



Natural language inference model for customer advocacy detection in online customer engagement

Bilal Abu-Salih^{1,5}  · Mohammed Alweshah² · Moutaz Alazab² · Manaf Al-Okaily³ · Muteeb Alahmari⁴ · Mohammad Al-Habashneh⁵ · Saleh Al-Sharaeh¹

Received: 7 March 2023 / Revised: 4 August 2023 / Accepted: 17 October 2023 /

Published online: 29 November 2023

© The Author(s) 2023

Abstract

Online customer advocacy has developed as a distinctive strategic way to improve organisational performance by fostering favourable reciprocal affirmative customer behaviours between the business and its customers. Intelligent systems that can identify online social advocates based on their social interaction and long-standing conversations with the brands are still lacking. This study adds to the burgeoning body of literature in this research area by developing a novel model to identify brand advocates using natural language inference (NLI) and artificial intelligence (AI) approaches. In particular, a hybridised deep learning model (BERT-BiLSTM-TextCNN) is proposed and adept at extracting the amount of entailment, contradiction, and neutrality obtained from the advocates' replies to the brands. This offers a new dimension to identify advocates based on the semantic similarities between the brands' tweets and customers' replies. The experimental results demonstrate the applicability of integrating the advantages of fine-tuned BERT, TextCNN, and BiLSTM using various evaluation metrics. Further, the proposed model is incorporated in a downstream task to verify and validate its effectiveness in capturing the correlation between brands and their advocates. Our findings contribute to the burgeoning body of literature in this research area and have important implications for identifying and engaging with brand advocates in online customer engagement.

Keywords Customer advocacy · Online customer engagement · Natural language inference · BERT · TextCNN · BiLSTM

1 Introduction

Online social networks (OSNs) have rapidly grown over the last decade (Castillo López et al., 2021). Their participatory nature encourages engaging and meaningful ties (Saldanha et al., 2020), facilitates dialogue between consumers and businesses, involves customers in the development of content and value (Goh & Arenas, 2020), and fosters long-lasting relationship exchanges (Castillo et al., 2021). Therefore, online Customer Engagement

Editors: Dino Ienco, Roberto Interdonato, Pascal Poncelet.

Extended author information available on the last page of the article

(CE) has become vital to a more comprehensive business success strategy (Han & Anderson, 2022). Online CE aims to understand the online customer's attitude toward a brand or company that extends beyond buying behaviour (Oh et al., 2017; Van Doorn et al., 2010), thereby providing them with a tailored experience that encourages them to become evangelists (Jacob & Johnson, 2021) by means of OSNs. Evangelists refer to customers who are so passionate about a brand or product that they actively promote it to others, often without any incentive or compensation from the brand. In the context of online advocacy, becoming an evangelist means that a customer becomes an online advocate by enthusiastically recommending a product or service to others through various online channels such as social media, reviews, or forums. Online advocates can become evangelists by sharing their positive experiences with a brand or product online, which can influence others to become customers as well. These customers may become brand loyalists and continue to promote the brand or product to others, creating a cycle of advocacy and evangelism. In essence, becoming an evangelist is the ultimate goal of online advocacy. It represents a level of customer loyalty and satisfaction that goes beyond mere satisfaction with a product or service, and instead reflects a deep connection with a brand or product that inspires customers to share their positive experiences with others (Kartajaya et al., 2016; Kim & Hwang, 2022). Therefore, businesses pay a great deal of attention to engaged customers and attempt to establish means to capture positive customer feedback (Lee & Kim, 2022; Nasr et al., 2017), thereby converting such happy consumers into brand advocates and gaining new consumers simultaneously (Song & Kim, 2022) (Mansoor & Paul, 2022).

By encouraging favourable reciprocal affinitive customer behaviours between the company and its consumers, online customer advocacy has evolved as an alternative strategic method to generate improved organisational performance results (Dutta et al., 2021; Malesev & Cherry, 2021). This implies that organisations succeed far more when they assist customers in locating and obtaining higher value in a market transaction. Customer advocacy is perceived as positive word-of-mouth (WOM) by which a consumer becomes exceptionally eager to enthusiastically recommend a product or service to other customers (Al-Okaily & Al-Okaily, 2022; Alsaad & Al-Okaily, 2021; Sweeney et al., 2020). Modelling online customer advocacy has been established in the literature by developing frameworks to extract engagement patterns (Potdar et al., 2018), explore positive and negative valence brand engagement (Hollebeek & Chen, 2014), formulate advocacy return on investment (Hollebeek et al., 2021), identify factors affecting online CE (Read et al., 2019), etc. Online advocacy is of the most crucial results of CE, yet relatively few empirical studies have examined the motivations behind consumer advocacy behaviours (Walz & Celuch, 2010). Further, there is a lack of intelligent systems that are able to identify online advocates based on their social interactivity and established dialogues with the brands. This study contributes to this flourishing literature stream by developing a novel model to identify brand advocates using natural language inference (NLI) and artificial intelligence (AI) techniques.

A significant contribution of AI research has been the development of practical inference and reasoning techniques. In particular, formal reasoning problems have benefited from using reasoning techniques and the knowledge bases that enable them (Wang et al., 2019a, 2019b). It is necessary to go beyond what Natural Language Processing (NLP) currently does in the age of AI, Deep Learning, and Big Data. Therefore, expanding NLP to do a wide range of tasks necessitates a variety of specialised techniques. For example, the ability to infer the links between various natural language claims is necessary for an AI system that aspires to comprehend and reason about natural language. NLI is an NLP task that assesses whether one statement implies or contradicts the other for a given set of

sentences to satisfy this demand. More specifically, NLI is frequently framed as a classification problem: given two statements, hypothesis and counter-hypothesis, the challenge is to categorise their connection into one of three categories: "entailment," "contradiction," or "neutral" (Bowman et al., 2015). Various applications have been created by syllogistic reasoning based on the facts that currently exist using NLI datasets. Examples of these applications include fake news detection (Sadeghi et al., 2022), question answering (Mishra et al., 2021), recommender systems (Zhu et al., 2021), opinion summarisation (Kim & Zhai, 2009), finance (So et al., 2022), healthcare (Herlihy & Rudinger, 2021), and politics (Mukherjee et al., 2022).

The broader social exchanges between businesses and their consumers are not thoroughly scrutinised, making it challenging to identify advocates from the textual inference drawn from these conversations. This study aims to apply the notion of NLI to these dialogues to measure the inferred textual entailment the advocates convey in their replies to the brand's tweets. Therefore, this article devises a novel technique that incorporates the notion of NLI, thereby obtaining a better understanding of these dialogues and better grasping the implied linkages between brand tweets and advocates' replies. In particular, a hybridised deep learning model (BERT-BiLSTM-TextCNN) is proposed that incorporates bidirectional encoder representations from transformers (BERT), bidirectional long short-term memory (BiLSTM) and text convolution neural network (TextCNN). This integrated model is designed to furnish a textual inference model that is adept at extracting the amount of entailment, contradiction, and neutrality obtained from the advocates' replies to the brands. This offers a new dimension to identify advocates based on the semantic similarities between the brands' tweets and the customer replies.

The experimental findings demonstrate that the hybridised model presented in this research can more effectively combine the benefits of TextCNN and BiLSTM; it captures local correlation while keeping context information and has performed well in various evaluation metrics. This is consolidated by integrating a fine-trained BERT model, which demonstrates superiority over state-of-the-art textual embedding models by capturing better contextual semantics comprehension. The proposed BERT-BiLSTM-TextCNN model has been applied to a downstream task in which the model's utility to capture the correlation between advocates and the brands is demonstrated and verified.

This paper is organised as follows: Sect. 2 provides a theoretical background on the paper's proposed topic. The methodology is discussed in Sect. 3. Section 4 demonstrates the utility of this paper via several evaluation measures. Section 5 discusses the experimental results and Sect. 6 concludes the article.

2 Theoretical background

2.1 Customer engagement and advocacy

The advent of social media and its potential for connecting with consumers and fostering relationships has had a revolutionary influence on communication that has thrilled marketing scholars and practitioners across the world and sparked a lot of interest in the notion of CE (Sashi et al., 2019). The Economist defines CE as an intimate long-term relationship between a seller and a customer (Unit, 2007). According to Van Doorn et al. (2010), these actions, which include word-of-mouth (WOM), reviews, recommendations, and ratings, are known as CE behaviours. The term WOM refers to the

favourable testimonials, remarks, and conversations customers have about a company's name, products, and services (Al-Okaily, 2021; Eisingerich et al., 2015). Positive word-of-mouth marketing (PWOM) is a much more intricate and multifaceted tactic than it may initially appear to be (Sun et al., 2021). The ability of positive reinforcement is an evident growth advantage, serving as both a promotional tool and a customer attitude survey. Advocacy is a unique instance of WOM since it is always beneficial and effective when customers are devoted and satisfied (Sashi, 2012). Online positive customer feedback is crucial for brand success but indicating brand advocates and gaining new consumers simultaneously requires deliberate action. Therefore, brand advocacy is characterised in the literature as the readiness to use new products and services, to promote them to others, but, most importantly, as the readiness to overlook mistakes made by the brand (Bilro et al., 2019).

2.2 Natural language inference

Natural language is characterised by various semantic expressions, which allow the same meaning to be communicated by or inferred from many texts. The many-to-many mapping between language expressions and meanings that results from this phenomenon may be seen as the dual problem of linguistic ambiguity. In order to detect that a particular intended meaning may be inferred from multiple text variations, a model for this variability phenomenon is required for many NLP applications. In Natural Language Inference (a.k.a Recognising Textual Entailment (RTE)), two sentences are given, namely "premise" and "hypothesis". We must determine if the offered hypothesis is True (i.e., two sentences convey the same meaning), False (i.e., two sentences share no similarity), or unrelated to the premise (Dagan et al., 2010). True indicates entailment, False indicates contradiction, and Undecided/unrelated indicates neutrality. This can be comprehended by using the examples below:

- Entailment: sentence1="A tennis game with multiple players playing", sentence2="Some people are playing a sport"
- Contradiction: sentence1="Two children are observing the moon together", sentence2="Two children are standing with their eyes shut"
- Neutral: sentence1=" An umbrella is being held by a happy costumed woman", sentence2=" A cheerful woman with an umbrella is dressed as a fairy".

There are well-known standard annotated benchmark datasets corpora dedicated to NLI tasks. Few state-of-the-art models trained on these collections acquire decent accuracy. These datasets include Stanford Natural Language Inference (SNLI) corpus, Multi-Genre Natural Language Inference (MultiNLI) corpus, SciTail, and Adversarial Natural Language Inference (ANLI). The description of these collections can be found in Table 1.

NLI is utilised in a variety of industries, including finance, retail, and other sectors (Araci, 2019; Pillai et al., 2022). It is frequently employed when it is necessary to determine if created or acquired end-user results adhere to the hypothesis. Tasks for automated auditing are included in one of the use cases. By deciding if sentences in the created document are consistent with those in the reference texts, NLI can partially replace human auditing.

Table 1 A summary of commonly used NLI annotated datasets

Dataset	Reference	Description	Size
SNLI	Bowman et al. (2015)	A dataset of sentence pairs labelled for entailment, contradiction, and semantic independence	570 k pairs
MultiNLI	Williams et al. (2017)	A collection of hypothesis/premise pairs including a range of genres of spoken and written text and supports cross-genre evaluation	433 k pairs
SciTail	Khot et al. (2018)	This dataset is constructed using phrases that already exist "in the wild," as opposed to the SNLI and MultiNLI, which were crowdsourced. Science questions and the related response candidates served as the basis for the creation of hypotheses, while relevant web phrases from a huge corpus served as the premises	27 k pairs
ANLI	Nie et al. (2019)	ANLI is a brand-new, sizable NLI benchmark dataset that was created using an adversarial, repetitive, human-and-model-in-the-loop method. SNLI and MNLI were easier than ANLI, which is significantly more challenging. There are three rounds in it. Train/dev/test divides are present for each round	163 k pairs

2.3 Related works

OSNs are no longer a passing trend. Every day, millions of messages from customers voice their thoughts on businesses using these virtual means, which has steadily changed how these brands interact with their customers and do business. In fact, keeping up with consumers' opinions of a company may be facilitated by rapid analysis of customer messages on various platforms (Yang et al., 2019). Brands have faithfully followed the trend of interacting and communicating with consumers on OSNs platforms in past years. Customers' comments on these platforms may transmit important messages, which is essential for businesses to build strong customer relationships (Ibrahim & Wang, 2019). Therefore, various studies have been proposed to tackle the ever-increase in social customer-brand interactions. One research direction addressed the need for guidance on improving CE (Bansal et al., 2022; Matosas-López & Romero-Ania, 2021; Simon & Tossan, 2018). Amongst these attempts, AI has been incorporated to benefit from the continuous propagation of social data. For example, Perez-Vega et al. (2021) intended to provide a conceptual framework that explains how businesses and consumers may improve the results of both firm-solicited and firm-unsolicited online customer interaction behaviours. The authors indicated various forms of firm-solicited and firm-unsolicited online CE behaviours that serve as stimuli for AI to analyse customer-related information and result in both AI and human responses, which, in turn, shape the contexts of future online CE. In the same direction, Gao Li, et al. (2022) examined the impacts of AI stimuli on CE and value co-creation. Additionally, the moderating effect of consumer ability preparation is looked at.

The incorporation of sentiment analysis to better understand customers' behaviour has been broadly examined in the literature (Abu-Salih et al., 2021, 2022; Wongthontham & Abu-Salih, 2018). Businesses may use sentiment analysis to learn how customers feel about their products, services, and company. Additionally, digital evaluations that include ratings and textual details on a range of topics give prospective customers crucial information about a product or service usage. For example, Jain et al. (2022) developed a multi-label ensemble forecasting model to predict recommendations using machine learning techniques. Various classifiers were utilised in their approach, and their performance was reported and compared. Five sentiment classes, including highly negative, negative, neutral, positive, and strongly positive, and three polarity characteristics, including a verb, adverb, and adjective, were proposed by Kausar et al. (2019) as a strategy for categorising the sentiment polarity of online product evaluations. A hybrid CNN-LSTM model was suggested by Rehman et al. (2019) to enhance the performance of sentiment analysis on movie reviews. In their study, Dashtipour et al. (2021) focused on Persian film reviews, the official language of Iran and Afghanistan. The results showed that the stacked-BiLSTM model beats all other techniques when they compared the performance of shallow learning algorithms with deep learning algorithms.

Modelling online customer advocacy has been established in the literature by developing frameworks to extract engagement patterns (Potdar et al., 2018), explore positive and negative valence brand engagement (Hollebeek & Chen, 2014), formulate advocacy return on investment (Hollebeek et al., 2021), identify factors affecting online CE (Read et al., 2019), etc. However, there is a lack of thorough scrutiny of the overall social conversations between brands and their customers, thereby detecting advocates based on the textual inference obtained from these dialogues. Therefore, this paper proposed a novel approach that incorporates NLI to better understand the inferred relationships between the brands' tweets and customers' replies. Next section discusses the methodology proposed in this study.

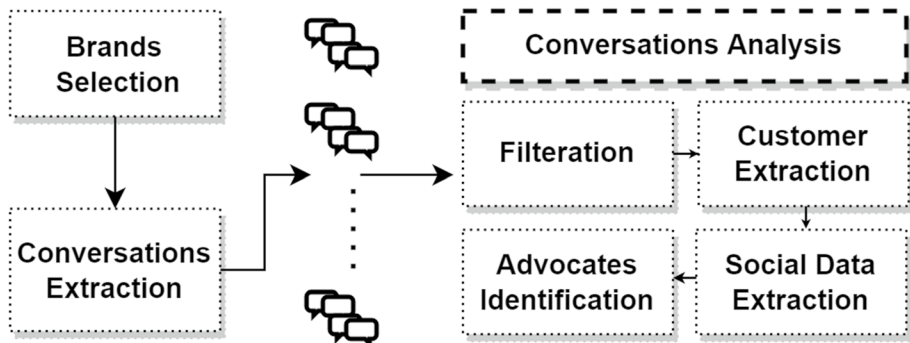


Fig. 1 Steps followed to extract social data & metadata for brands and customers

3 Method

This section discusses the proposed methodology. This methodology comprises two main stages, namely data preparation and model development. The following subsections discuss each of these stages.

3.1 Data preparation

3.1.1 Social conversation analysis

Since Twitter’s launch in 2006, it has offered a rich data collection of more than 500 million tweets per day or around 200 billion tweets annually. Also, providing customer service using this designated platform has proven significant, as reported by 85% of small and medium businesses (SMB) (Sayce, 2022). This implies the necessity to continuously implement sophisticated tools to understand customers, thereby obtaining the hoped-for added value. Therefore, the social dataset used in this study targets online CE via the Twitter platform. Figure 1 illustrates the steps followed to extract social data & metadata for brands and customers. In particular, we select 14 Twitter accounts that belong to official Australian brands. The tweets of these accounts are collected using Twitter APIs, and these tweets are analysed to extract the conversations between the brand and the customers. These conversations are filtered to eliminate poor and limited conversations. This ensures that the proposed model will be carried out on customers with a proper brand dialogue. The Twitter accounts of those customers, and thus their social data (tweets) and metadata (accounts information), are obtained. The contents of those customers are analysed manually to indicate advocate customers who exhibit positive opinions (recommendations) toward the brand, namely who recommend the brand’s product or service via practising positive WOM in their tweets.

3.1.2 Social data preprocessing

Data pre-processing is often applied at the beginning of an ML and AI development pipeline to prepare the data into a format that can be processed more quickly and efficiently. The collected tweets from the previous stage are pre-processed using the following activities: (i) data cleansing: This is the process of replacing missing values and removing erroneous,

incomplete, duplicate, and incorrect data from the collected tweets; (ii) aggregation: This process entails aggregating the curated tweets of both the brands and customers in one data silo; (iii) data reduction: In this process, we eliminate some of the historical and outdated tweets which do not depict the recent behaviours of customers; (iv) transformation: This process aims to transform the collected tweets and metadata into a tabular format namely the MySQL format; (v) tokenisation: Tokenisation is a technique used in natural language processing to break down phrases and paragraphs into simpler language-assignable elements (words/tokens). We incorporate WordPiece to tokenise words (Song et al., 2020).

3.2 Model development

This section discusses different modules that are incorporated so as to extract relevant features that will be used in the prediction module.

Figure 2 illustrates these modules and the set of supportive technologies and datasets that are used in these modules which are summarised as follows.

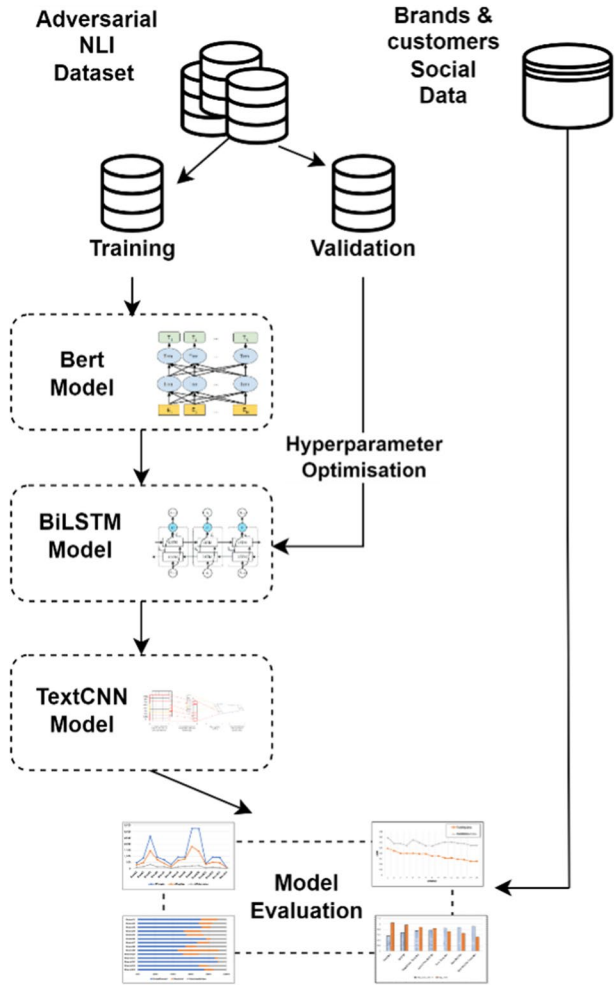
3.2.1 BERT models

BERT stands for Bidirectional Encoder Representation of Transformers. It is an unpervised deep learning-based language representation model created by Google AI (Devlin et al., 2018). The primary technological advancement of BERT is the application of Transformer's bidirectional training to language modelling. In the majority of NL tasks, transformer models have demonstrated outstanding outcomes (Wang et al., 2019a, 2019b; Wolf et al., 2019). Modern NLP has adopted the power of transfer learning together with large-scale transformer language models as a standard. In contrast, earlier research looked at text sequences from either a left-to-right or a mixed left-to-right and right-to-left training perspective. BERT demonstrates that bidirectionally trained language models can comprehend the context and flow of language more deeply than single-direction language models. BERT has stirred up controversy in the ML field by offering cutting-edge findings across a wide range of NLP tasks, such as text classification, natural language inference (NLI), and others. We incorporate four different transformer-based pre-trained models in this study, namely $BERT_{Base}$, $BERT_{Large}$, $BERTTweet_{Base}$, and $BERTTweet_{Large}$ which are described in Table 2.

3.2.1.1 Fine-tuning BERT model for NLI Task The BERT neural network architecture has a sizable number of parameters—between 100 million and more than 300 million—and is very complex. Consequently, overfitting would occur if a BERT model were trained from the start on a limited dataset (such as our collected dataset). As a result, it is preferable to employ a pre-trained BERT model that was trained on a sizable dataset. The model can then be refined by running additional training on our comparatively smaller sample. To fine-tune the BERT model for our NLI task, we freeze all the layers of the incorporated BERT model and attach a dense layer and a softmax layer to the architecture and train this new model. Figure 3 demonstrates the proposed approach.

To feed our classifier, a single vector must represent the whole input text. In BERT, the choice is made to use the initial token's concealed state to represent the entire phrase. An extra token must be manually added to the input phrase in order to do this. For this, the token [CLS] is selected in the original implementation. A phrase of a specific length is provided as input to the BERT model. The size of a phrase is often

Fig. 2 Proposed methodology



determined by the facts we are dealing with. We will have to add paddings (empty tokens) to the sentences if they are less than this maximum length to make up the length. To express sentence padding, the token [PAD] is used. Each token received a different ID when the BERT model was trained. Therefore, in order to employ a pre-trained BERT model, each token in the input phrase must first be converted into its unique matching ID. The intended replacement for tokens not found in the original vocabulary is a special token called [UNK], which stands for unknown token. The input data will lose a great deal of information if all unseen tokens are changed to [UNK]. As a result, and as indicated in the data pre-processing, WordPiece method is used, which divides a word into numerous subwords, allowing the model to represent frequently occurring subwords additionally. These steps of text preparation are illustrated in Fig. 4.

Table 2 A summary of incorporated Pretrained BERT Models

Model	Reference	Model description	Architecture		Dataset description
			#Param	Max length	
BERT _{Base}	Devlin et al. (2018)	Pretrained model on English language using a masked language modeling (MLM) objective	110M	128	The dataset of 11,038 unpublished books from BookCorpus and English Wikipedia served as the pretraining data (excluding lists, tables and headers)
BERT _{Large}			340M	512	
BERT _{tweet_{Base}}	Nguyen et al. (2020)	large-scale pre-trained language model for English Tweets	135M	128	English Tweets streamed from 01/2012 to 08/2019 and 5 M Tweets associated to the COVID-19 pandemic
BERT _{tweet_{Large}}			355M	512	

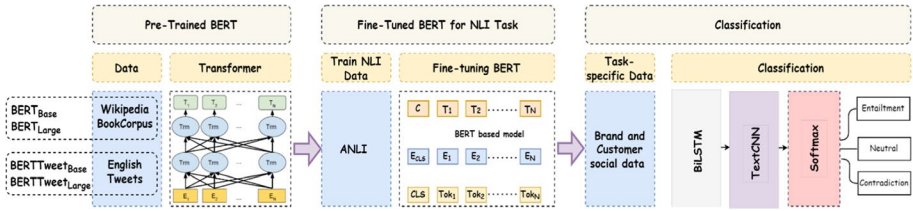


Fig. 3 Fine-tuning of BERT model for NLI task



Fig. 4 Sentence preparation for BERT models

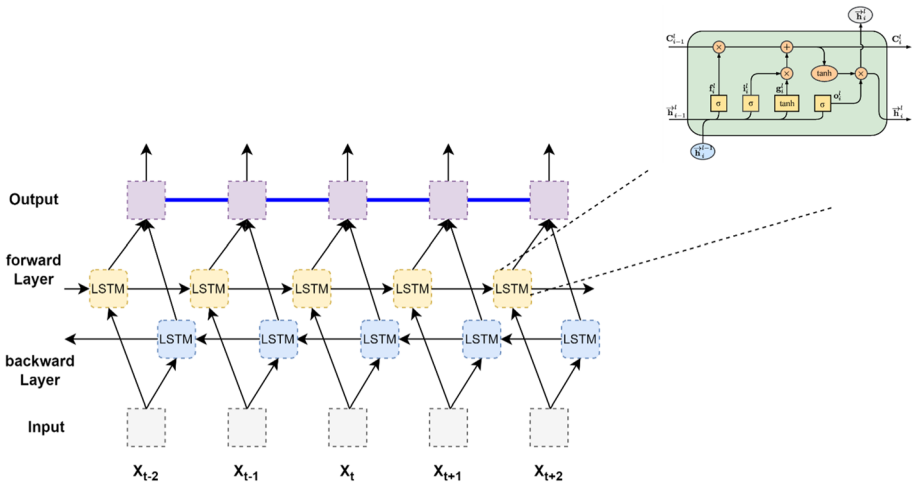


Fig. 5 Schematic representation of BiLSTM

3.2.2 BiLSTM model

A Bidirectional Long Short-Term Memory, or BiLSTM (Graves & Schmidhuber, 2005), is a sequence processing model that consists of two LSTMs, one of which receives input forward and the other of which receives it backward. With the help of BiLSTMs, the network can access more information, which benefits the algorithm’s context (e.g., telling what words immediately follow and lead a word in a particular sentence). In practice, the architecture of BiLSTM excels, particularly in NLP. The primary justification is that each element of an input sequence contains data from the past and the present. For this reason, by integrating LSTM layers from both directions, BiLSTM can generate more meaningful output. Further, because of its LSTM features, the BiLSTM prevents gradient vanishing that commonly occurs in RNN. The same LSTM equations are used in both forward and backward LSTM networks.

As illustrated in Fig. 5, the forward hidden layer and the reverse hidden layer are two separate hidden layers in BiLSTM. The input is taken into account by the forward hidden layer

h_t^f in ascending order, i.e., $t = 1, 2, 3, \dots, T$. The backward hidden layer h_t^b , on the other hand, takes into account the input in decreasing sequence, i.e., $t = T, \dots, 3, 2, 1$. Finally, output y_t is produced by combining h_t^f and h_t^b . The implementation of BiLSTM can be described in the following equations:

$$h_t^f = \tan h\left(w_{xh}^f x_t + w_{hh}^f h_{t-1}^f + b_h^f\right) \quad (1)$$

$$h_t^b = \tan h\left(w_{xh}^b x_t + w_{hh}^b h_{t-1}^b + b_h^b\right) \quad (2)$$

$$y_t = w_{hy}^f h_t^f + w_{hy}^b h_t^b + b_y \quad (3)$$

The embedding result from BERT is used as an input vector by BiLSTM layer to extract sentence characteristics. The ultimate output [Hf, Hb] of the hidden state of BiLSTM will be the concatenation of the forward LSTM h_t^f and backward LSTM h_t^b networks.

3.2.3 TextCNN

A variation of the CNN model is the TextCNN model. It allows CNN to fully utilise its parallel processing capabilities, and training is completed more quickly. The ability to extract text features is added, along with the properties of the original CNN. TextCNN, which has a great capacity to extract shallow text features, employs one-dimensional convolution to obtain the n-gram feature representation of the phrase. The task's linguistic n-gram can be recognised by the TextCNN model. Each layer of the hierarchical structure's convolution structure can allow n-grams with comparable components to share the expected behaviour when it comes across a specific n-gram that has not been registered throughout the prediction process. CNN considers the longer n-grams in the text so that the model can respond to non-continuous n-grams more effectively. The capacity of TextCNN to flexibly process different temporal information from the extensive vocabulary is enhanced by altering the height of the convolution kernel. There are no structural modifications between TextCNN and the conventional image CNN network (or even simpler). In actuality, TextCNN merely uses one layer of convolution, one layer of maxpooling, and ultimately connects the output to softmax for n classification.

Since TextCNN uses multiple convolution kernels of various sizes to extract key information from sentences (similar to n-grams with multiple window sizes), which can better capture local correlation, and BiLSTM can capture text information from both the forward direction and the reverse direction at the same time, this paper combines the two models. The obtained word vector was fed into the BiLSTM layer for bidirectional learning, the outcome was fed into the convolutional layer, and feature extraction was carried out. They then go through the pooling layer while keeping the essential features, drastically cutting into the number of parameters and overfitting. Finally, the output fully connected layer is used to combine the results and output the NLI result.

4 Experimental results

In this section, we discuss how we evaluate our proposed system. First, we measure the utility of the BERT-BiLSTM-TextCNN model. In particular, we compare similar findings produced with BERT using the datasets presented with their corresponding best prediction performances from the literature. We begin by outlining the BERT fine-tuning settings and providing specifics on the requirements for the tests. The outcomes of the experiment will then be presented and explained. In the next experiment, we will examine the proposed model on a downstream task based on the collected social data. All data used in this study have been stored, processed and analysed using the Pawsey Supercomputing Research Centre facilities.¹

4.1 Evaluation metrics

To assess the prediction accuracy and generalizability of BERT-BiLSTM-TextCNN model, certain evaluation metrics are incorporated. Assuming TP as true positive, FP as false positive, TN as true negative, and FN as a false negative. Precision, Recall and F-measure are used as a quantitative evaluation of the prediction performance of the models, where $precision = \frac{TP}{(TP+FP)}$, $Recall = \frac{TP}{(TP+FN)}$, and $F1 - measure = 2 \times \frac{(Precision \times Recall)}{Precision + Recall}$. One of the objectives of this study is to evaluate the proposed model for detecting the NLI inferred from two sentences. Hence, let's assume i is the category of NLI ($i=1, 2, 3$ where 1, 2, 3 denote to entailment, neutral, and contradiction, respectively). One binary classifier will be created for each category, and the effectiveness of each classifier will be evaluated using each of the classification metrics (such as Precision). For example, there are three Precision scores as a consequence. We need a mechanism to express Precision across all classes since we only need one precision score, not three, to compare one model to another. The use of averaging methods is now necessary. Three specific averaging methods are commonly used for multiclass classification:

- **Macro-average:** This is simple arithmetic used to calculate the average of all metrics for all classes. This method offers each class the same weight, making it a viable choice for tasks requiring balanced categorisation. The macro-average for the incorporated metrics can be computed as:

$$Precision - macro(P^{mac}) = \sum_{i=1}^n \frac{Precision^i}{n} \quad (4)$$

$$Recall - macro(R^{mac}) = \sum_{i=1}^n \frac{Recall^i}{n} \quad (5)$$

$$F - measure - macro(F^{mac}) = 2 \times \frac{P^{Mac} \times R^{Mac}}{P^{Mac} + R^{Mac}} \quad (6)$$

¹ <https://pawsey.org.au/> (accessed 25/09/2022).

- Weighted-average:** By calculating the average of binary metrics weighted by the number of samples from each class in the target, weighted adjusts for class imbalance. When three classes have accuracy scores of 0.85 for class 1, 0.80 for class 2, and 0.89 for class 3, the weighted average will be determined by multiplying each score by the proportion of instances of each class and dividing by the total number of samples. The weighted average for the incorporated metrics can be formulated as:

$$Precision - weighted(P^W) = \frac{\sum_{i=1}^n (|y_i| \times Precision^i)}{\sum_{i=1}^n (|y_i|)} \tag{7}$$

$$Recall - weighted(R^W) = \frac{\sum_{i=1}^n (|y_i| \times Recall^i)}{\sum_{i=1}^n (|y_i|)} \tag{8}$$

$$F - measure - weighted(F^W) = \frac{\sum_{i=1}^n (|y_i| \times F - measure^i)}{\sum_{i=1}^n (|y_i|)} \tag{9}$$

where y is the set of predictions.

- Micro-average:** Micro-averaging calculates the percentage of adequately categorised observations among all observations. This metric is used to determine total correctness and can be calculated as follows:

$$Precision - micro(P^{mic}) = \frac{\sum_{i=1}^n TP^i}{\sum_{i=1}^n TP^i + \sum_{i=1}^n FP^i} \tag{10}$$

$$Recall - micro(R^{mic}) = \frac{\sum_{i=1}^n TP^i}{\sum_{i=1}^n TP^i + \sum_{i=1}^n FN^i} \tag{11}$$

$$F - measure - micro(F^{mic}) = 2 \times \frac{P^{mic} \times R^{mic}}{P^{mic} + R^{mic}} \tag{12}$$

- Cohen’s Kappa (K):** A version of accuracy that has been turbocharged and includes metrics of chance and class imbalance. It is used to assess inter-rater reliability as well as intra-rater reliability. Since it considers the potential that the agreement may have happened by chance, it is typically believed to be a more reliable measurement than a simple % agreement estimate. Cohen’s Kappa can be computed as:

$$k = \frac{P_0 - P_e}{1 - P_e} \tag{13}$$

where P_0 is the proportionate agreement between actual and expected values. This would be the sum of any confusion matrix’s diagonal cells divided by the sum of its non-diagonal cells. P_e is the likelihood that correct values and false values will coincide by accident.

- Matthew’s correlation coefficient (MCC):** A statistical tool for evaluating models. Its function, which is equal to chi-square statistics for a 2×2 contingency table, assesses or measures the difference between the expected and actual values. It is formulated as:

Table 3 Performance evaluation of BERT models on ANLI

	Model			
	BERT _{Base}	BERT _{Large}	BERTTweet _{Base}	BERTTweet _{Large}
Evaluation metric				
Macro average				
P^{mac}	0.684	0.894	0.714	0.876
R^{mac}	0.692	0.850	0.764	0.891
F^{mac}	0.688	0.870	0.738	0.883
Weighted average				
P^W	0.600	0.893	0.764	0.915
R^W	0.600	0.873	0.774	0.905
F^W	0.590	0.883	0.769	0.910
Micro average				
P^{mic}	0.715	0.861	0.694	0.911
R^{mic}	0.694	0.843	0.716	0.921
F^{mic}	0.704	0.852	0.705	0.916

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (14)$$

4.2 BERT fine-tuning performance results

As indicated in the methodology section, ANLI corpus is used to measure the performance of the BERT-BiLSTM-TextCNN model. In the fine-tuning process of BERT, we follow the recommendations of the BERT’s designers (Devlin et al., 2018). In particular, it is recommended to keep all the parameters constant except the batch-size, learning rate (Adam), and epoch-size. The advised values for these designated parameters are Batch-size: 16,32, Learning Rate: 5e-5, 3e-5, 2e-5, and Epoch-size: 2,3,4. The fine-tuning dataset is split into training (80%) and validation (20%), whereby the training set is used to train the model and establish its learning parameters and the validation set is used to identify the optimised hyperparameters for both the BiLSTM layer and TextCNN layer. The learning rate is set to 0.001, and the number of training cycles is set to 15 when configuring the BiLSTM layer’s settings. The batch size is set to 300 given a large number of texts, Adam is utilised as the optimisation function to hasten convergence, and cross entropy is employed as the loss function.

In the TextCNN layer, after identifying the best single convolution kernel, it is decided that using a mixture of convolution kernels (3, 4, and 5) to aggregate the findings is preferable than using a single best convolution kernel. There are 300 convolution kernels in all, with each kernel having a count of 100. Dropout is adjusted to 0.5 to lessen the chance of overfitting. The 1-max pooling technique yields superior results, the activation function employs the ReLU function, and the L2 regularisation is set to 3. The classification performance of the BERT models is seen in Table 3. As illustrated in the table, BERTTweet_{Large} outperforms all other embedding model using ANLI dataset. This can be understood due to the robustness of the mechanism followed to create the dataset. In particular, annotators were incorporated as adversaries and urged to identify flaws that would cause the model to

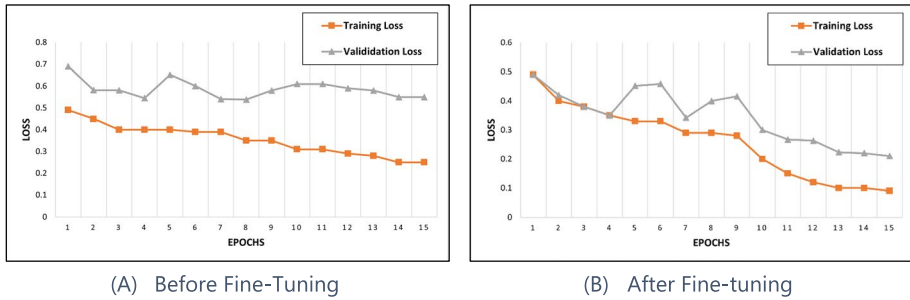


Fig. 6 Training Loss vs Validation Loss comparison where (A) is the loss per epoch before fine tuning and (B) is the loss per epoch after fine-tuning

incorrectly categorise something that a human would classify adequately. The authors discovered that non-expert annotators are incredibly inventive at identifying and taking advantage of flaws in this gamified environment and with the right incentives. Three rounds were gathered, and each round's test sets grew more challenging and the models more reliable as the rounds went on. The state-of-the-art on the current NLI benchmarks was produced by training on new data.

For a total of 15 rounds, the training set and validation set in this study are both defined by loss before and after tuning. Figure 6A. B illustrate the loss comparison of the training set and validation set before and after tuning, respectively. After adjusting, the training set and validation set's loss values are smoother, converge more quickly, and the model works better.

4.3 Baseline comparison

The proposed BERT-BiLSTM-TextCNN is compared with certain baseline methods to verify its utility, including TextCNN, BiLSTM, Bert-TextCNN, and Bert-BiLSTM. This study aims to evaluate the effectiveness and performance of the proposed BERT-BiLSTM-TextCNN model for the task at hand. To ensure a comprehensive evaluation, we compared our model with several relevant baseline methods commonly used in text classification and sentiment analysis tasks. The inclusion of TextCNN, BiLSTM, Bert-TextCNN, and Bert-BiLSTM as baselines serves multiple purposes. Firstly, TextCNN and BiLSTM are widely recognized and established models in natural language processing (NLP) tasks. They provide a benchmark for traditional neural network architectures that have been extensively studied and widely used in various text classification tasks. Additionally, considering the popularity and success of transformer-based models like BERT in recent NLP research, we included Bert-TextCNN and Bert-BiLSTM as baselines. These models combine the power of pre-trained language representations with convolutional or recurrent neural networks to capture both local and global contextual information. We will certainly consider incorporating more recent and complex approaches in future work as baselines for comparison. Our goal is to continue to improve the accuracy and effectiveness of our model in identifying online advocates, and we appreciate your feedback in helping us to achieve this. Also, we incorporate Word2Vec technique which employs a neural network model to learn word embeddings, thereby obtaining meaningful relationships and converting the similarity into a vector. The idea of relatedness between words or

Table 4 A comparison with baseline models using the collected dataset

Model		P	R	$F1$	F^{mac}	F^{mic}	K	MCC
TextCNN	Entailment	0.49	0.66	0.57	0.39	0.44	0.078	0.083
	Neutral	0.52	0.57	0.54				
	Prediction	0.26	0.04	0.07				
BiLSTM	Entailment	0.52	0.60	0.56	0.43	0.48	0.207	0.213
	Neutral	0.52	0.64	0.57				
	Prediction	0.37	0.11	0.17				
Word2Vec- TextCNN	Entailment	0.58	0.67	0.62	0.52	0.53	0.308	0.312
	Neutral	0.59	0.63	0.61				
	Prediction	0.47	0.25	0.33				
Word2Vec-BiLSTM	Entailment	0.61	0.67	0.64	0.60	0.58	0.358	0.362
	Neutral	0.61	0.69	0.65				
	Prediction	0.54	0.29	0.38				
BERT-TextCNN	Entailment	0.74	0.65	0.70	0.71	0.71	0.49	0.494
	Neutral	0.68	0.72	0.70				
	Prediction	0.70	0.74	0.72				
BERT-BiLSTM	Entailment	0.71	0.67	0.69	0.72	0.74	0.579	0.579
	Neutral	0.70	0.65	0.67				
	Prediction	0.75	0.82	0.78				
BERT-BiLSTM-TextCNN	Entailment	0.80	0.76	0.78	0.78	0.79	0.677	0.677
	Neutral	0.77	0.78	0.78				
	Prediction	0.79	0.81	0.80				

items, such as semantic relatedness, synonym recognition, concept classification, selection preferences, and analogies, is extracted using the Word2Vec model. We incorporate the embeddings obtained by Word2Vec and train them using TextCNN (Word2Vec-TextCNN) and BiLSTM (Word2Vec-BiLSTM) as baseline methods for the NLI task. The following parameter settings are set for this experiment; Embedding size = 300, Input length = 512, optimise = Adam, Learning rate = 0.01, epoch = 15, and loss-function = cross-entropy. Table 4 demonstrates the obtained evaluation metrics from the baseline models including P , $RF1$, F^{mac} , F^{mic} , K , and MCC . As depicted in the table, the benchmark comparison verifies the utility of the proposed BERT-BiLSTM-TextCNN over other models. Micro-average of F1 score measures the overall quality of a multi-classifier model since it computes the F1-score of the accumulated contributions of all classes. It can also be seen as the overall likelihood of true positive classifications. F^{mic} score that is measured for the designed BERT-BiLSTM-Text attains the highest value which indicates the model capacity to emphasise on the common labels in the given dataset. Unlike micro F1-score, macro F1-score achieves low values for models that only excel in popular classes while utterly failing at uncommon ones. As illustrated in Table 4, most of the baseline models perform poorly in terms of both F^{mac} and F^{mic} .

When evaluating the effectiveness of ML classification models, K score is used to compare the perfect agreement and agreement by chance between the real-world observations and the classifier's predicted values. K value can be interpreted as follows: values 0 denote "no agreement", 0.01–0.20 indicate "no to slight agreement", 0.21–0.40 indicate "fair

agreement”, 0.41–0.60 indicate “moderate agreement”, 0.61–0.80 indicate “substantial agreement”, and 0.81–1.00 denote “almost perfect agreement” (Cohen, 1960). By applying this interpretation on the baseline comparison results, it implies that TextCNN model demonstrates slight agreement. BiLSTM, Word2Vec-TextCNN, and Word2Vec-BiLSTM reveal fair agreement. Both Bert-TextCNN and Bert-BiLSTM show moderate agreement, and the proposed model exhibits substantial agreement between the actual observations and predicted ones.

MCC is a more dependable statistical measure that only yields a high score if the prediction performed well in each of the four categories of the confusion matrix (true positives, false negatives, true negatives, and false positives), proportionally to the size of the dataset’s positive and negative samples. The *MCC*’s range of values is from -1 to $+1$. An ideal model has a score of 1, whereas an imperfect model has a score of -1 . One of the main benefits of *MCC* is its quality since it makes it simple to comprehend. As can be observed in Table 4, BERT-BiLSTM-TextCNN model verifies its utility by overshadowing other baseline models in this metric as well.

To offer a deeper insight into the performance of each baseline model in a comparison with the proposed model, AUC-ROC curve is extracted for each model. By revealing the True Positive Rate (TPR) and the False Positive Rate, this curve demonstrates the classifier’s behaviour at each threshold (FPR). The purpose of the ROC Curve is to demonstrate the effectiveness of the model for each potential threshold as a relationship between TPR and FPR. These variables are computed for each threshold and displayed on a plane in order to depict the curve. Figure 7 illustrates the ROC curve for each baseline model. As depicted in the figure, BERT-TextCNN-BiLSTM attains a skilful multiclass classification.

The aggregated ROC-AUC scores along with the \log_loss values for each model in its multi-classification task is demonstrated in Fig. 8. An excellent model has a ROC-AUC close to 1, indicating that it has a high level of separability. An AUC close to 0, which implies the poorest measure of separability, indicates a bad model—It signals that it is reversing the outcome. If ROC-AUC is close to 0.5, the model has absolutely no potential for class separation. As observed in Fig. 8, the proposed BERT-BiLSTM-TextCNN model attains the lowest \log_loss value and best accumulated ROC-AUC score value nears to ‘1’. In contrary, TextCNN baseline model obtains the worst assessment in both ROC-AUC score and \log_loss values. The model has no distinguishing capacity to separate between classes.

This conducted ablation study provides a deeper understanding of the model’s internal mechanisms and sheds light on the importance of the different architectural elements. It allows us to draw conclusions about the impact of individual components on the overall performance and helps validate the design choices made in our proposed model. Specifically, we evaluated the performance of two ablated variants: BERT-BiLSTM and BERT-TextCNN. These variants allowed us to isolate the effects of the BiLSTM and TextCNN components when combined with the BERT embeddings. By comparing the performance of these ablated models against the full BERT-BiLSTM-TextCNN model, we were able to assess the added value and significance of each component. Next section also verifies the utility of the proposed model on a downstream task.

4.4 Downstream task experiment results

This section discusses the experiment that has been carried out to apply the proposed model in a downstream task. The collected dataset is described, followed by providing the outcomes of this experiment.

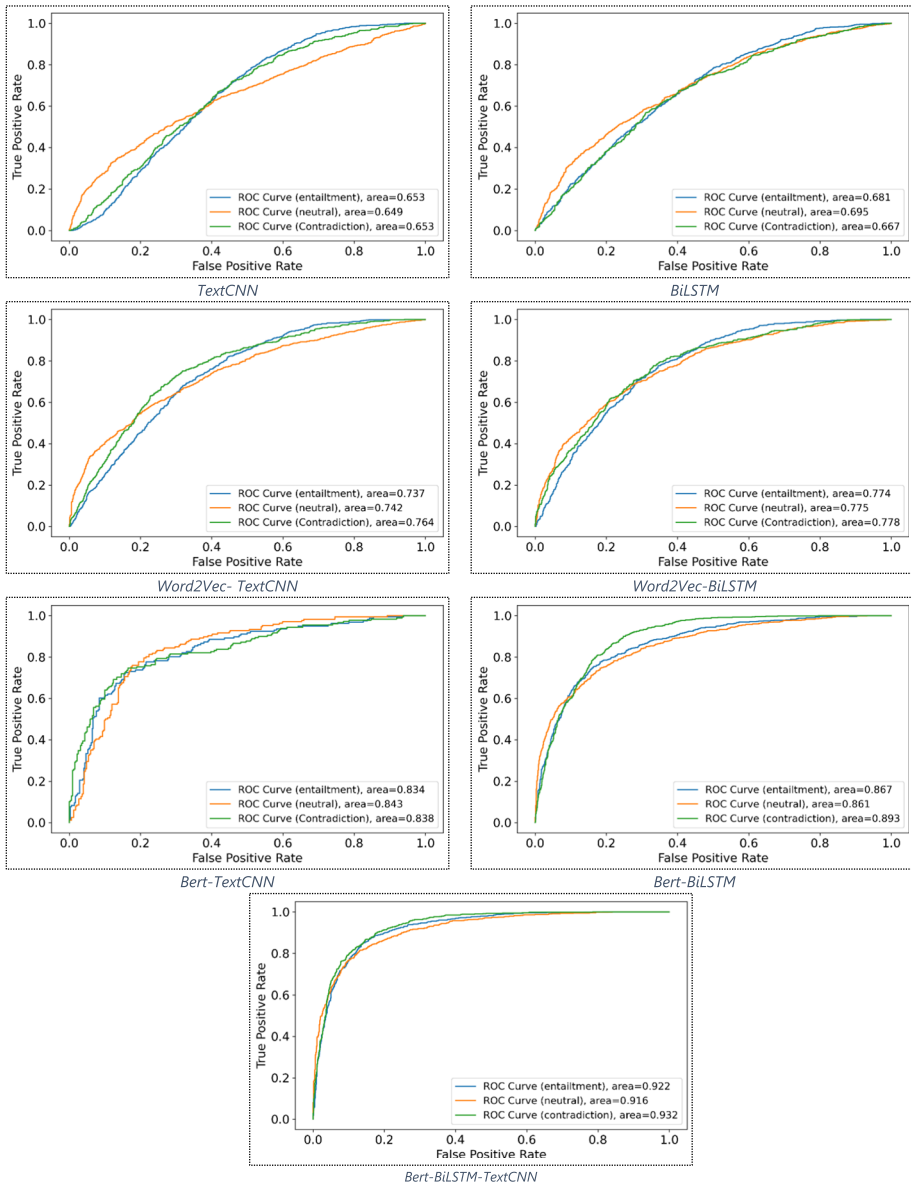


Fig. 7 ROC Curves for the baseline models

4.4.1 Social dataset description

As mentioned in the data preparation section, 14 Australian brands have been selected, whereby conversations with their customers are extracted and analysed. These Australian brands belong to three sectors, namely Banking, Airlines, and Sports. To provide temporal scrutiny to the conversations that occur between brands and customers, this study examines

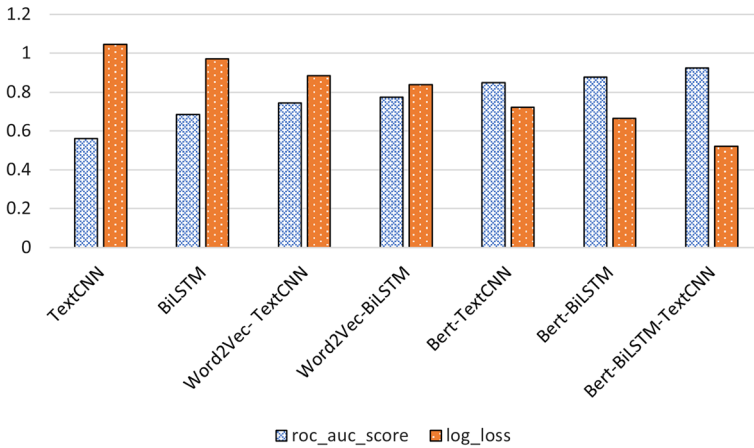


Fig. 8 ROC_auc_score and log_loss for the baseline models

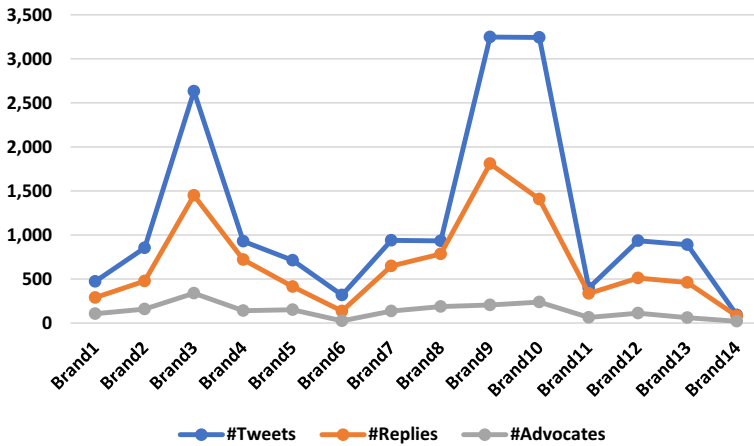


Fig. 9 The number of filtered tweets and their replies for 14 Australian brands from Jan to Sep 2022

the recently posted brands’ tweets, namely those posted from January to September 2022. These tweets and their replies are collected, followed by the followers extraction process. Finally, the social data of customers are obtained and filtered to infer users who convey positive attitudes toward the brands. Figure 9 demonstrates the number of filtered tweets and their replies as well as the inferred advocate of 14 Australian brands.

4.4.2 Correlation between NLI model and CE

To demonstrate the significance of the proposed NLI model on the designated task, we extracted the correlation between advocate customers and semantic similarities of their replies to brands’ tweets. In particular, we aim to answer the following research question:

RQ. Is there an inference relationship between the NLI approach and customer advocacy?

Table 5 The number of inferred NLI labels from the replies of brands' tweets

	#replies	#entailments (%)	#neutrals (%)	#contradictions (%)
Brand1	289	76 (26%)	96 (33%)	117 (41%)
Brand2	476	259 (54%)	78 (16%)	139 (29%)
Brand3	1449	422 (29%)	76 (5%)	951 (66%)
Brand4	720	452 (63%)	22 (3%)	246 (34%)
Brand5	414	72 (17%)	305 (74%)	37 (9%)
Brand6	136	115 (84%)	1 (1%)	20 (15%)
Brand7	648	447 (69%)	80 (12%)	121 (19%)
Brand8	784	644 (82%)	92 (12%)	48 (6%)
Brand9	1810	1258 (70%)	183 (10%)	369 (20%)
Brand10	1407	155 (11%)	997 (71%)	255 (18%)
Brand11	335	173 (52%)	67 (20%)	95 (28%)
Brand12	511	378 (74%)	38 (7%)	95 (19%)
Brand13	460	125 (27%)	215 (47%)	120 (26%)
Brand14	80	24 (30%)	20 (25%)	36 (45%)

To answer this research question, we compute the amount of entailment, neutrality, and contradiction the customer conveys to the brands' content. To accomplish this, we pose a brand's tweet as a 'Premise' and its replies as 'Hypotheses'. Then we incorporate the proposed BERT-BiLSTM-TextCNN model to classify each reply to one of the designated labels, namely entailment, neutral, and contradiction. Then we measure the correlation between these inferred labels and those categorised as an advocate and non-advocate customers, thereby indicating the relationship between the NLI approach and customer advocacy. Table 5 offers the number of inferred NLI labels obtained from the replies to brands' tweets using our proposed model. The table shows discrepancies in the inferred semantics of all brands. This implies that entailment, neutrality, and contradiction of customers' replies to brands' tweets vary amongst brands. This observation will be examined in future work, thereby obtaining clues or reasons behind such inconsistency.

Figure 10 depicts a deeper look at the analysis conducted on replies of the extracted brands' advocates. This figure depicts the direct relationships between advocates' replies and the brands' tweets. These Inferences show the meaning of otherwise disparate text fragments, demonstrating a cohesive discourse correlation between the brand and its advocates which answers the research question.

5 Discussion

The previous section describes the evaluation of a proposed NLI model named BERT-BiLSTM-TextCNN which has been compared with other baseline models such as TextCNN, BiLSTM, Bert-TextCNN, and Bert-BiLSTM. Word2Vec technique is also incorporated, which learns word embeddings through a neural network model and extracts the idea of relatedness between words or items. The embeddings obtained by Word2Vec are then trained using TextCNN (Word2Vec-TextCNN) and BiLSTM (Word2Vec-BiLSTM) as baseline methods for the NLI task. The models are evaluated based on metrics such as precision, recall, F1-score, micro-average F1-score, macro-average F1-score, K score, and

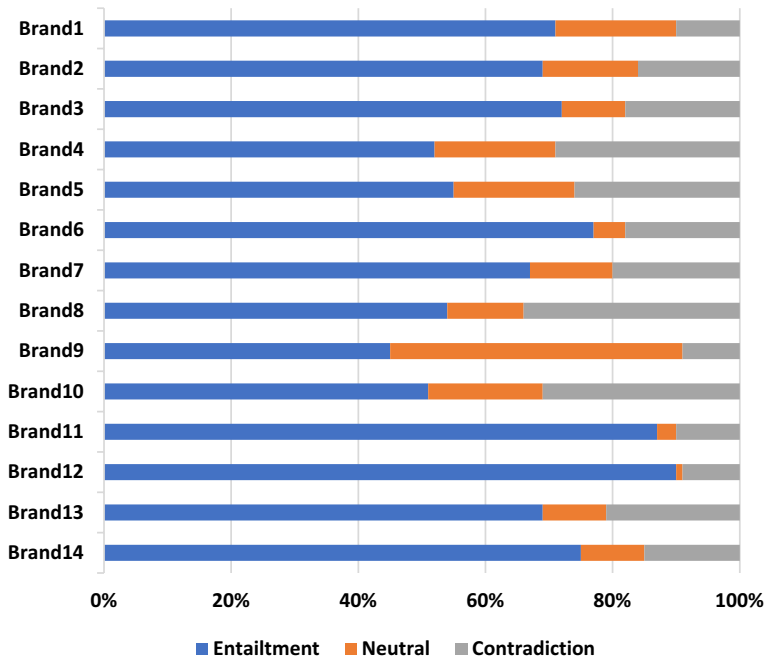


Fig. 10 NLI judgements obtained from advocates' replies to the brands using our model

MCC. The results show that the proposed model outperforms other baseline models in terms of various evaluation metrics, including F^{mic} , F^{mac} , $Kscore$, and MCC .

By combining the three architectures, BERT-BiLSTM-TextCNN is able to capture both the local and global features of text data. The BERT component provides the model with a strong understanding of the context of the text, while the BiLSTM and TextCNN components are able to capture the local features of the text. This combination of architectures allows the model to learn a more robust representation of the text data and therefore outperforms other models. In addition, BERT-BiLSTM-TextCNN also incorporates attention mechanisms, which allow the model to focus on the most important parts of the text. This further enhances the model's ability to understand the context of the text and improves its performance on natural language processing tasks.

The trade-off between model complexity and performance is a critical consideration in any machine learning task. On the one hand, complex models like BERT-BiLSTM-TextCNN have a high capacity to learn complex patterns and relationships in the data, leading to better performance. On the other hand, such models are computationally expensive and may require significant computational resources and time to train and evaluate. Additionally, more complex models are more difficult to interpret and may not provide a clear understanding of the factors driving their predictions. Our proposed approach strikes a balance between model complexity and performance by using a hybridized deep learning model that combines the strengths of BERT, BiLSTM, and TextCNN. While this approach does increase the model's complexity compared to simpler models, it achieves superior performance in identifying brand advocates in online customer engagement. Moreover, our approach can be fine-tuned to extract more relevant and precise features from the textual data, which helps to improve model

interpretability. Additionally, by incorporating various evaluation metrics, we show that our proposed approach outperforms existing approaches while maintaining a manageable level of complexity.

This study provides insights into the performance of deep learning models for text classification tasks. While BERT-BiLSTM-TextCNN outperforms the other models in terms of the designated task over the indicated evaluation metrics, future work will focus on adapting the model to new domains. Also, future work will incorporate more recent and complex approaches as baselines for comparison. This will include state-of-the-art models such as RoBERTa and GPT-3, which are based on BERT but have shown significant improvements in language understanding tasks. Further, other advanced techniques such as transformers, attention mechanisms, and ensembling will be explored to further enhance the performance of the model.

6 Conclusion

In a hyperconnected world where consumers are more empowered, businesses have struggled to develop innovative ways to serve their clients through valuable informational exchanges. Many NLP applications, such as semantic search and question-answering, depend on NLI. Due to the availability of massive, complex datasets, the NLI problem has attracted much attention. The majority of current approaches to the issue concentrate on learning-based techniques that employ textual data to categorise whether a given premise entails, contradicts, or is neutral with regard to a particular hypothesis. This paper proposes a novel approach to identify brand advocates using NLI and AI techniques. The model is based on the integration between BERT, BiLSTM, and TextCNN. This integrated model aims to provide a textual inference model capable of extracting the level of entailment, contradiction, and neutrality seen in the advocates' responses to the brands. Based on the semantic similarity between the brand tweets and the customers' replies, this adds a new layer to identifying advocates. The results of the experiments show that the hybridised model proposed in this study more successfully combines the advantages of TextCNN and BiLSTM; it captures local correlation while preserving context information and has fared well in many evaluation metrics. This is strengthened by including a fine-trained BERT model, which outperforms cutting-edge textual embedding models by better understanding contextual semantics. The developed BERT-BiLSTM-TextCNN model is used in a downstream job to show and validate its effectiveness in capturing the relationship between brands and their advocates.

Acknowledgements Not applicable.

Author contributions BAS: methodology, conceptualisation, implementation, and writing; MAW, MAA, and MAO: writing, literature review, MAAH, MAH,; GM, KC, and AH: editing and supervision. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability The dataset used in this research can be accessed at: <https://cloudstor.aarnet.edu.au/plus/s/4Ya0yIXcnSwk5on>.

Code availability Complete source code to reproduce the framework will be made open source, and the link will be provided with the camera-ready version of the paper.

Declarations

Conflict of interest The authors declare no conflicts of interest.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication I, the undersigned, give my consent for the publication of identifiable details, which can include photograph(s) and/or videos and/or case history and/or details within the text.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abu-Salih, B., Qudah, D. A., Al-Hassan, M., Ghafari, S. M., Issa, T., Aljarah, I., Beheshti, A., & Alqahtani, S. (2022). An intelligent system for multi-topic social spam detection in microblogging. *Journal of Information Science*, 01655515221124062.
- Abu-Salih, B., Wongthongtham, P., Zhu, D., Chan, K. Y., & Rudra, A. (2021). *Social big data analytics*. Springer.
- Al-Okaily, M. (2021). Assessing the effectiveness of accounting information systems in the era of COVID-19 pandemic. *VINE Journal of Information and Knowledge Management Systems*.
- Al-Okaily, M., & Al-Okaily, A. (2022). An empirical assessment of enterprise information systems success in a developing country: the Jordanian experience. *The TQM Journal*.
- Alsaad, A., & Al-Okaily, M. (2021). Acceptance of protection technology in a time of fear: the case of Covid-19 exposure detection apps. *Information Technology & People*.
- Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.
- Bansal, R., Pruthi, N., & Singh, R. (2022). Developing customer engagement through artificial intelligence tools: roles and challenges. In *Developing Relationships, Personalization, and Data Herald in Marketing 5.0* (pp. 130–145). IGI Global.
- Bilro, R. G., Loureiro, S. M. C., & Guerreiro, J. (2019). Exploring online customer engagement with hospitality products and its relationship with involvement, emotional states, experience and brand advocacy. *Journal of Hospitality Marketing & Management*, 28(2), 147–171.
- Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.
- Castillo López, A., Llorens Montes, F. J., & Braojos Gómez, J. (2021). Impact of social media on the firm's knowledge exploration and knowledge exploitation: The role of business analytics talent.
- Castillo, A., Benítez, J., Llorens, J., & Luo, X. R. (2021). Social media-driven customer engagement and movie performance: Theory and empirical evidence. *Decision Support Systems*, 145, 113516.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Dagan, I., Dolan, B., Magnini, B., & Roth, D. (2010). Recognizing textual entailment: Rational, evaluation and approaches—erratum. *Natural Language Engineering*, 16(1), 105–105.
- Dashtipour, K., Gogate, M., Adeel, A., Larjani, H., & Hussain, A. (2021). Sentiment analysis of persian movie reviews using deep learning. *Entropy*, 23(5), 596.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

- Dutta, K., Sharma, K., & Goyal, T. (2021). Customer's digital advocacy: the impact of reviews and influencers in building trust for tourism and hospitality services. *Worldwide Hospitality and Tourism Themes*.
- Eisingerich, A. B., Chun, H. H., Liu, Y., Jia, H. M., & Bell, S. J. (2015). Why recommend a brand face-to-face but not on Facebook? How word-of-mouth on online social sites differs from traditional word-of-mouth. *Journal of Consumer Psychology*, 25(1), 120–128.
- Gao, L., Li, G., Tsai, F., Gao, C., Zhu, M., & Qu, X. (2022). The impact of artificial intelligence stimuli on customer engagement and value co-creation: the moderating role of customer ability readiness. *Journal of Research in Interactive Marketing*(ahead-of-print).
- Goh, J. M., & Arenas, A. E. (2020). IT value creation in public sector: How IT-enabled capabilities mitigate tradeoffs in public organisations. *European Journal of Information Systems*, 29(1), 25–43.
- Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5–6), 602–610.
- Han, S., & Anderson, C. K. (2022). The dynamic customer engagement behaviors in the customer satisfaction survey. *Decision Support Systems*, 154, 113708.
- Herlihy, C., & Rudinger, R. (2021). MedNLI is not immune: Natural language inference artifacts in the clinical domain. *arXiv preprint arXiv:2106.01491*.
- Hollebeek, L. D., & Chen, T. (2014). Exploring positively-versus negatively-valenced brand engagement: a conceptual model. *Journal of Product & Brand Management*.
- Hollebeek, L. D., Sharma, T. G., Pandey, R., Sanyal, P., & Clark, M. K. (2021). Fifteen years of customer engagement research: a bibliometric and network analysis. *Journal of Product & Brand Management*.
- Ibrahim, N. F., & Wang, X. (2019). Decoding the sentiment dynamics of online retailing customers: Time series analysis of social media. *Computers in Human Behavior*, 96, 32–45.
- Jacob, M. E., & Johnson, J. (2021). Conceptualising Digital Content Marketing for Greater Consumer Brand Engagement. *Colombo Business Journal*, 12(2).
- Jain, P. K., Pamula, R., & Yekun, E. A. (2022). A multi-label ensemble predicting model to service recommendation from social media contents. *The Journal of Supercomputing*, 78(4), 5203–5220.
- Kartajaya, H., Kotler, P., & Setiawan, I. (2016). *Marketing 4.0: moving from traditional to digital*. Wiley.
- Kausar, S., Huahu, X., Ahmad, W., & Shabir, M. Y. (2019). A sentiment polarity categorization technique for online product reviews. *IEEE Access*, 8, 3594–3605.
- Khot, T., Sabharwal, A., & Clark, P. (2018). Scitail: A textual entailment dataset from science question answering. Proceedings of the AAAI Conference on Artificial Intelligence.
- Kim, H. D., & Zhai, C. (2009). Generating comparative summaries of contradictory opinions in text. Proceedings of the 18th ACM conference on Information and knowledge management.
- Kim, J.-Y., & Hwang, J. (2022). Who is an evangelist? Food tourists' positive and negative eWOM behavior. *International Journal of Contemporary Hospitality Management*, 34(2), 555–577.
- Lee, C.-H., & Kim, H.-R. (2022). Positive and negative switching barriers: Promoting hotel customer citizenship behaviour through brand attachment. *International Journal of Contemporary Hospitality Management*(ahead-of-print).
- Malesev, S., & Cherry, M. (2021). Digital and social media marketing-growing market share for construction SMEs. *Construction Economics and Building*, 21(1), 65–82.
- Mansoor, M., & Paul, J. (2022). Mass prestige, brand happiness and brand evangelism among consumers. *Journal of Business Research*, 144, 484–496.
- Matosas-López, L., & Romero-Ania, A. (2021). How to improve customer engagement in social networks: A study of Spanish brands in the automotive industry. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(7), 3269–3281.
- Mishra, A., Patel, D., Vijayakumar, A., Li, X. L., Kapanipathi, P., & Talamadupula, K. (2021). Looking beyond sentence-level natural language inference for question answering and text summarization. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.
- Mukherjee, A., Bhattacharyya, S., Ray, K., Gupta, B., & Das, A. K. (2022). A Study of Public Sentiment and Influence of Politics in COVID-19 Related Tweets. In *Computational Intelligence in Pattern Recognition* (pp. 655–665). Springer.
- Nasr, L., Burton, J., & Gruber, T. (2017). Developing a deeper understanding of positive customer feedback. *Journal of Services Marketing*.
- Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). BERTweet: A pre-trained language model for English Tweets. *arXiv preprint arXiv:2005.10200*.
- Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2019). Adversarial NLI: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

- Oh, C., Roumani, Y., Nwankpa, J. K., & Hu, H.-F. (2017). Beyond likes and tweets: Consumer engagement behavior and movie box office in social media. *Information & Management*, 54(1), 25–37.
- Perez-Vega, R., Kaartemo, V., Lages, C. R., Razavi, N. B., & Männistö, J. (2021). Reshaping the contexts of online customer engagement behavior via artificial intelligence: A conceptual framework. *Journal of Business Research*, 129, 902–910.
- Pillai, P., Ryali, S., Maniar, H., Mangsuli, P., & Abubakar, A. (2022). NLP applications in the oil and natural gas industry. Second International Meeting for Applied Geoscience & Energy
- Potdar, V., Joshi, S., Harish, R., Baskerville, R., & Wongthongtham, P. (2018). A process model for identifying online customer engagement patterns on Facebook brand pages. *Information Technology & People*.
- Read, W., Robertson, N., McQuilken, L., & Ferdous, A. S. (2019). Consumer engagement on Twitter: perceptions of the brand matter. *European Journal of Marketing*.
- Rehman, A. U., Malik, A. K., Raza, B., & Ali, W. (2019). A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. *Multimedia Tools and Applications*, 78(18), 26597–26613.
- Sadeghi, F., Bidgoly, A. J., & Amirkhani, H. (2022). Fake news detection on social media using a natural language inference approach. *Multimedia Tools and Applications*, pp. 1–21.
- Saldanha, T. J., Sahaym, A., Mithas, S., Andrade-Rojas, M. G., Kathuria, A., & Lee, H.-H. (2020). Turning liabilities of global operations into assets: IT-enabled social integration capacity and exploratory innovation. *Information Systems Research*, 31(2), 361–382.
- Sashi, C. M. (2012). Customer engagement, buyer-seller relationships, and social media. *Management Decision*.
- Sashi, C., Brynildsen, G., & Bilgihan, A. (2019). Social media, customer engagement and advocacy: An empirical investigation using Twitter data for quick service restaurants. *International Journal of Contemporary Hospitality Management*.
- Sayce, D. (2022). *The Number of tweets per day in 2022*. Retrieved 24/09/2022 from <https://www.dsayce.com/social-media/tweets-day/>
- Simon, F., & Tossan, V. (2018). Does brand-consumer social sharing matter? A relational framework of customer engagement to brand-hosted social media. *Journal of Business Research*, 85, 175–184.
- So, R., Chu, C. F. C., & Lee, C. W. J. (2022). Extract Aspect-based Financial Opinion Using Natural Language Inference. Proceedings of the 2022 International Conference on E-business and Mobile Commerce,
- Song, X., Salcianu, A., Song, Y., Dopson, D., & Zhou, D. (2020). Fast wordpiece tokenization. *arXiv preprint arXiv:2012.15524*.
- Song, S., & Kim, H.-Y. (2022). Is social media marketing worth it for luxury brands? The dual impact of brand page satisfaction and brand love on word-of-mouth and attitudinal loyalty intentions. *Journal of Product & Brand Management*.
- Sun, X., Foscht, T., & Eisingerich, A. B. (2021). Does educating customers create positive word of mouth? *Journal of Retailing and Consumer Services*, 62, 102638.
- Sweeney, J., Payne, A., Frow, P., & Liu, D. (2020). Customer advocacy: A distinctive form of word of mouth. *Journal of Service Research*, 23(2), 139–155.
- Unit, E. I. (2007). Beyond loyalty: Meeting the challenge of customer engagement. Retrieved November, 15, 2014.
- Van Doorn, J., Lemon, K. N., Mittal, V., Nass, S., Pick, D., Pirner, P., & Verhoef, P. C. (2010). Customer engagement behavior: Theoretical foundations and research directions. *Journal of Service Research*, 13(3), 253–266.
- Walz, A., & Celuch, K. (2010). The effect of retailer communication on customer advocacy: The moderating role of trust. *The Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, 23.
- Wang, X., Kapanipathi, P., Musa, R., Yu, M., Talamadupula, K., Abdelaziz, I., Chang, M., Fokoue, A., Makni, B., & Mattei, N. (2019). Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the AAAI conference on artificial intelligence*
- Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F., & Chao, L. S. (2019). Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*.
- Williams, A., Nangia, N., & Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., & Funtowicz, M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- Wongthongtham, P., & Abu-Salih, B. (2018). Ontology-based approach for semantic data extraction from social big data: state-of-the-art and research directions. *arXiv preprint arXiv:1801.01624*.

- Yang, M., Ren, Y., & Adomavicius, G. (2019). Understanding user-generated content and customer engagement on Facebook business pages. *Information Systems Research*, 30(3), 839–855.
- Zhu, J., Patra, B. G., & Yaseen, A. (2021). Recommender system of scholarly papers using public datasets. *AMIA Summits on Translational Science Proceedings, 2021*, 672.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Bilal Abu-Salih^{1,5}  · Mohammed Alweshah² · Moutaz Alazab² · Manaf Al-Okaily³ · Muteeb Alahmari⁴ · Mohammad Al-Habashneh⁵ · Saleh Al-Sharaeh¹

✉ Bilal Abu-Salih
b.abusalih@ju.edu.jo

Mohammed Alweshah
weshah@bau.edu.jo

Moutaz Alazab
m.alazab@bau.edu.jo

Manaf Al-Okaily
m.alokaily@jadara.edu.jo

Muteeb Alahmari
msaad@iau.edu.sa

Mohammad Al-Habashneh
mohammad.alhabashneh@curtin.edu.au

Saleh Al-Sharaeh
ssharaeh@ju.edu.jo

- ¹ King Abdullah II School of Information Technology, The University of Jordan, Amman, Jordan
- ² Prince Abdullah Bin Ghazi Faculty of Information and Communication Technology, Al-Balqa Applied University, Al-Salt, Jordan
- ³ Faculty of Business, Jadara University, Jarash, Jordan
- ⁴ Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia
- ⁵ School of Management and Marketing, Curtin University, Perth, Australia