID:10731224

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. doi:10.1109/TPAMI.2013.50 PMID:23787338

Blitzer, J., McDonald, R., & Pereira, F. (2006). Domain Adaptation with Structural Correspondence Learning. In *EMNLP '06, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 120–128). Association for Computational Linguistics. doi:10.3115/1610075.1610094

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. doi:10.1162/tacl_a_00051

Chen, P., Sun, Z., Bing, L., & Yang, W. (2017). Recurrent Attention Network on Memory for Aspect Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 452–461). Copenhagen, Denmark: Association for Computational Linguistics. 10.18653/v1/D17-1047

Cheng, X., Xu, W., & Wang, T., Chu, W., Huang, W., Chen, K., & Hu, J. (2019). Variational Semi-Supervised Aspect-Term Sentiment Analysis via Transformer. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* (pp. 961–969). Hong Kong, China: Association for Computational Linguistics. 10.18653/v1/K19-1090

Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (pp. 103–111). Doha, Qatar: Association for Computational Linguistics. 10.3115/v1/W14-4012

Choi, W., Choi, S. J., Park, S., & Lee, S. (2019). Adversarial Style Transfer for Long Sentences. *2019 International Conference on Electronics, Information, and Communication (ICEIC)*. 10.23919/ELINFOCOM.2019.8706482

Christy Daniel, D., & Shyamala, L. (2019). An insight on sentiment analysis research from text using deep learning methods. *International Journal of Innovative Technology and Exploring Engineering*, *8*(10), 2033–2048. doi:10.35940/ijitee.J9316.0881019

Cooper, P. (2019). *28 Twitter Statistics All Marketers Need to Know in 2019*. Retrieved from https://blog.hootsuite.com/twitter-statistics/

Cummins, N., Amiriparian, S., Ottl, S., Gerczuk, M., Schmitt, M., & Schuller, B. (2018). Multimodal Bag-of-Words for Cross Domains Sentiment Analysis. In *ICASSP-2018, Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4954–4958). IEEE. 10.1109/ICASSP.2018.8462660

Dai, Z., & Yang, Z., Yang, Y., Carbonell, J., Le, Q., & Salakhutdinov, R. (2019). Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2978–2988). Florence, Italy: Association for Computational Linguistics. 10.18653/v1/P19-1285

Della Pietra, S., Della Pietra, V., & Lafferty, J. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(4), 380–393. doi:10.1109/34.588021

Deng, D., Jing, L., Yu, J., & Sun, S. (2019). Sparse Self-Attention LSTM for Sentiment Lexicon Construction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(11), 1777–1790. doi:10.1109/TASLP.2019.2933326

Desjardins, J. (2019). *What Happens in an Internet Minute in 2019?* Retrieved from https://www.visualcapitalist.com/what-happens-in-an-internet-minute-in-2019/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* CoRR, abs/1810.04805

Du, C., Tsai, M., & Wang, C. (2019). Beyond Word-level to Sentence-level Sentiment Analysis for Financial Reports. *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 10.1109/ICASSP.2019.8683085

Du, J., Gui, L., Xu, R., & He, Y. (2018). A Convolutional Attention Model for Text Classification. In X. Huang, J. Jiang, D. Zhao, Y. Feng, & Y. Hong (Eds.), Lecture Notes in Computer Science: Vol. 10619. *Natural Language Processing and Chinese Computing* (pp. 183–195). Springer International Publishing. doi:10.1007/978-3-319-73618-1_16

Duan, X., Zhou, Y., Jing, C., Zhang, L., & Chen, R. (2018). Cross-domain Sentiment Classification Based on Transfer Learning and Adversarial Network. In *Proceedings of the 2018 IEEE 4th International Conference on Computer and Communications (ICCC)* (pp. 2302–2306). IEEE. 10.1109/CompComm.2018.8780771

Ekman, P. (1999). Basic emotions. Handbook of Cognition and Emotion, 98(45-60), 16.

Evgeniou, T., & Pontil, M. (2001). *Support Vector Machines: Theory and Applications* (Vol. 2049). Springer. doi:10.1007/3-540-44673-7_12

Facebook. (2019). *Facebook Reports Second Quarter 2019 Results.* Retrieved from https://investor.fb.com/investor-news/press-release-details/2019/Facebook-Reports-Second-Quarter-2019-Results/default.aspx

Fu, X., Wei, Y., Xu, F., Wang, T., Lu, Y., Li, J., & Huang, J. Z. (2019). Semi-supervised Aspect-level Sentiment Classification Model based on Variational Autoencoder. *Knowledge-Based Systems*, *171*, 81–92. doi:10.1016/j.knosys.2019.02.008

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., . . . Lempitsky, V. S. (2016). Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.*, *17*, 59:1-59:35. Retrieved from http://jmlr.org/papers/v17/15-239.html

Gao, Z., Feng, A., Song, X., & Wu, X. (2019). Target-Dependent Sentiment Classification With BERT. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 154290–154299. doi:10.1109/ACCESS.2019.2946594

Gautam, G., & Yadav, D. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In M. Parashar (Ed.), *Proceedings of the 2014 Seventh International Conference on Contemporary Computing (IC3): 7 - 9 Aug. 2014, Noida, India* (pp. 437–442). Piscataway, NJ: IEEE. 10.1109/IC3.2014.6897213

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural Computation*, *12*(10), 2451–2471. doi:10.1162/089976600300015015 PMID:11032042

Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain Adaptation for Large-scale Sentiment Classification: A Deep Learning Approach. In *ICML'11, Proceedings of the 28th International Conference on International Conference on Machine Learning* (pp. 513–520). Omnipress. Retrieved from https://dl.acm.org/citation.cfm?id=3104482.3104547

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2014). Generative Adversarial Nets. In *NIPS'14, Proceedings of the 27th International Conference on Neural Information Processing Systems* (Vol. 2, pp. 2672–2680). MIT Press. Retrieved from https://dl.acm.org/citation.cfm?id=2969033.2969125

Graves, A. (2012). Supervised Sequence Labelling with Recurrent Neural Networks (2nd ed.). In Studies in Computational Intelligence: Vol. 385. Berlin: Springer Berlin Heidelberg. doi:10.1007/978-3-642-24797-2

Han, J., Zhang, Z., Cummins, N., & Schuller, B. (2019). Adversarial Training in Affective Computing and Sentiment Analysis: Recent Advances and Perspectives [Review Article]. *IEEE Computational Intelligence Magazine*, *14*(2), 68–81. doi:10.1109/MCI.2019.2901088

Hao, Y., Mu, T., Hong, R., Wang, M., Liu, X., & Goulermas, J. Y. (2019). Cross-domain Sentiment Encoding through Stochastic Word Embedding. *IEEE Transactions on Knowledge and Data Engineering*, *1*, 1. Advance online publication. doi:10.1109/TKDE.2019.2913379

Hassan, A., & Mahmood, A. (2018). Convolutional Recurrent Deep Learning Model for Sentence Classification. *IEEE Access: Practical Innovations, Open Solutions*, *6*, 13949–13957. doi:10.1109/ACCESS.2018.2814818

Hatzivassiloglou, V., & McKeown, K. (1997). Predicting the Semantic Orientation of Adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 174–181). Madrid, Spain: Association for Computational Linguistics. 10.3115/976909.979640

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97. doi:10.1109/MSP.2012.2205597

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735 PMID:9377276

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (*Volume 1*: Long Papers).* Melbourne, Australia: Association for Computational Linguistics. 10.18653/v1/P18-1031

Hsu, C.-C., & Ku, L.-W. (2018). SocialNLP 2018 EmotionX Challenge Overview: Recognizing Emotions in Dialogues. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media* (pp. 27–31). Melbourne, Australia: Association for Computational Linguistics. 10.18653/v1/W18-3505

Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. In *KDD '04, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 168–177). New York, NY: ACM. 10.1145/1014052.1014073

Hussein, D. M. E.-D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University -. Engineering and Science*, *30*(4), 330–338. doi:10.1016/j.jksues.2016.04.002

Jiang, M., Wu, J., Shi, X., & Zhang, M. (2019). Transformer Based Memory Network for Sentiment Analysis of Web Comments. *IEEE Access: Practical Innovations, Open Solutions*, *1*, 179942–179953. Advance online publication. doi:10.1109/ACCESS.2019.2957192

John, V., Mou, L., Bahuleyan, H., & Vechtomova, O. (2019). Disentangled Representation Learning for Non-Parallel Text Style Transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 424–434). Florence, Italy: Association for Computational Linguistics. 10.18653/v1/P19-1041

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 27. doi:10.118640537-019-0192-5

Jose, R., & Chooralil, V. S. (2016). Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble Approach. In *Proceedings of the 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)* (pp. 64–67). IEEE. 10.1109/SAPIENCE.2016.7684133

Joshi, O. S., & Simon, G. (2018). Sentiment Analysis Tool on Cloud: Software as a Service Model. In *Proceedings of the 2018 International Conference On Advances in Communication and Computing Technology (ICACCT)* (pp. 459–462). Sangamner, India: Springer. 10.1109/ICACCT.2018.8529649

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 427–431). Valencia, Spain: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/E17-2068

Joyce, B., & Deng, J. (2017). Sentiment analysis of tweets for the 2016 US presidential election. In *Proceedings of the 2017 IEEE MIT Undergraduate Research Technology Conference (URTC)* (pp. 1–4). IEEE. 10.1109/URTC.2017.8284176

Kaur, H., & Mangat, V., & Nidhi (2017). A survey of sentiment analysis techniques. *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).* 10.1109/I-SMAC.2017.8058315

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746–1751). Doha, Qatar: Association for Computational Linguistics. 10.3115/v1/D14-1181

Kingma, D. P., & Welling, M. (2013, December 20). *Auto-Encoding Variational Bayes*. Retrieved from https://arxiv.org/pdf/1312.6114v10

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS'12, Proceedings of the 25th International Conference on Neural Information Processing Systems* - Volume 1 (pp. 1097–1105). Curran Associates Inc.

Lampert, C. H., Nickisch, H., & Harmeling, S. (2014). Attribute-Based Classification for Zero-Shot Visual Object Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(3), 453–465. doi:10.1109/TPAMI.2013.140 PMID:24457503

Larochelle, H., Erhan, D., & Bengio, Y. (2008). Zero-data Learning of New Tasks. In *AAAI'08, Proceedings of the 23rd National Conference on Artificial Intelligence* (Vol. 2, pp. 646–651). AAAI Press. Retrieved from https://dl.acm.org/citation.cfm?id=1620163.1620172

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. doi:10.1038/nature14539 PMID:26017442

Letarte, G., Paradis, F., Giguère, P., & Laviolette, F. (2018). Importance of Self-Attention for Sentiment Analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 267–275). Brussels, Belgium: Association for Computational Linguistics. 10.18653/v1/W18-5429

Li, N., Zhai, S., & Zhang, Z., & Liu, B. (2017). Structural Correspondence Learning for Cross-lingual Sentiment Classification with One-to-many Mappings. In *AAAI'17, Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (pp. 3490–3496). AAAI Press. Retrieved from https://dl.acm.org/citation.cfm?id=3298023.3298075

Li, Z., & Wei, Y., Zhang, Y., & Yang, Q. (2018). Hierarchical Attention Transfer Network for Cross-Domain Sentiment Classification. *AAAI Conference on Artificial Intelligence; Thirty-Second AAAI Conference on Artificial Intelligence*. Retrieved from https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16873

Li, Z., & Zhang, Y., Wei, Y., Wu, Y., & Yang, Q. (2017). End-to-end Adversarial Memory Network for Cross-domain Sentiment Classification. In C. Sierra (Ed.), *IJCAI'17, Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 2237–2243). AAAI Press. Retrieved from https://dl.acm.org/citation.cfm?id=3172077.3172199

Liu, G., & Guo, J. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing*, *337*, 325–338. doi:10.1016/j.neucom.2019.01.078

Liu, Q., Zhang, H., Zeng, Y., Huang, Z., & Wu, Z. (2018). Content Attention Model for Aspect Based Sentiment Analysis. In *WWW '18, Proceedings of the 2018 World Wide Web Conference* (pp. 1023–1032). Geneva, Switzerland: International World Wide Web Conferences Steering Committee. 10.1145/3178876.3186001

Liu, R., Shi, Y., Ji, C., & Jia, M. (2019). A Survey of Sentiment Analysis Based on Transfer Learning. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 85401–85412. doi:10.1109/ACCESS.2019.2925059

Liu, W., & Fu, X. (2018). Introduce More Characteristics of Samples into Cross-domain Sentiment Classification. In *ICPR 2018, Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR)* (pp. 25–30). IEEE. 10.1109/ICPR.2018.8545331

Liu, Y., Ott, M., Goyal, N., Du Jingfei, Joshi, M., Chen, D., . . . Stoyanov, V. (2019, July 26). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. Retrieved from https://arxiv.org/pdf/1907.11692v1

Liu, Z., Bai, X., Cai, T., Chen, C., Zhang, W., & Jiang, L. (2019). Improving Sentence Representations with Local and Global Attention for Classification. In *IJCNN 2019, Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–7). Curran Associates, Inc. 10.1109/IJCNN.2019.8852436

Long, M., Wang, J., Cao, Y., Sun, J., & Yu, P. S. (2016). Deep Learning of Transferable Representation for Scalable Domain Adaptation. *IEEE Transactions on Knowledge and Data Engineering*, *28*(8), 2027–2040. doi:10.1109/TKDE.2016.2554549

Lu, G., Zhao, X., Yin, J., & Yang, W., & Li, B. (2018). Multi-task learning using variational auto-encoder for sentiment classification. *Pattern Recognition Letters*. Advance online publication. doi:10.1016/j.patrec.2018.06.027

Luo, X., Zimet, G., & Shah, S. (2019). A natural language processing framework to analyse the opinions on HPV vaccination reflected in twitter over 10 years (2008 - 2017). *Human Vaccines & Immunotherapeutics*, *15*(7-8), 1496–1504. doi:10.1080/21645515.2019.1627821 PMID:31194609

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 142–150). Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P11-1015

Maitra, P., & Sarkhel, R. (2018). A K-Competitive Autoencoder for Aggression Detection in Social Media Text. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)* (pp. 80–89). Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W18-4410

Makhzani, A., & Frey, B. (2015). *Winner-take-all autoencoders*. MIT Press.

Manshu, T., & Bing, W. (2019). Adding Prior Knowledge in Hierarchical Attention Neural Network for Cross Domain Sentiment Classification. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 32578–32588. doi:10.1109/ACCESS.2019.2901929

Manshu, T., & Xuemin, Z. (2019). CCHAN: An End to End Model for Cross Domain Sentiment Classification. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 50232–50239. doi:10.1109/ACCESS.2019.2910300

Mathews, S. M. (2019). Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review. *Advances in Intelligent Systems and Computing*, *998*, 1269–1292. doi:10.1007/978-3-030-22868-2_90

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. Retrieved from https://arxiv.org/pdf/1301.3781.pdf

Mikolov, T., Sutskever, I., & Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In *NIPS'13, Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (pp. 3111–3119). Curran Associates Inc. Retrieved from https://dl.acm.org/citation.cfm?id=2999792.2999959

Müller, M. M., & Salathé, M. (2019). Crowdbreaks: Tracking health trends using public social media data and crowdsourcing. *Frontiers in Public Health*, *7*(APR), 81. Advance online publication. doi:10.3389/fpubh.2019.00081 PMID:31037238

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(44), 22071–22080. doi:10.1073/pnas.1900654116 PMID:31619572

Myagmar, B., Li, J., & Kimura, S. (2019). Cross-Domain Sentiment Classification With Bidirectional Contextualized Transformer Language Models. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 163219–163230. doi:10.1109/ACCESS.2019.2952360

Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)* (pp. 1–5). IEEE. 10.1109/ICCCNT.2013.6726818

Nigam, K., & Lafferty, J., & Mccallum, A. (1999). Using maximum entropy for text classification. In *IJCAI-99, Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filtering* (pp. 61–67). AAAI Press.

Niu, X., & Hou, Y. (2017). Hierarchical Attention BLSTM for Modeling Sentences and Documents. Lecture Notes in Computer Science, 10635, 167–177. doi:10.1007/978-3-319-70096-0_18

Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., . . . Dean, J. (2014). Zero-Shot Learning by Convex Combination of Semantic Embeddings. In *2nd International Conference on Learning Representations, ICLR 2014*. Conference Track Proceedings.

Pan, S. J., Ni, X., Sun, J.-T., Yang, Q., & Chen, Z. (2010). Cross-domain Sentiment Classification via Spectral Feature Alignment. In *WWW '10, Proceedings of the 19th International Conference on World Wide Web* (pp. 751–760). New York, NY: ACM. 10.1145/1772690.1772767

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. doi:10.1109/TKDE.2009.191

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment classification using machine learning techniques*. Association for Computational Linguistics. Retrieved from https://dl.acm.org/ft_gateway.cfm?id=1118704&type=pdf

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Retrieved from https://www.aclweb.org/anthology/D14-1162.pdf

Peters, B., Niculae, V., & Martins, A. F. T. (2018). Interpretable Structure Induction via Sparse Attention. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 365–367). Brussels, Belgium: Association for Computational Linguistics. 10.18653/v1/W18-5450

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellerman (Eds.), *Theories of Emotion* (pp. 3–33). Academic Press. doi:10.1016/B978-0-12-558701-3.50007-7

Rameshbhai, C. J., & Paulose, J. (2019). Opinion mining on newspaper headlines using SVM and NLP. *Iranian Journal of Electrical and Computer Engineering*, *9*(3), 2152–2163. doi:10.11591/ijece.v9i3.pp2152-2163

Ran, J. (2019). A Self-attention Based LSTM Network for Text Classification. *Journal of Physics: Conference Series*, *1207*, 12008. doi:10.1088/1742-6596/1207/1/012008

Ranjit, S., Shrestha, S., Subedi, S., & Shakya, S. (2018). Foreign Rate Exchange Prediction Using Neural Network and Sentiment Analysis. *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN).* 10.1109/ICACCCN.2018.8748819

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *KDD '16, Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). New York, NY: ACM. 10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (Eds.). (2018). *Semantically equivalent adversarial rules for debugging NLP models.* Retrieved from https://www2.scopus.com/inward/record.uri?eid=2-s2.0-85061785761&partnerID=40&md5=be8d9d4a9111c0f0f6ba388f3dcc16bb

Rizos, G., Hemker, K., & Schuller, B. (2019). Augment to Prevent: Short-Text Data Augmentation in Deep Learning for Hate-Speech Classification. In *CIKM '19, Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 991–1000). New York, NY: ACM. 10.1145/3357384.3358040

Rosenthal, S., Farra, N., & Nakov, P. (2017). SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 502–518). Vancouver, Canada: Association for Computational Linguistics. 10.18653/v1/S17-2088

Sagha, H., Cummins, N., & Schuller, B. (2017). Stacked denoising autoencoders for sentiment analysis: A review. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, *7*(5), e1212. doi:10.1002/widm.1212

Save, A., & Shekokar, N. (2017). Analysis of cross domain sentiment techniques. *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT).* 10.1109/ICEECCOT.2017.8284637

Shah, D., Isah, H., & Zulkernine, F. (2018). Predicting the Effects of News Sentiments on the Stock Market. *2018 IEEE International Conference on Big Data (Big Data)*. 10.1109/BigData.2018.8621884

Shmueli, B., & Ku, L.-W. (2019). *SocialNLP EmotionX 2019 Challenge Overview: Predicting Emotions in Spoken Dialogues and Chats*. Retrieved from https://arxiv.org/abs/1909.07734

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1631–1642). Seattle, WA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D13-1170

Stappen, L., Cummins, N., Meßner, E.-M., Baumeister, H., Dineley, J., & Schuller, B. W. (2019). Context Modelling Using Hierarchical Attention Networks for Sentiment and Self-assessed Emotion Detection in Spoken Narratives. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6680–6684). Brighton: IEEE. 10.1109/ICASSP.2019.8683801

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? In M. Sun, X. Huang, H. Ji, Z. Liu, & Y. Liu (Eds.), *LNCS sublibrary. SL 7, Artificial intelligence:* v. 11856. *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings* (pp. 194–206). Cham: Springer. 10.1007/978-3-030-32381-3_16

Taboada, M., Anthony, C., & Voll, K. (2006). Methods for Creating Semantic Orientation Databases. *Proceeding of LREC-06, the 5th International Conference on Language Resources and Evaluation.* Retrieved from https://www.microsoft.com/en-us/research/publication/methods-for-creating-semantic-orientation-databases/

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307. doi:10.1162/COLI_a_00049

Tang, D., Qin, B., & Liu, T. (2015). Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1422–1432). Lisbon, Portugal: Association for Computational Linguistics. 10.18653/v1/D15-1167

Tang, D., Qin, B., & Liu, T. (2016). Aspect Level Sentiment Classification with Deep Memory Network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 214–224). Austin, TX: Association for Computational Linguistics. 10.18653/v1/D16-1021

Tang, G., Müller, M., Rios, A., & Sennrich, R. (2018). Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4263–4272). Brussels, Belgium: Association for Computational Linguistics. 10.18653/v1/D18-1458

Tian, Z., Rong, W., Shi, L., Liu, J., & Xiong, Z. (2018). Attention Aware Bidirectional Gated Recurrent Unit Based Framework for Sentiment Analysis. In W. Liu, F. Giunchiglia, & B. Yang (Eds.), *Knowledge Science, Engineering and Management* (pp. 67–78). Springer International Publishing. doi:10.1007/978-3-319-99365-2_6

Turney, P. D. (2002). *Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews*: Association for Computational Linguistics. Retrieved from https://dl.acm.org/ft_gateway.cfm?id=1073153&type=pdf

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. Springer New York., doi:10.1007/978-1-4757-3264-1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and Composing Robust Features with Denoising Autoencoders. In *ICML '08, Proceedings of the 25th International Conference on Machine Learning* (pp. 1096–1103). New York, NY: ACM. 10.1145/1390156.1390294

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 353–355). Brussels, Belgium: Association for Computational Linguistics. 10.18653/v1/W18-5446

Wang, X., Jiang, W., & Luo, Z. (2016). Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Retrieved from https://www.aclweb.org/anthology/C16-1229.pdf

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data, 3*(1), 1817. doi:10.118640537-016-0043-6

Wiebe, J. (2000). Learning Subjective Adjectives from Corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence* (pp. 735–740). AAAI Press. Retrieved from https://dl.acm.org/citation.cfm?id=647288.721121

Wiebe, J., Bruce, R., & O'Hara, T. P. (1999). Development and Use of a Gold-Standard Data Set for Subjectivity Classifications. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. Retrieved from https://www.aclweb.org/anthology/P99-1032.pdf

Wu, X., Cai, Y., Li, Q., Xu, J., & Leung, H.-F. (2018). Combining Contextual Information by Self-attention Mechanism in Convolutional Neural Networks for Text Classification. Lecture Notes in Computer Science, 11233, 453–467. doi:10.1007/978-3-030-02922-7_31

Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2019). Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(9), 2251–2265. doi:10.1109/TPAMI.2018.2857768 PMID:30028691

Xu, N. (2017). Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In *Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics (ISI)* (pp. 152–154). IEEE. 10.1109/ISI.2017.8004895

Xu, W., Sun, H., Deng, C., & Tan, Y. (2017). Variational Autoencoder for Semi-Supervised Text Classification. In *AAAI'17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (Vol. 4, pp. 3358–3364). San Francisco, CA: AAAI Press.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1480–1489). San Diego, CA: Association for Computational Linguistics. 10.18653/v1/N16-1174

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems 32* (pp. 5754–5764). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding.pdf

Yin, H., Liu, P., Zhu, Z., Li, W., & Wang, Q. (2019). Capsule Network With Identifying Transferable Knowledge for Cross-Domain Sentiment Classification. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 153171–153182. doi:10.1109/ACCESS.2019.2948628

Yu, J., & Jiang, J. (2016). Learning Sentence Embeddings with Auxiliary Tasks for Cross-Domain Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 236–246). Austin, TX: Association for Computational Linguistics. 10.18653/v1/D16-1023

Zhang, Q., Lu, R., Wang, Q., Zhu, Z., & Liu, P. (2019). Interactive Multi-Head Attention Networks for Aspect-Level Sentiment Classification. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 160017–160028. doi:10.1109/ACCESS.2019.2951283

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-Level Convolutional Networks for Text Classification. In *NIPS'15, Proceedings of the 28th International Conference on Neural Information Processing Systems* - Volume 1 (pp. 649–657). Cambridge, MA: MIT Press.

Zhang, Y., Barzilay, R., & Jaakkola, T. (2017). Aspect-augmented Adversarial Networks for Domain Adaptation. *Transactions of the Association for Computational Linguistics*, *5*(1), 515–528. doi:10.1162/tacl_a_00077

Zhang, Z., & Saligrama, V. (2015). Zero-Shot Learning via Semantic Similarity Embedding. In *ICCV'15, Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 4166–4174). ACM. 10.1109/ICCV.2015.474

Zhou, G., Zhu, Z., He, T., & Hu, X. T. (2016). Cross-lingual sentiment classification with stacked autoencoders. *Knowledge and Information Systems*, *47*(1), 27–44. doi:10.100710115-015-0849-0

Zhu, Y., Gao, X., Zhang, W., Liu, S., & Zhang, Y. (2018). A bi-directional LSTM-CNN model with attention for Aspect-level text classification. *Future Internet.* Advance online publication. doi:10.3390/fi10120116

Zvarevashe, K., & Olugbara, O. O. (2018). A framework for sentiment analysis with opinion mining of hotel reviews. *Proceedings of the 2018 Conference on Information Communications Technology and Society (ICTAS)*. 10.1109/ICTAS.2018.8368746

ADDITIONAL READING

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Baltimore, Maryland: Association for Computational Linguistics. 10.3115/v1/P14-5010

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 *(Long Papers)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics. 10.18653/v1/N18-1202

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98–125. doi:10.1016/j.inffus.2017.02.003

Thongtan, T., & Phienthrakul, T. (2019). Sentiment Classification Using Document Embeddings Trained with Cosine Similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (pp. 407–414). Florence, Italy: Association for Computational Linguistics. 10.18653/v1/P19-2057

Zimbra, D., Abbasi, A., Zeng, D., & Chen, H. (2018). The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. *ACM Trans. Manage. Inf. Syst.*, 9(2), 5:1-5:29. doi:10.1145/3185045

KEY TERMS AND DEFINITIONS

Adversarial Learning: A learning paradigm based on two models attempting to achieve opposing goals.

Attention: A mechanism which allows a model to place additional emphasis on specific features.

Autoencoder: A network composed of an encoder and a decoder that can learn compact representations of its input data in a self-supervised manner.

Data Augmentation: A technique for improving the performance of a model by enriching the training data, e.g. by generating additional instances of minority classes.

Deep Learning: A form of machine learning which uses multi-layered architectures to automatically learn complex representations of the input data. Deep models deliver state-of-the-art results across many fields, e.g. computer vision and NLP.

Explainable AI: An emerging area of research whose goal is to make the decision-making processes of deep models understandable for humans.

Sentence Modelling: The task of converting a text into a representation that can be processed by a machine learning algorithm.

Sentiment Analysis: The task of discovering the underlying feelings expressed in a text. Methods are commonly classified by their scope, i.e. whether they consider aspects, sentences, or the entire document.

Transfer Learning: A collective term for machine learning techniques concerned with adapting a model across different domains and/or tasks.

Transformer: A type of deep model with an encoder-decoder structure that combines self-attention with feedforward networks.

ENDNOTES

[1] The code for Word2vec has been made publicly available at https://code.google.com/archive/p/word2vec/.

[2] The code for GloVe, along with pre-trained word vectors, is publicly available at https://github.com/stanfordnlp/GloVe.

[3] The code for fastText is publicly available at https://github.com/facebookresearch/fastText.

[4] The IMDB dataset is available at http://ai.stanford.edu/~amaas/data/sentiment/. It is also included in Tensorflow https://www.tensorflow.org/datasets/catalog/imdb_reviews.

[5] The Yelp dataset is available at https://github.com/zzhang83/Yelp_Sentiment_Analysis or in Tensorflow https://www.tensorflow.org/datasets/catalog/yelp_polarity_reviews.

[6] The SST dataset is publicly available at http://nlp.stanford.edu/~socherr/stanfordSentimentTreebank.zip.

[7] Implementations of both BERT-Base and BERT-Large are publicly available at https://github.com/google-research/bert.