

Article

Activity Classification in E-Commerce Product Reviews Using Deep Learning and Transformer Models

Tinashe Wamambo, Arooj Fatima , Bethwel Kiplagat, Mahdi Maktab Dar Oghaz  and Cristina Luca * 

Faculty of Science and Engineering, Anglia Ruskin University, Cambridge CB1 1PT, UK; tinashe.wamambo@hotmail.co.uk (T.W.); arooj.fatima@aru.ac.uk (A.F.); bk538@student.aru.ac.uk (B.K.); mahdi.maktabdar@aru.ac.uk (M.M.D.O.)

* Correspondence: cristina.luca@aru.ac.uk

Abstract

Existing research on e-commerce product reviews has primarily focused on analysing consumers' opinions, emotions, sentiments and associated star ratings. Whilst these approaches provide insights into consumers' perceptions of products, they offer limited understanding of how products are used in real-world contexts. Therefore, they do little to enhance the e-commerce experience by helping consumers make more informed purchasing decisions based on products' intended uses without requiring them to read numerous reviews during the decision-making process. To address this problem, this paper investigates the feasibility of automatically identifying and classifying product usage activities from e-commerce reviews. A methodology combining natural language processing, manual activity-level annotation and deep learning-based text classification was developed and evaluated. An initial dataset of 60,000 Amazon product reviews was manually labelled according to six activity classes: run, walk, hike, swim, climb and unknown. Following quality inspection and data cleaning, a final dataset of 50,843 reviews was used for model training and evaluation. Multiple classification approaches were assessed, including CNN, LSTM, hybrid LSTM-CNN architectures and transformer-based models (DistilBERT and DistilBERT-CNN). Experimental evaluation was conducted using multiple random seeds to ensure robustness and reproducibility. The results indicate that activity classification from e-commerce reviews is a challenging task due to ambiguity and overlapping usage descriptions, with all evaluated models achieving comparable performance on the full dataset. Among the evaluated models, the hybrid LSTM-CNN-GloVe architecture achieved the highest performance on the keyword-filtered dataset, whilst the DistilBERT-CNN model also demonstrated strong results. The findings demonstrate the feasibility of extracting activity-oriented information from product reviews and highlight activity classification as a distinct and under-explored natural language processing task that complements traditional sentiment analysis. The proposed methodology provides a foundation for improving product discovery and supporting usage-oriented search and recommendation systems in e-commerce environments.



Academic Editor: Olga Kurasova

Received: 14 October 2025

Revised: 14 July 2026

Accepted: 17 July 2026

Published: 23 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: activity classification; deep learning; transformer models; e-commerce reviews; word embeddings; natural language processing

1. Introduction

In the modern world, the rapid growth of e-commerce platforms has generated unprecedented volumes of product reviews written by users, providing valuable information

about their experiences and product usage. E-commerce retailers such as Amazon receive thousands of product reviews and consumer feedback on a daily basis, outlining consumers' use and experiences with the products. This has led to increased research interest in sentiment analysis and text classification [1].

Consumers often rely on product reviews to learn whether a product fulfils its promises, meets their expectations and aligns with their needs. However, due to the sheer volume of reviews generated for a given product, it is unfeasible for the average consumer to investigate them all and gain a comprehensive understanding of both the product use and the sentiments of previous users. Currently, the quickest and easiest decision-making element tied to product reviews is the star rating. The higher the star rating, the more likely a consumer is to purchase the product. However, a star rating is ambiguous and prone to grade inflation. For example, a product with a 4.8-star rating may not necessarily be exceptional, and the difference between a product with a 4.5-star rating and one with a 4.8-star rating could be significant. This makes it challenging for consumers to distinguish between adequate products and very good ones.

An additional drawback of star ratings is their superficial nature as they provide little insight into the specific opinions expressed in reviews or a product's suitability for certain tasks [2]. For example, a pair of shoes with a high star rating may be suitable for running but not for hiking. This information may not be readily available to potential consumers looking to purchase those shoes.

Research on product reviews has so far focused on sentiment analysis, which is the process of identifying and categorising the sentiment expressed in a text document [3]. Using natural language processing (NLP) and machine learning (ML) techniques, prior studies have classified reviews as positive, negative or neutral [4], identified subjective or objective opinions [5,6], performed aspect-based sentiment analysis [7] and compared the performance of various machine learning algorithms for classifying text documents [8,9]. Although hybrid deep learning architectures such as CNN-LSTM and embedding-based models have achieved strong performance in sentiment classification tasks, these approaches largely concentrate on identifying emotional responses and opinions rather than understanding how products are used in practice. Therefore, useful information on product use and functional suitability remains unexplored. This creates a clear research gap in extracting and structuring activity-oriented information from product reviews.

The problem addressed in this paper is the automatic classification of product reviews according to the activity being performed using the product. In this context, an activity is defined as an action performed by the consumer using a product as described in a review. For example, in the sentence "I bought these boots for walking my dogs around the local park, they have been fit for purpose thus far," the identified activity is "walking". Unlike sentiment analysis, which focuses on identifying how consumers feel about a product, activity classification aims to determine how products are used in real-world contexts. Although both tasks involve analysing user-generated text using NLP techniques, activity classification represents a conceptually distinct problem that complements rather than replaces sentiment analysis.

To address this research gap, this study investigates deep learning approaches for activity classification in e-commerce product reviews. The study proposes a machine learning based approach incorporating data pre-processing, manual activity annotation and text classification. An initial dataset of 60,000 Amazon product reviews was manually labelled and then refined through quality inspection and cleaning, resulting in a final dataset of 50,843 reviews used to evaluate multiple deep-learning architectures, including LSTM, CNN, hybrid LSTM-CNN-GloVe models and transformer-based approaches (DistilBERT and DistilBERT-CNN).

The experimental results demonstrate that activity classification from user reviews is a challenging task due to ambiguity and overlap between activity descriptions. On the full dataset, all evaluated models achieved comparable performance, with only marginal improvements observed for transformer-based approaches. However, on the keyword-filtered dataset, the hybrid LSTM-CNN-GloVe architecture achieved the highest performance, whilst the DistilBERT-CNN model also demonstrated strong results.

The contribution of this work is threefold. First, it introduces activity classification as a distinct application scenario within e-commerce review analysis by shifting the focus from sentiment prediction towards understanding product usage behaviour. Second, it presents a manually annotated dataset with activity labels, supporting the modelling of real-world product usage contexts. Third, it provides a comparative evaluation of multiple deep learning and transformer-based architectures under rigorous experimental conditions, including five-fold cross-validation and multiple random seeds to ensure robustness and reproducibility.

The rest of this paper is organised as follows: Section 2 reviews the relevant background and related research on e-commerce review analysis and text classification; Section 3 describes the proposed activity classification methodology, including dataset construction, annotation procedures, text pre-processing and the evaluated classification models; Section 4 presents a comparative evaluation and analysis of the investigated deep learning and transformer-based approaches; finally, the main findings, the limitations of the study and directions for future research are summarised in Section 5.

2. Literature Review

Research on product reviews has evolved significantly, with studies primarily focusing on sentiment analysis, review helpfulness and text classification. With advances in NLP and deep learning, researchers have developed increasingly sophisticated approaches for extracting meaningful insights from large volumes of user-generated content. The existing literature can be broadly categorised into four main themes: (1) sentiment and product review analysis, (2) usage context and activity recognition, (3) neural network-based text classification, and (4) feature representation and word embedding techniques.

2.1. Sentiment and Product Review Analysis

A substantial body of research has focused on analysing sentiment and evaluating the usefulness of product reviews. Early work in this area relied on traditional machine learning techniques combined with feature engineering methods such as Part-of-Speech (POS) tagging and n-gram extraction.

Wu and Wang [10] investigated the helpfulness of Amazon reviews for shoe products by classifying them into helpful or unhelpful categories using a Convolutional Neural Network (CNN) combined with syntactic features. Their results demonstrated that incorporating linguistic features improved classification performance. Similarly, Sun et al. [11] classified reviews according to their informativeness by distinguishing between search and experience products and introducing classification thresholds.

Wu and Chen [12] extended the work carried out by [10] and employed a BERT-based model to generate contextual sentence embeddings and analyse temporal voting patterns for the same reviews. Their findings showed that voting results for online reviews were not constant, with many reviews receiving few or no votes over time.

Sentiment classification has also been widely studied for detecting misleading or unfair reviews. Elmurngi and Gherbi [13] compared several supervised learning algorithms, including Naïve Bayes, Decision Trees, Logistic Regression and Support Vector Machines

and found Logistic Regression to be the most effective for detecting unfair sentiment in reviews.

Other studies have explored the use of star ratings as labels for sentiment analysis. Zhao et al. [14] and Sindhura et al. [15] used star ratings as weak labels to classify the sentiment and emotions expressed in product reviews. Sadhana et al. [16] disregarded the star ratings entirely and focused on target opinion words to determine the polarity of product reviews. Nasim et al. [17] also ignored star ratings and instead combined Term Frequency–Inverse Document Frequency (TF-IDF) with n-grams to vectorise features. Their goal was to extract unigram and bigram features. However, they discovered that processing high-value n-gram features resulted in a high-dimensional feature space. To limit ambiguity and ensure precision within the features, they only extracted bi-gram features.

Other researchers have investigated inconsistencies between textual sentiment and numerical ratings. Yao et al. [18] utilised a Statistics-based Outlier Detection and Correction Method (SODCM) to identify mismatches between sentiment and ratings, whilst Wahyudi and Kusumaningrum [19] used Latent Dirichlet Allocation (LDA) to filter out irrelevant aspects, such as delivery services, to better evaluate product quality.

Whilst these studies provide valuable insights into sentiment polarity, review quality and reliability, they primarily focus on what consumers feel rather than how products are used. As a result, they offer limited understanding of usage behaviour in real-world contexts.

2.2. Usage Context and Activity Recognition

Compared with sentiment analysis, relatively few studies have focused on identifying product usage contexts or consumer activities. This area is particularly important for understanding how products are used in practice and how they meet user needs.

Suryadi and Kim [20] proposed a data-driven framework for identifying product usage contexts in online reviews. Their approach combined machine learning and NLP techniques with aspect-based sentiment analysis to capture sentiment associated with specific usage contexts. Whilst their results demonstrated the ability to extract meaningful contextual information, they found that overall ratings were not a strong indicator of consumer sentiment towards a particular usage context due to the moderate linear correlation for most of the usage contexts in their study. More recently, Amirifar et al. [21] explored the relationship between product features and product popularity using deep learning models (i.e., DFFNN, PNN and RBFNN) combined with Named Entity Recognition (NER). Their approach highlighted the potential for extracting actionable insights beyond sentiment classification by focusing on product characteristics.

Hayatin et al. [22] presented a systematic literature review of sentiment summarisation research, especially focusing on aspect extraction techniques published between 2004 and 2023, reviewing more than 150 studies. Their review highlighted that most existing work focuses on explicit aspect extraction, whilst implicit aspects remain under-explored despite being common in reviews.

2.3. Neural Network-Based Text Classification

The rapid development of deep learning has significantly advanced text classification in product review analysis. Neural network architectures such as CNNs, Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks and transformer-based models have been widely adopted due to their ability to capture complex linguistic patterns.

CNN-based models have been effectively used to extract local features from text, as demonstrated by Wu and Wang [10]. Recurrent architectures, particularly LSTM and

Bidirectional LSTM (BiLSTM), have been employed to capture sequential dependencies and long-range contextual information in text data [23].

Transformer-based models, especially BERT, have further improved performance by generating context-aware embeddings. Ahmed and Wang [7] enhanced BERT by incorporating Normalised Google Distance (NGD) to improve semantic representations, whilst Xie et al. [24] used a pre-trained BERT model with aspect-target recognition capabilities. Al-Garadi et al. [25] concluded that fusion transformer models significantly outperformed traditional ML and deep learning baselines in their research to detect prescription medication misuse and abuse from Twitter posts.

Ali et al. [26] analysed Amazon product reviews by classifying them into positive, negative and neutral sentiment categories. The authors compared the performance of several sentiment analysis approaches, including traditional machine learning models such as Logistic Regression, Random Forest, Decision Tree and Naïve Bayes, as well as deep learning models including CNN and BiLSTM, and transformer-based models such as BERT and XLNet. The findings showed that transformer-based techniques achieved superior performance in sentiment classification. However, the study focused only on overall sentiment prediction and did not perform aspect-based sentiment analysis, meaning that individual product features or attributes could not be analysed separately. The authors therefore recommended aspect-based sentiment analysis as an area for future research.

Hybrid architectures have also gained popularity by combining the strengths of different models. For example, Xiaoyan et al. [27] proposed a GloVe-CNN-BiLSTM model, whilst Pimpalkar and Raj R [28] combined CNN with BiLSTM to capture both local and long-range dependencies. Bellar et al. [29] further evaluated various deep learning models and word embedding techniques for sentiment prediction. Alghaligah et al. [30] proposed an optimised hybrid deep learning model aimed at improving classification accuracy and efficiency. The findings demonstrated that the proposed model outperformed traditional machine learning methods. However, the study primarily focused on binary classification between spam and legitimate messages and did not address more advanced challenges such as context-aware analysis of review content.

These models achieve high performance in classification tasks but, they are predominantly applied to sentiment analysis and review categorisation. Their application to identifying usage activities remains limited.

2.4. Feature Representation and Word Embedding Techniques

Feature representation plays a critical role in text classification performance. Traditional approaches such as Bag-of-Words (BoW), TF-IDF and n-grams have been widely used due to their simplicity and interpretability. However, these methods often result in high-dimensional and sparse feature spaces.

Nasim et al. [17] highlighted this limitation when using n-gram features and addressed it by restricting their model to bigrams. Similarly, Arthamevia and Purbolaksono [31] combined TF-IDF with n-gram features and used Support Vector Machines for aspect-based sentiment classification. More recent approaches use dense word embeddings to capture semantic relationships between words. Models such as GloVe, Word2Vec, FastText and BERT embeddings provide richer text representations. For example, Xiaoyan et al. [27] integrated GloVe embeddings into a hybrid deep learning model, whilst Bellar et al. [29] compared multiple word-embedding techniques for sentiment classification.

2.5. Research Gap Analysis

The existing literature demonstrates significant progress in sentiment analysis, review helpfulness evaluation and text classification using both traditional and deep learning

approaches. However, most studies focus on analysing opinions, sentiments or product features rather than understanding how products are used in real-world contexts. Although some research have explored usage context, it remains limited and does not explicitly address activity-based product usage. Furthermore, whilst advanced neural network architectures and word-embedding techniques have improve classification performance, they are rarely applied to identifying consumer activities. Table 1 summarises the key characteristics and limitations of the most relevant existing approaches discussed in this section. To address these limitations, this study introduces an activity classification methodology for automatically identifying the specific activities consumers perform using a product based on information extracted from e-commerce reviews. This enables users to make more informed purchasing decisions based on real-world usage scenarios, reducing the need to manually analyse large volumes of reviews.

Table 1. Comparison of the most relevant existing systems.

Study	Task	Method	Focus	Limitation
Wu and Wang [10]	Helpfulness classification	CNN + syntactic features	Review usefulness	No usage context
Wu and Chen [12]	Helpfulness prediction	BERT Temporal	voting behaviour	No activity extraction
Sun et al. [11]	Review informativeness	ML classifiers	Helpful reviews	No context modelling
Elmurngi and Gherbi [13]	Sentiment classification	LR, SVM, NB	Detect unfair reviews	No usage insight
Suryadi and Kim [20]	Usage context detection	NLP + ML	Context & sentiment	No activity-level detail
Zhao et al. [14]	Sentiment classification	Weak supervision	Sentiment polarity	Relies on ratings
Ahmed and Wang [7]	Aspect sentiment	BERT + NGD	Semantic understanding	No usage activity
Ali et al. [26]	Sentiment classification	ML + DL models	Model comparison	Not task-focused
Amirifar et al. [21]	Feature impact analysis	DL + NER	Product features	Not usage behaviour

3. Activity Classification Methodology

3.1. Overview

This section presents the methodology used to identify and classify product usage activities from Amazon reviews. As illustrated in Figure 1, the approach consists of four main stages: dataset construction and annotation, text pre-processing, activity classification and model evaluation.

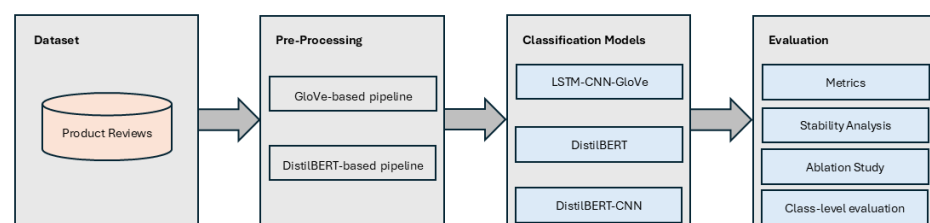


Figure 1. Overview of the proposed activity classification methodology.

The dataset used in this study contains product reviews extracted from Amazon, where each review is assigned an activity label representing the primary activity performed using the product. The reviews are pre-processed using NLP techniques including text cleaning, tokenisation and feature extraction at the document level. Features are derived from words and phrases containing nouns and verbs that indicate product usage context. Activities are expected to be identified through action verbs such as “run”, “walk” and “hike”. These features are provided as input to the Classification component, which is responsible for training and evaluating multiple models for activity classification. The evaluated approaches include CNN, LSTM, hybrid LSTM-CNN models and transformer-based architectures. The goal is to produce a trained machine learning model capable of predicting the activity with the highest prediction probability. The trained models are evaluated using accuracy, precision, recall, macro F1-score, confusion matrix analysis, and overfitting gap measurements. To improve robustness and reproducibility, model evaluation is conducted across multiple experimental runs using different random seeds.

The objective of this study is to investigate activity classification as an underexplored NLP task within the analysis of Amazon product reviews. The activity classification methodology provides a structured pipeline for transforming raw reviews into activity predictions. The following subsections describe each stage of the methodology and the associated experiments in detail.

3.2. Dataset Construction and Annotation

Amazon product reviews were used to develop and evaluate the activity classification models proposed in this paper. The reviews were collected from a publicly available repository containing approximately 5 million Amazon product reviews (<https://nijianmo.github.io/amazon/index.html>, accessed on 16 July 2026).

For the purposes of this study, the following datasets (<https://figshare.com/s/d7cf7839755a33c682de>, accessed on 16 July 2026) were created from the collected product reviews:

1. The **extensive dataset** contains 5 million product reviews.
2. The **annotated activity subset** was initially created by manually labelling 60,000 reviews extracted from the **extensive dataset**. Following manual quality inspection, reviews that were unsuitable for activity classification were removed. The remaining reviews were then balanced across the six activity classes where possible to minimise class imbalance during model training, resulting in a final annotated dataset of 50,843 reviews. This dataset was used to train and evaluate the classification models described in Section 3.4.
3. The **keyword-filtered activity subset** contains 30,000 reviews in which the assigned activity label is explicitly referenced in the review text. This dataset was created to support an additional experiment investigating the impact of reduced label ambiguity on classification performance.
4. The **independent validation subset** contains 1200 product reviews extracted from the **extensive dataset** as a subset reserved for independent model evaluation.

The **annotated activity subset** and **independent validation subset** contain different product reviews to ensure that model training and validation were conducted on separate data, therefore reducing potential bias.

The dataset contains reviews for products in the following categories: shoes, boots, trainers, shorts and clothing. These products are commonly linked to the activity classes investigated in this research. Each review was manually labelled with one of the target activity classes: *run*, *walk*, *hike*, *swim*, *climb* and *unknown*.

The subsets were manually cleaned and labelled prior to experimentation. The cleaning process involved inspecting the content of the reviews to ensure their relevance and suitability for this research. Reviews were then manually analysed and labelled according to the primary activity performed using the product.

The annotation process was conducted manually by a single annotator with a background in machine learning and natural language processing. The annotator examined each review to identify the primary activity performed using the product, based on the contextual meaning of the text. The annotation process followed consistent guidelines to enhance reproducibility and maintain labelling consistency.

Activities were identified using action-oriented keywords (verbs such as “run”, “walk”, “climb”) and their contextual use within the review. Each review was assigned a single label corresponding to the most prominent activity described, reflecting the single-label classification setting adopted in this study.

In cases where multiple activities were mentioned within a single review, the annotator selected the most clearly emphasised or frequently referenced activity. Reviews where no clear or dominant activity could be identified or where the activity was ambiguous were labelled as *unknown*. This approach ensured consistency in handling unclear or noisy data. Analysis of the dataset showed that 15.9% of reviews contained overlapping activities, highlighting structural ambiguity within the annotation task.

The **independent validation subset** was reserved for evaluating model performance on previously unseen reviews, therefore assessing the model’s ability to generalise beyond the training data. Annotation reliability was evaluated through a follow-up survey study involving 100 participants. Participants were asked to independently identify activities within a subset of product reviews. The results of this survey were compared with the manually annotated labels and showed strong alignment, providing supporting evidence for the reliability of the annotation process.

The final **annotated activity subset** achieved a near-balanced distribution across the six activity classes. This distribution was selected to reduce class imbalance during model training and minimise potential bias towards any particular activity class. Due to the limited availability of suitable *climb* reviews after quality inspection, this class contains 8343 instances, whilst the remaining activity classes contain 8500 instances each. Table 2 presents the distribution of activity labels within the dataset, whilst Table 3 provides examples of product reviews and the corresponding activity labels.

Table 2. Distribution of classified product activities.

Activity	Description	Number of Reviews	Percentage
Run	Activities involving running	8500	16.7%
Walk	Activities involving walking	8500	16.7%
Hike	Hiking activities	8500	16.7%
Swim	Activities involving swimming	8500	16.7%
Climb	Climbing activities	8343	16.4%
Unknown	No clear activity identified	8500	16.7%

The **keyword-filtered activity subset** was constructed to investigate the impact of reduced label ambiguity on classification performance. Only reviews containing an explicit reference to the assigned activity label were retained. The keywords included base forms and common inflections, for example, “run”, “running” and “runner” for the *run* class, “walk” and “walking” for the *walk* class and similar variations for other activity categories (e.g., *hike*: “hiking”, “trail”, “trek”, “summit”). Reviews labelled as *unknown* were excluded from this subset, as this class does not correspond to a specific activity keyword. This

filtering process was intended to reduce ambiguity by ensuring alignment between the textual content and the assigned activity label. As a result, the dataset size was reduced to approximately 30,000 reviews for training and 803 reviews for validation.

Table 3. Examples of reviews and the target activities.

Review	Target Activity
"It is still quite stiff and not stretched so it hurts when I wear it to climb. I had to keep changing to another climbing shoe."	Climb
"Love these fior swimming in weedy or rocky areas; also great for kayaking and canoeing"	Swim
"Extremely comfortable. No breaking in needed. Very light and not chunky at all. I use them on my mountain bike as they have a good strong sole and are great for when I have to jump off and cross deep rivers or climb up steep sand banks etc. Very, very happy I chose this shoe."	Climb
"They are comfortable and all my friends comment how much they like them. Vegan shoes! Now my conscience feels a bit better and I joke that if I get lost in the woods I could eat them for dinner. They do not have the support of regular athletic shoes so I don't walk all over town in them. For me they are good for most everything else."	Walk

3.3. Text Pre-Processing

Prior to model training, the product reviews were pre-processed to improve the quality and consistency of the data. The pre-processing pipeline varied depending on the model architecture, reflecting the differences between traditional embedding-based approaches and transformer-based models. In both cases, the goal was to transform the raw review text into a suitable numerical representation for input to the classification models.

3.3.1. Pre-Processing for GloVe-Based Models

For the GloVe-based models (CNN, LSTM and hybrid LSTM-CNN architectures), a range of NLP techniques was employed to support activity classification, including tokenisation, Part-of-Speech (POS) tagging, stop word removal and feature extraction.

Tokenisation was done using the Keras Tokenizer with a vocabulary size of 10,000 words and the infrequent words occurring fewer than two times were mapped to an out-of-vocabulary (OOV) token. POS tagging was used to identify linguistically relevant terms associated with user activities, particularly verbs and noun phrases that describe the activity a consumer performed using a product.

Common English stop words were removed using the NLTK stopword corpus to reduce sequence length and focus the models on semantically meaningful terms associated with product usage activities. Features were extracted primarily in the form of nouns and verbs, based on the assumption that activities are commonly expressed through verbs (e.g., *run, walk, hike*). The resulting feature set was used as input to the evaluated ML models.

3.3.2. Pre-Processing for DistilBERT-Based Models

For the DistilBERT-based models, pre-processing was minimal and aligned with the requirements of transformer architectures. Unlike the GloVe-based pipeline, this approach did not rely on explicit feature extraction, as contextual representations were learned directly by the model.

The input product reviews were first tokenised using the DistilBERT WordPiece tokenizer. WordPiece tokenisation splits rare or unseen words into subword units, allowing

the model to handle misspellings, informal language and domain-specific vocabulary commonly found in online product reviews. Each review was converted into token identifiers and accompanied by an attention mask to distinguish real from padding tokens. Reviews were padded or truncated to a fixed sequence length to ensure consistent tensor dimensions during batch training. The tokenised reviews were then passed through the frozen DistilBERT encoder, which generated contextual token-level representations for the subsequent classification process.

3.4. Classification Approaches

3.4.1. Experimental Setup

A range of machine learning and deep learning experiments were conducted to investigate the feasibility of activity classification in e-commerce product reviews and to compare the performance of different text classification techniques. The evaluated methods included traditional feature representations (Bag of Words and GloVe Word Embeddings), neural network architectures (CNN, LSTM, hybrid LSTM-CNN-GloVe models) and transformer-based models (DistilBERT and hybrid DistilBERT-CNN). Their comparative performance and effectiveness are presented and discussed in Section 4.

To ensure robustness and reproducibility, a repeated evaluation strategy with multiple random seeds was adopted. Each model was trained five times using different random seeds (42, 123, 456, 789 and 1011) on a fixed 90/10 training–validation split and the reported results correspond to the mean and standard deviation across the five runs. Performance was assessed using classification accuracy, precision, recall and F1-score.

All experiments were implemented in Python 3.8 using TensorFlow 2.x with the Keras API. Transformer-based models employed the Hugging Face Transformers library for DistilBERT tokenisation and contextual embedding generation. Natural language pre-processing, including POS tagging, lemmatisation and stop-word removal, was performed using NLTK, whilst NumPy and pandas were used for data handling and scikit-learn was used to compute the evaluation metrics. All experiments were conducted on a single GPU.

The principal hyperparameters used for the evaluated deep learning architectures are summarised in Table 4. For the GloVe-based models, pre-trained 100-dimensional GloVe embeddings were used and kept fixed during training. Reviews were padded or truncated to fixed sequence lengths to enable batch processing. The LSTM layer contains 100 hidden units and optimisation was performed using the Adam optimiser with sparse categorical cross-entropy loss for up to 10 epochs, with early stopping (patience = 5) to reduce overfitting. Results were averaged across five runs using different random seeds.

The Bag of Words baseline model employed a vocabulary of 400 terms represented using binary term encoding, followed by two fully connected Dense layers of 64 neurons each with a ReLU activation function and a softmax activated output layer. All ablation variants used the same training configuration as the full LSTM-CNN-GloVe architecture, with only the investigated architectural component removed.

The GloVe Word Embeddings model [32] is composed of five layers, including Input, Embedding, Global Max Pooling and two Dense layers. The Embedding layer employs pre-trained 100-dimensional GloVe embeddings to capture semantic relationships between words [33,34], whilst the pooling layer uses one-dimensional sequence to extract salient textual features from the review sequences [35,36].

The CNN-GloVe model is effective at capturing complex patterns and semantic relationships between words in text. Its architecture includes convolutional and pooling layers that learn and extract features from the text [35,37], followed by an output layer for classification. The implemented model consists of six layers: Input, Embedding, Conv1D, Global Max Pooling and two Dense layers. The Conv1D layer has 128 output filters and a

kernel size of 5, following [38], and is placed on top of an Embedding layer that uses 100-dimensional GloVe embeddings similar to [34,39]. ReLU and softmax activation functions were used in the hidden and output layers, respectively [37].

Table 4. Experimental configuration of the evaluated deep learning architectures.

Parameter	LSTM-CNN-GloVe	DistilBERT-CNN
Embedding source	GloVe (glove.6B.100d)	DistilBERT-base
Embedding dimension	100	768 (token-level)
Embedding trainability	Frozen	Frozen
Trainable parameters	175,000	492,422
Tokenizer	Keras	WordPiece
Sequence length	100 tokens	64 tokens
Conv1D filters/kernel	128 filters, kernel = 5, ReLU	128 filters, kernel = 5, ReLU
MaxPooling1D	pool size = 4	pool size = 4
GlobalMaxPooling1D	No	Yes
Dropout	None	0.3 ($\times 2$)
Output layer	Dense (6, softmax)	Dense (6, softmax)
Loss function	Sparse categorical cross-entropy	Sparse categorical cross-entropy
Optimizer	Adam (lr = 0.001)	Adam (lr = 2×10^{-5})
Batch size	32	16
Training epochs	Up to 10	Up to 5
Early stopping patience	5	2
Random seeds	42, 123, 456, 789, 1011	
Target classes	Run, Walk, Hike, Swim, Climb, Unknown	

The LSTM architecture consists of five layers: Input, Embedding, LSTM, Flatten and Dense. The Embedding layer transforms reviews into vector representations, whilst the LSTM layer captures sequential dependencies and high-level features [38,40]. The Flatten layer converts these features into a vector [39] and the Dense output layer uses the softmax activation function for classification. Prior studies have demonstrated the effectiveness of LSTM architectures for text classification tasks involving review datasets [35,41,42].

The LSTM-CNN-GloVe model is a hybrid neural network architecture that combines LSTM, CNN and GloVe Word Embeddings similarly to [40–42]. The model was designed to classify product reviews by capturing both local contextual features through CNN layers and longer-range sequential dependencies through LSTM layers [35,43]. This architecture predicts a single activity label for each product review.

This study also evaluated DistilBERT as a transformer-based approach for activity classification. DistilBERT is a compressed variant of BERT that applies knowledge distillation during pre-training to retain much of BERT's contextual language understanding capability while reducing computational complexity and inference time [44]. The model uses a stack of six transformer encoder layers with multi-head self-attention mechanisms to capture long-range contextual dependencies between words, making it well suited for identifying activity-related patterns within complex product reviews.

In addition to the standalone DistilBERT evaluation, a hybrid DistilBERT-CNN architecture was also tested to enable a direct comparison with the GloVe-based CNN models. In this configuration, DistilBERT replaces GloVe as the embedding source but retains the CNN classification head, enabling a direct comparison between static and contextual embedding approaches within a consistent architectural framework. Both the full dataset and a keyword-filtered subset were used for training.

Figure 2 illustrates the overall activity classification pipeline used in this study. Product review text is first pre-processed, as described in Section 3.3. The resulting text representations are then provided as input to the evaluated classification models, including the

LSTM-CNN-GloVe and DistilBERT architectures, to predict the most probable activity performed by a consumer using the product.

Raw product review

"Needed a water and outdoor shoe , taking them on a cruise and climbing Dunn's River. They are very ,very comfortable and all around summer shoe. True to size. Btw does not itch. And love the color choices !"

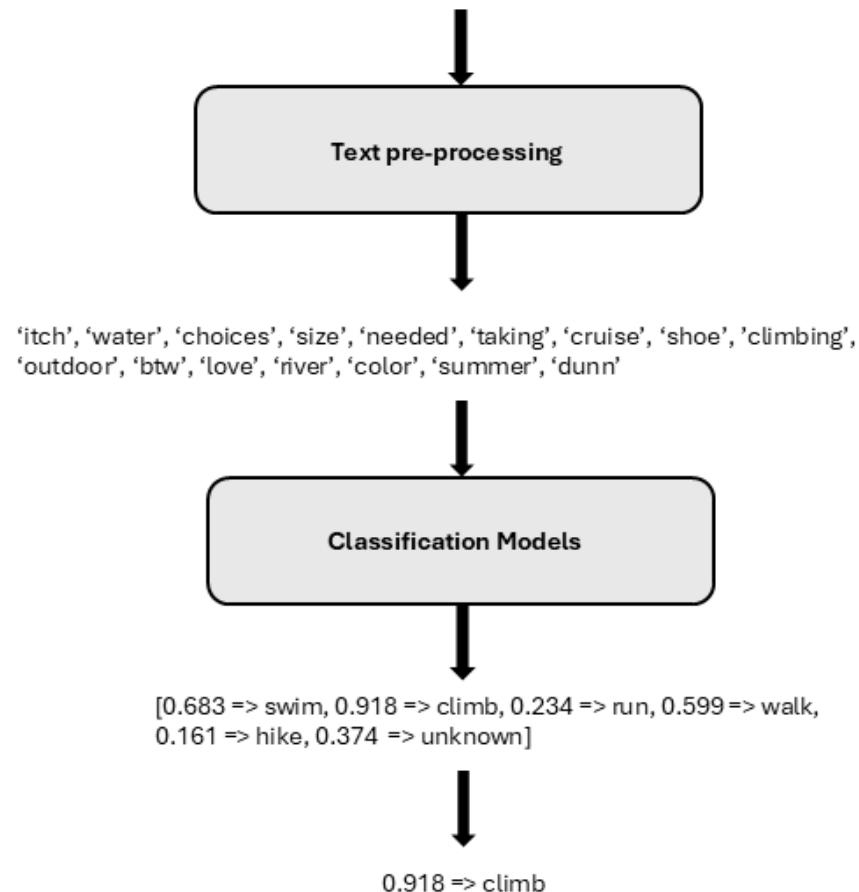


Figure 2. Activity classification pipeline example.

The following subsections describe the hybrid LSTM-CNN-GloVe and DistilBERT-CNN architectures in greater detail, as these were the primary deep learning configurations evaluated in this study.

3.4.2. LSTM-CNN-GloVe Architecture

The hybrid LSTM-CNN-GloVe architecture was evaluated as a lightweight deep learning model for activity classification in e-commerce product reviews, using the subset of product reviews described in Section 3.2. The model combines pre-trained GloVe word embeddings with convolutional and recurrent neural network layers to capture both local linguistic patterns and long-range contextual dependencies within review text.

Figure 3 illustrates the architecture of the hybrid model. The product reviews were pre-processed using the GloVe-based pipeline described in Section 3.3.1 at the document level, which includes tokenisation, POS tagging, stop-word removal and feature extraction. The Keras Tokenizer was then used to transform the reviews and their labels into a sequence of numerical values that were padded with zeros to ensure reviews are of equal length.

The architecture was implemented using the Keras Sequential API as a stack of neural network layers, with each pre-processed review represented as a sequence of word indices mapped to their corresponding GloVe word embeddings [45].

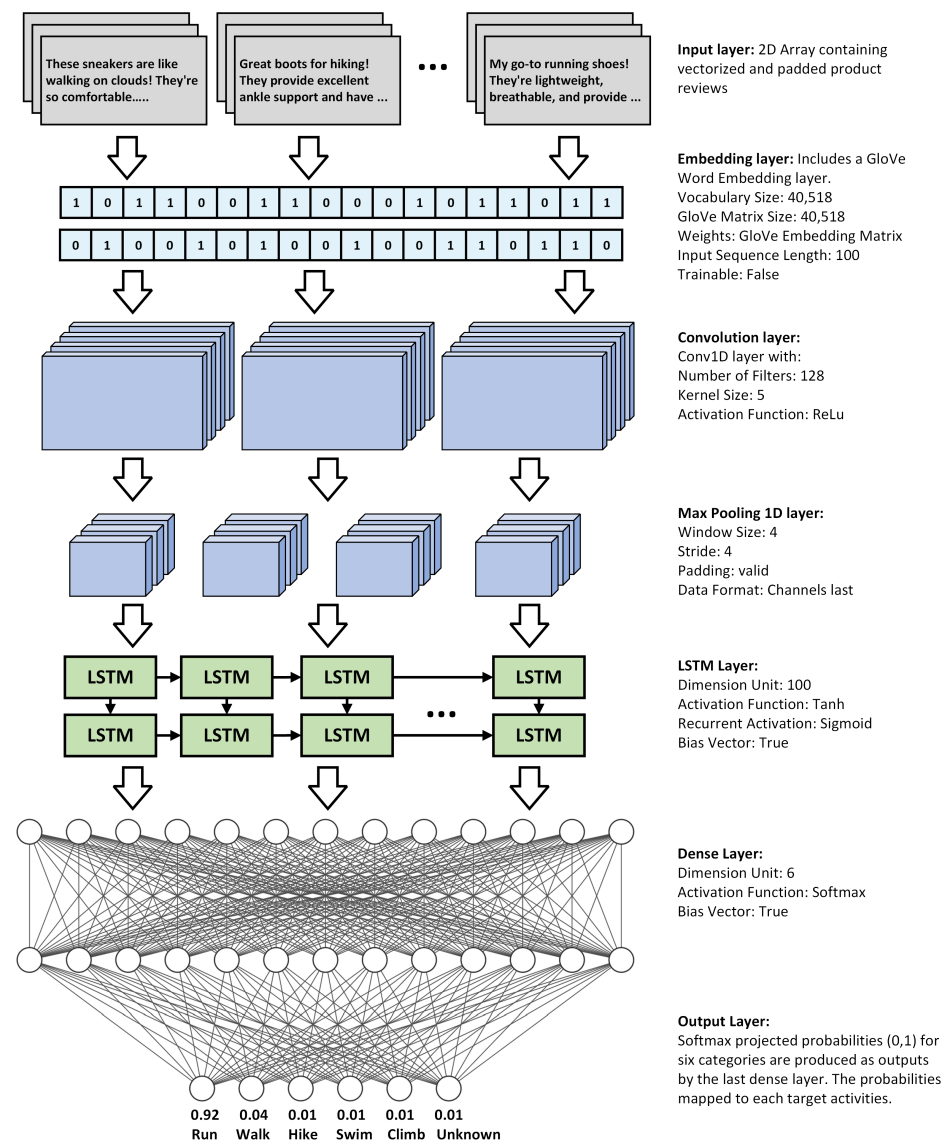


Figure 3. LSTM-CNN-GloVe neural network structure.

The Embedding layer used 100-dimensional GloVe word embeddings to map word sequences to word vectors, allowing statistical information from word-to-word occurrences within text to be used. Stopword removal, performed during pre-processing, helps reduce noise and improves the quality of the learned representations. The Embedding layer produced word representations with each pre-processed review represented as a sequence of word vectors. Equation (1) shows the mathematical formulation that concatenates word vectors w_i corresponding to the i^{th} word in the product review [45].

$$T = w_1 \oplus w_2 \oplus \dots \oplus w_n \in \mathbb{R}^{n \times d} \quad (1)$$

where $\oplus$ denotes vector concatenation, so that T is the embedding matrix whose i -th row is the vector of the i -th token.

A one-dimensional convolutional layer (Conv1D) with 128 output filters, a kernel size of 5 and a ReLU activation function was used in this model. The layer applied filters of

different widths to the embedding matrix to extract distinctive features, capturing syntactic or semantic associations between phrases that are far apart in the review text. Equation (2) shows the mathematical calculation used to filter through the embedding matrix and produce a feature map using a convolution operation. $F \in \mathbb{R}^m$ represents a filter with a width of m that shifts through the embedding matrix to produce a feature map where each c_i is calculated as described in [45]. The symbol $*$ represents the convolution operation, whilst $T_{i:i+m-1}$ represents the word vectors from w_i to w_{i+m-1} . The review text is represented by T , which is processed by the filter F . The bias term is represented by b_i and the ReLU activation function is represented by f .

$$c_i^{(j)} = f(\langle F^{(j)}, T_{i:i+m-1} \rangle + b^{(j)}) \quad i = 1 \dots n - m + 1 \quad (2)$$

where $T_{i:i+m-1} \in \mathbb{R}^{m \times d}$ is the window of token vectors from w_i to w_{i+m-1} , $\langle \cdot, \cdot \rangle$ denotes the sum of the element-wise product (the convolution operation), $b^{(j)} \in \mathbb{R}$ is the scalar bias shared across all positions of the filter j and $f(\cdot) = \max(0, \cdot)$ is the ReLU activation function. Therefore, each filter produces a feature map $c^{(j)} \in \mathbb{R}^{n-m+1}$. A Max Pooling layer with a pool size of 4 was added on top of a CNN layer to extract high-level features from a product review feature map. The resulting maximum-pooled feature vector was concatenated and provided as input to an LSTM layer, reducing the number of required parameters and helping to control overfitting. Equation (3) shows the operation performed by this layer in mathematical terms.

$$\hat{c}_t^{(j)} = \max(c_{p(t-1)+1}^{(j)}, c_{p(t-1)+2}^{(j)} \dots c_{pt}^{(j)}) \quad t = 1 \dots T' \quad (3)$$

where $T' = \lfloor (n - m + 1) / p \rfloor$ is the length of the pooled sequence. Stacking the K pooled feature maps in the position yields a sequence of feature vectors $x_t = (\hat{c}_t^{(1)} \dots \hat{c}_t^{(K)}) \in \mathbb{R}^K$, $t = 1, \dots, T'$, which serves as the input sequence to the LSTM layer. The combination reduces the number of downstream parameters and helps control overfitting.

The LSTM layer was used to capture long-range sequential dependencies within the review text by modelling contextual information across the input sequence. It is composed of three gates: the Forget Gate (f_t), the Input Gate (i_t) and the Output Gate (o_t), represented by Equations (4), (5) and (6), respectively.

$$f_t = \sigma(W_f [h_{t-1}; x_t] + b_f) \quad (4)$$

$$i_t = \sigma(W_i [h_{t-1}; x_t] + b_i) \quad (5)$$

$$o_t = \sigma(W_o [h_{t-1}; x_t] + b_o) \quad (6)$$

where $\sigma(\cdot)$ is the logistic sigmoid function, $[h_{t-1}; x_t]$ denotes the concatenation of the previous hidden state and the current input, and W_f, W_i, W_o and b_f, b_i, b_o are the corresponding weight matrices and bias vectors.

The final layer of the LSTM-CNN-GloVe architecture is a fully-connected dense layer that uses the softmax activation function (defined in Equation (7)) to produce the class probability distribution for multi-class classification. This layer was used to classify data into chunks across hidden layers, with every neuron in the input layer connected to every neuron in the output layer. The output layer contains 6 neurons representing the target activities for this classification, which are used to predict the activity a consumer performed based on probability scores.

$$P(y = j | h_{T'}) = \frac{\exp(W_j h_{T'} + b_j)}{\sum_{k=1}^C \exp(W_k h_{T'} + b_k)} \quad j = 1 \dots C \quad (7)$$

where $C = 6$ is the number of activity classes, and W_j and b_j denote the j^{th} row of the output weight matrix and the j^{th} element of the output bias vector, respectively. The predicted activity is the class with the highest probability. The Adam optimiser and sparse categorical cross-entropy loss function were used to train the model over 10 epochs, minimising the loss through backpropagation which helps optimise model performance. The loss function is calculated using Equation (8), with x^i as the i^{th} element of the dataset, y^i as the predicted label, t as the number of training samples, and θ as the parameters.

$$J(\theta) = -\frac{1}{t} \sum_{i=1}^t \log \left(P(Y = y^i | x^i, \theta) \right) \quad (8)$$

3.4.3. DistilBERT-CNN Architecture

In addition to the conventional deep learning architectures, this study evaluated a DistilBERT-CNN hybrid model that combines transformer-based contextual embeddings with a convolutional classification head. This architecture, shown in Figure 4, was designed to enable a direct comparison with the hybrid LSTM-CNN-GloVe model whilst also investigating whether contextualised language representations improve activity classification performance. The model follows the DistilBERT-based pre-processing pipeline described in Section 3.3.2, which relies on subword tokenisation and contextual representations rather than explicit linguistic feature extraction.

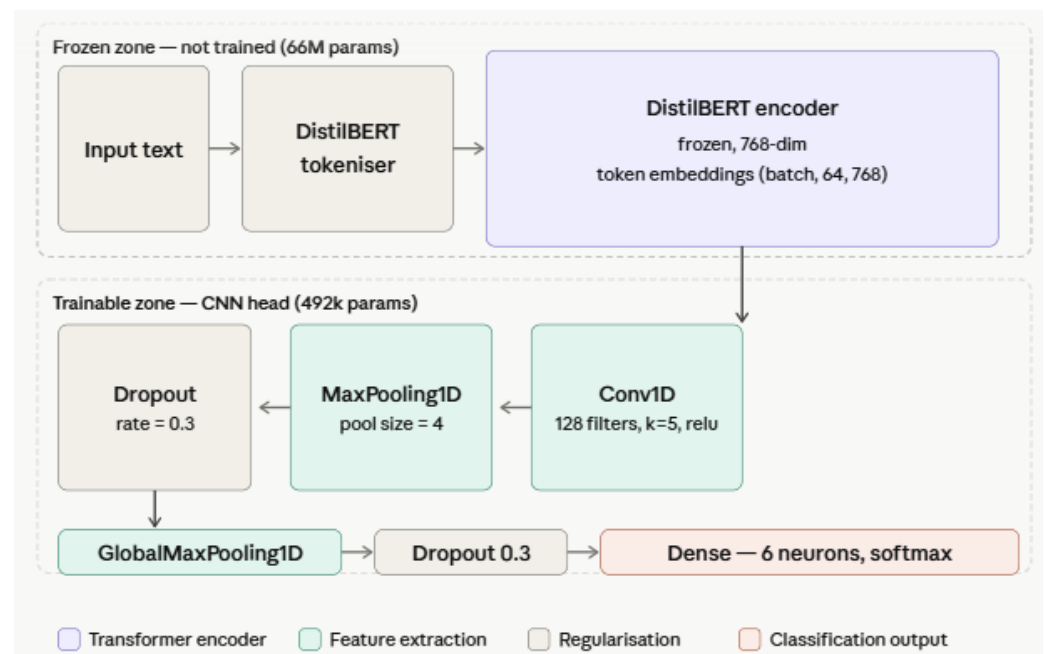


Figure 4. DistilBERT-CNN architecture.

In this configuration, DistilBERT replaced GloVe as the embedding source, whilst the CNN and max-pooling components were preserved from the original pipeline. The DistilBERT weights were frozen during training, meaning that only the CNN-based classification head was updated. Under this configuration, the model contained approximately 492,422 trainable parameters, reducing computational cost whilst still allowing the model to benefit from transformer-derived contextual embeddings.

The output embeddings from DistilBERT were passed to a one-dimensional convolutional layer containing 128 filters with a kernel size of 5 to detect local activity-related n-gram patterns from the contextual token sequence. This layer was intended to identify short contextual phrases and activity-related linguistic patterns, such as “went for a walk”, “used for running”, “hiking shoes”, or “climbing grip”.

A MaxPooling1D layer with a pool size of 4 was then applied to reduce the dimensionality of the feature maps and retain the most salient local features. Dropout with a rate of 0.3 was used after pooling to reduce overfitting by randomly disabling a proportion of activations during training.

A GlobalMaxPooling1D layer was then used to aggregate the strongest feature responses across the full review sequence, producing a fixed-length representation suitable for classification. A second Dropout layer with a rate of 0.3 was applied before the final Dense layer. The final Dense layer contained six output neurons corresponding to the target activity classes and used the softmax activation to generate class probability distributions. The predicted class was selected as the activity with the highest probability.

4. Results and Analysis

4.1. Overall Model Performance

This section presents the experimental evaluation of the investigated machine learning and deep learning approaches for activity classification in e-commerce product reviews. The objective of the experiments was not only to compare predictive performance across architectures but also to assess the robustness of activity classification under noisy and ambiguous real-world review data.

The experiments were conducted using the **annotated activity subset**, described in Section 3.2. Performance was evaluated using classification accuracy, precision, recall and macro F1-score. Class-level analysis, confusion matrices and ablation studies were also performed to further investigate model behaviour and identify sources of classification difficulty.

Table 5 presents the overall performance of the evaluated architectures on the **independent validation subset** (see Section 3.2). The reported values correspond to the mean and standard deviation across five training runs with different random seeds. The results show that despite their substantial differences in architectural complexity, the evaluated models achieved relatively similar performance levels on the original dataset.

Table 5. Performance of evaluated models.

Model	Accuracy (%)	Macro F1 (%)	Overfitting Gap (%)
LSTM-CNN-GloVe	63.70 ± 0.49	59.29 ± 2.49	0.017
DistilBERT	65.77 ± 0.58	62.01 ± 0.97	−0.005
DistilBERT-CNN	57.58 ± 0.46	56.63 ± 0.52	0.030

The hybrid LSTM-CNN-GloVe architecture achieved 63.70% accuracy and a macro F1-score of 59.29%, demonstrating stable performance and low overfitting across repeated runs. DistilBERT achieved the highest overall performance, with an accuracy of 65.77% and a macro F1-score of 62.01%, representing a modest improvement over the LSTM-CNN-GloVe architecture. This suggests that transformer-based contextual embeddings provide additional discriminative information, but do not fully resolve the ambiguity associated with activity prediction from e-commerce product reviews.

The DistilBERT-CNN architecture achieved lower performance than both DistilBERT and LSTM-CNN-GloVe. Although the model benefits from contextual embeddings, the addition of a CNN-based classification head did not improve activity classification performance on the **annotated activity subset**. This suggests that local convolutional feature extraction, when applied on top of contextual embeddings, is insufficient to address the substantial label ambiguity present in the reviews.

The relatively small standard deviations observed across experiments indicate that the evaluated models exhibit stable and reproducible behaviour under repeated training

conditions. The modest performance differences between architectures suggest that dataset ambiguity and overlapping activity labels impose a stronger limitation on classification performance than model complexity alone. Therefore, improvements in data quality and label separability may provide greater gains in performance than further increases in model complexity.

4.2. Ablation Study

The ablation experiments were conducted to evaluate the individual contribution of the CNN, LSTM and embedding components within the proposed architectures. Unlike the results reported in Table 5, which represent the mean performance across five random-seed runs on the **independent validation subset**, the results in Table 6 correspond to a single evaluation using a fixed training–validation split. Consequently, the numerical values are not expected to match exactly and should be interpreted as comparative results within the ablation study.

Table 6. Ablation Study of Results for the Hybrid LSTM–CNN–GloVe Architecture.

Model Variant	Accuracy (%)	Macro F1 (%)
LSTM-CNN-GloVe	64.85	57.04
LSTM-only	65.07	57.05
CNN-only	64.45	59.44
GloVe-avg	63.75	56.82

The results show that all evaluated model variants achieved comparable performance, suggesting that a substantial portion of the predictive signal is captured by the pre-trained embeddings. Removing the CNN component resulted in only a minor reduction in performance, whilst the LSTM-only configuration achieved performance comparable to the full LSTM-CNN-GloVe architecture (65.07% vs. 64.85% accuracy). This suggests that sequential modelling contributes meaningfully, although it is not strictly necessary to achieve competitive performance. The CNN-only model also produced competitive results, suggesting that local n-gram feature extraction alone captures a substantial proportion of the relevant information in the reviews. Its higher macro F1-score suggests that convolutional features may better capture class-balanced patterns.

The GloVe-avg configuration, which removes sequential and convolutional components, shows a modest reduction in performance (63.75% accuracy, 56.82% macro F1). This suggests that simple embedding averaging already captures much of the semantic information, whilst structured modelling components provide only incremental improvements.

The ablation results suggest that pre-trained embeddings contribute the majority of the model's effectiveness, whilst the CNN and LSTM components provide complementary but comparatively modest refinements.

4.3. Stability and Statistical Reliability

The repeated-seed evaluation further complements the comparative analysis by assessing the consistency of model performance across independent training runs.

Unlike many previous studies that report results based on a single experiment run, this study uses repeated experiments with different random seeds to obtain more reliable estimates.

The results show some variation across runs, confirming that performance is affected by random initialisation and training dynamics. However, the models are generally stable, as the observed variability remains relatively small.

These findings suggest that performance differences of small magnitude should be interpreted with caution. They also highlight the importance of reporting performance variability, as results based on a single run may lead to overly optimistic or non-reproducible conclusions.

4.4. Per-Class Performance Analysis

The class-wise results in Table 7 show substantial variation in performance across activity categories. The *walk*, *climb* and *swim* classes achieved the highest F1-scores (0.884, 0.901 and 0.936, respectively), indicating that these categories are more easily distinguishable. This is likely due to the presence of more specific and consistent vocabulary associated with these activities. In contrast, the *run*, *hike* and *unknown* classes showed considerably lower performance (F1-scores below 0.30). Both precision and recall remained low for these categories, suggesting that models found it difficult to distinguish between them. These differences indicate that classification performance is strongly influenced by the distinctiveness of activity-specific language. Classes with overlapping or ambiguous terminology present a greater challenge for the models, resulting in reduced classification performance.

Table 7. Class-wise Performance of DistilBERT-CNN on the Full 6-Class Dataset.

Class	Precision	Recall	F1-Score
Unknown	0.240	0.183	0.208
Run	0.252	0.231	0.241
Walk	0.848	0.924	0.884
Hike	0.258	0.315	0.284
Climb	0.903	0.899	0.901
Swim	0.929	0.943	0.936
Macro Average	0.572	0.583	0.576

To further investigate these class-specific differences, the following subsection analyses the confusion matrix to identify the principal sources of misclassification.

4.5. Confusion Matrix Analysis

Figure 5 presents the normalised confusion matrix averaged across five experimental runs. The matrix provides a detailed view of the model's performance across activity classes and highlights patterns of misclassification, particularly between activity classes that share similar language.

The strongest performance was observed for the *walk*, *swim* and *climb* classes, which achieved very high recall (approximately 0.98, 0.98 and 0.96, respectively). These classes are associated with more distinctive vocabulary, allowing the model to identify them with minimal confusion. Misclassifications for these categories are rare and are typically assigned to the *unknown* category.

The majority of prediction errors occurred for *run*, *hike* and *unknown* categories. The confusion matrix shows that instances from these classes were frequently predicted as *run*, with around 0.59 of both *hike* and *unknown* examples misclassified in this way.

These patterns suggest considerable overlap in the language used to describe these activities. The model appears to capture a broader notion of outdoor activity, which is often associated with the *run* class, leading to systematic misclassification.

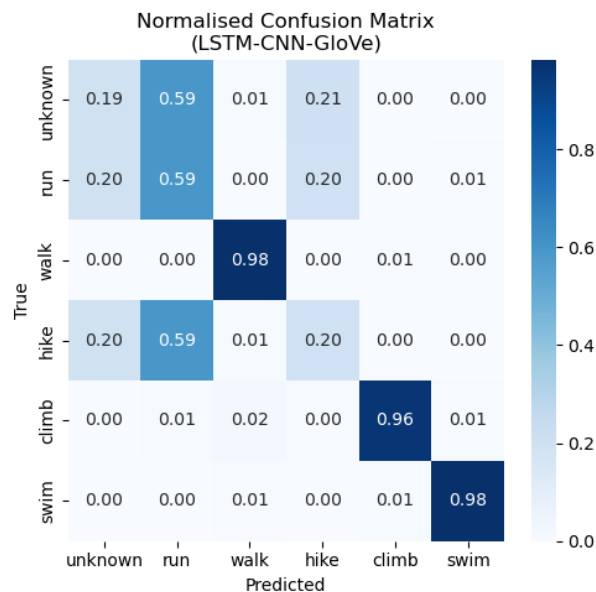


Figure 5. Normalised confusion matrix.

These findings confirm that the effectiveness of activity classification models depends strongly on the distinctiveness of activity-specific vocabulary within the review texts.

4.6. Experiments on Keyword-Filtered Dataset

To further investigate the impact of label ambiguity on model performance, an additional set of experiments was conducted using the **keyword-filtered subset** described in Section 3.2.

Both the LSTM-CNN-GloVe and DistilBERT-CNN architectures were trained using the same experimental protocol as in the main experiments to ensure comparability.

The results presented in Table 8 show a substantial improvement compared with the unfiltered dataset. The LSTM-CNN-GloVe model achieved an accuracy of 95.52%, whilst the DistilBERT-CNN model achieved an accuracy of 92.90% and a macro F1-score of 88.56%. These results represent a significant increase from the 63–66% accuracy range observed in the experiments conducted on the full dataset.

Table 8. Performance Comparison on the Full and Keyword-Filtered Datasets.

Model	Accuracy (%)	Macro F1 (%)
LSTM-CNN-GloVe	63.70	59.29
LSTM-CNN-GloVe (filtered)	95.52	76.61
DistilBERT-CNN	57.58	56.94
DistilBERT-CNN (filtered)	92.90	88.56

This substantial improvement indicates that model performance in the original task is primarily constrained by dataset ambiguity rather than architectural limitations. When explicit activity indicators are present, both traditional hybrid architectures and transformer-based models are able to learn discriminative patterns effectively. However, when activity descriptions are implicit, overlapping or context-dependent, performance degrades significantly across all models.

These results also demonstrate that increasing model complexity does not necessarily lead to improved performance in the presence of ambiguous data. The comparable performance of LSTM-CNN-GloVe and DistilBERT-CNN on the filtered dataset suggests that, given clear linguistic signals, simpler architectures can achieve results similar to those of

more computationally expensive models. Therefore, the experiments on the filtered dataset provide strong evidence that the primary challenge in activity classification from product reviews lies in the nature of the data rather than the choice of model. This highlights the importance of dataset quality, label clarity and task formulation in the development of robust activity prediction systems.

4.7. Comparison with Related Work

As discussed in Section 2, existing research on e-commerce product reviews has predominantly focused on sentiment analysis [7,15,26], opinion mining, review helpfulness prediction [10–12,21] and the use of deep learning techniques for extracting contextual information from reviews to enhance recommendation systems [46]. These approaches primarily aim to identify consumers' attitudes towards products, extract product features or assess the usefulness of reviews. Whilst such methods provide valuable insights into customer satisfaction and purchasing behaviour, they do not address how products are actually used in real-world contexts.

Research using deep learning has led to significant improvements in review classification tasks. Hybrid architectures combining convolutional and recurrent neural networks (e.g., CNN, LSTM, BiLSTM) and transformer-based models such as BERT and DistilBERT have demonstrated strong performance in sentiment analysis and opinion mining. However, these tasks differ fundamentally from activity classification. Sentiment categories are often characterised by explicit emotional expressions, whilst activity classification requires the identification of behavioural patterns and usage contexts that are frequently described ambiguously.

The results of this study indicate that activity classification from e-commerce reviews is a substantially more challenging task. All evaluated architectures achieved comparable performance on the **annotated activity subset**, with accuracy values ranging between 63% and 66% (see Table 5). Although DistilBERT achieved the highest overall performance, the improvement over the proposed LSTM-CNN-GloVe architecture was limited. This suggests that increased model complexity alone is insufficient to address the ambiguity inherent in real-world product reviews.

The confusion matrix analysis revealed considerable overlap between the *run*, *hike* and *unknown* categories, whilst activities associated with more distinctive vocabularies, such as *walk*, *swim* and *climb*, were classified with higher accuracy. This indicates that classification difficulty appears to be driven more by semantic overlap in activity descriptions rather than architectural limitations.

The experiments on the keyword-filtered subset reinforce this interpretation. When restricting the dataset to reviews containing explicit activity-related terminology, the DistilBERT-CNN model achieved an accuracy of 92.9% and a macro F1-score of 88.56%. The significant improvement observed under these conditions demonstrates that deep learning models are capable of accurately identifying product usage activities when explicit and unambiguous terms are present. The performance gap between the filtered and unfiltered datasets therefore highlights the challenges associated with real-world review data, where users describe activities in indirect or ambiguous ways.

The main contribution of this study lies in the automatic identification of product usage activities rather than sentiment or opinion classification. Unlike previous work that focused mainly on sentiment or rating prediction, this study extracts behavioural information that can support activity-oriented product search, recommendation and product discovery in e-commerce platforms.

Table 9 summarises the key differences between the proposed approach and the most relevant studies. Whilst prior work employs similar deep learning architectures, including

CNNs, LSTMs, BiLSTMs and transformer-based models, these approaches are primarily designed for sentiment classification, rating prediction or contextual representation learning. In contrast, this study applies comparable modelling techniques to a more complex task, the classification of product usage activities.

Table 9. Comparison of proposed approach with related work.

Papers	Task	Technique	Relevance to Activity Classification
Sindhura et al. [15]	sentiment analysis	WDE-CNN-LSTM	does not capture product usage
Ahmed and Wang [7]	sentiment analysis	BiLSTM-BERT-CNN	no activity-related information
Ali et al. [26]	sentiment analysis	BiLSTM-BERT	focused on opinion, not usage
Suryana et al. [46]	review representation for recommendation	GloVe-LSTM	does not perform activity classification
Amirifar et al. [21]	product rating prediction	RBFNN and NER	no activity-related information
This study	activity classification	LSTM-CNN-GloVe, DistilBERT	identifies product usage activity

4.8. E-Commerce Experience Evaluation

This section discusses the potential implications of the proposed approach for improving the e-commerce experience in comparison with existing studies outlined in Section 4.7.

The primary motivation of this research is to enhance the user experience of e-commerce platforms by enabling consumers to identify products that are suitable for specific activities. In contrast to traditional review analysis methods, this approach focuses on extracting activity-related information directly from product reviews, therefore reducing the need for manual review inspection.

Previous studies have primarily focused on sentiment analysis and recommendation tasks. For example, Sindhura et al. [15] focused on classifying product reviews to determine the emotional sentiment expressed by consumers towards products. Ahmed and Wang [7] identified sentiment expressed by consumers towards specific key aspects of a product. Although these approaches provide valuable insights into customer opinions, they do not explicitly capture how products are used in practice. As a result, users may still need to manually examine review content to determine whether a product is suitable for their intended activity.

The approach presented in this study enables the automatic identification of product usage activities directly from product review text. When integrated into an e-commerce platform, this functionality could provide summarised activity-based insights, allowing users to quickly assess product suitability without reading large volumes of reviews. This capability has the potential to support more efficient and targeted decision-making, particularly in scenarios where product suitability depends on specific usage contexts. By highlighting activity-related information, the proposed approach complements existing sentiment analysis and recommendation systems, contributing to a more comprehensive and user-centred e-commerce experience.

5. Conclusions

In this research, we investigated the feasibility of product activity classification from e-commerce reviews using deep learning and transformer-based models. Unlike sentiment analysis, which focuses on identifying users' opinions, this study addresses the more complex task of inferring product usage activities from unstructured review text.

The results of our experiments demonstrate that activity classification is a challenging problem, with all evaluated models achieving comparable performance (63–66% accuracy) on the full dataset. The analysis shows that performance limitations are primarily caused by ambiguity and semantic overlap in activity descriptions rather than model architecture. This interpretation is supported by experiments conducted on a keyword-filtered dataset, where performance increased significantly (up to 92.9% accuracy), demonstrating that modern deep learning models are capable of accurately identifying activities when explicit activity-related terms are present.

This paper establishes activity classification as a distinct and challenging natural language processing task and highlights the importance of dataset quality and label clarity in achieving reliable performance. The proposed approach has the potential to enhance the e-commerce experience by enabling users to identify products suitable for their intended activities without the need to manually inspect large volumes of product reviews on platforms such as Amazon.

Several limitations of this research should be acknowledged. First, the models were trained on a limited set of predefined activity classes, which may restrict their applicability to broader use cases. Extending the dataset to include a wider range of activities would improve the robustness and applicability of the approach. Second, the current implementation operates as an offline, standalone system, requiring manual input of product reviews. Future work could explore the development of an integrated application or browser-based interface to enable real-time usage within e-commerce platforms. Third, improving the quality and consistency of activity annotations, as well as reducing semantic overlap between activity categories, may further enhance model performance. Future research could also investigate more advanced strategies for handling ambiguity, such as multi-label classification.

This study demonstrates that, while deep learning models can effectively capture activity-related patterns, the primary challenge lies in the nature of real-world data. Addressing data ambiguity and improving annotation quality are therefore key directions for advancing activity classification in e-commerce contexts.

Author Contributions: Conceptualization: T.W. and C.L.; Data curation: T.W.; Formal Analysis: T.W.; Investigation: T.W. and B.K.; Methodology: T.W., M.M.D.O., C.L., A.F. and B.K.; Project administration: C.L.; Software: T.W. and B.K.; Supervision: C.L., A.F. and M.M.D.O.; Validation: A.F., C.L., M.M.D.O. and B.K.; Visualisation: T.W.; Writing—original draft: T.W.; Writing—review and editing: T.W., M.M.D.O., C.L. and A.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data used in this paper is available from <https://figshare.com/s/d7cf7839755a33c682de> (accessed on 16 July 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Abou Setta, I.G.; Shehata, O.M.; Awad, M.A. Multivariate Prediction of Correct Lane for Autonomous Electric Vehicle Using Deep Learning Models. In Proceedings of the 2020 8th International Conference on Control, Mechatronics and Automation (ICCMA), Moscow, Russia, 6–8 November 2020; pp. 127–130. [\[CrossRef\]](#)
2. Singh, S.N.; Sarraf, T. Sentiment analysis of a product based on user reviews using random forests algorithm. In Proceedings of the 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 29–31 January 2020; pp. 112–116. [\[CrossRef\]](#)
3. Shrestha, N.; Nasoz, F. Deep learning sentiment analysis of amazon. com reviews and ratings. *arXiv* **2019**, arXiv:1904.04096. [\[CrossRef\]](#)
4. Rani, D.; Kumar, R.; Chauhan, N. Study and Comparision of Vectorization Techniques Used in Text Classification. In Proceedings of the 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 3–5 October 2022; pp. 1–6. [\[CrossRef\]](#)
5. Sindhu, C.; Adak, S.; Tigga, S.C. Opinionated text classification for hindi tweets using deep learning. In Proceedings of the 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 8–10 April 2021; pp. 1217–1222. [\[CrossRef\]](#)
6. Chao, C.; Jiang, W. Study on the Subjective and Objective Text Classification and Pretreatment of Chinese Network Text. In Proceedings of the 2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics, Nanchang, China, 26–27 August 2012; Volume 2, pp. 25–29. [\[CrossRef\]](#)
7. Ahmed, Z.; Wang, J. A fine-grained deep learning model using embedded-CNN with BiLSTM for exploiting product sentiments. *Alex. Eng. J.* **2023**, *65*, 731–747. [\[CrossRef\]](#)
8. Ma, T.M.; Yamamori, K.; Thida, A. A comparative approach to Naïve Bayes classifier and support vector machine for email spam classification. In Proceedings of the 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE), Kobe, Japan, 13–16 October 2020; pp. 324–326. [\[CrossRef\]](#)
9. Kumar, S.; Sharma, A.; Reddy, B.K.; Sachan, S.; Jain, V.; Singh, J. An intelligent model based on integrated inverse document frequency and multinomial Naive Bayes for current affairs news categorisation. *Int. J. Syst. Assur. Eng. Manag.* **2021**, *13*, 1341–1355. [\[CrossRef\]](#)
10. Wu, S.H.; Wang, J.W. Integrating neural and syntactic features on the helpfulness analysis of the online customer reviews. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Vancouver, BC, Canada, 27–30 August 2019; pp. 1013–1017. [\[CrossRef\]](#)
11. Sun, X.; Han, M.; Feng, J. Helpfulness of online reviews: Examining review informativeness and classification thresholds by search products and experience products. *Decis. Support Syst.* **2019**, *124*, 113099. [\[CrossRef\]](#)
12. Wu, S.H.; Chen, Y.K. Cross-domain helpfulness prediction of online consumer reviews by deep learning model. In Proceedings of the 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI), Las Vegas, NV, USA, 11–13 August 2020; pp. 412–418. [\[CrossRef\]](#)
13. Elmurngi, E.I.; Gherbi, A. Unfair reviews detection on amazon reviews using sentiment analysis with supervised learning techniques. *J. Comput. Sci.* **2018**, *14*, 714–726. [\[CrossRef\]](#)
14. Zhao, W.; Guan, Z.; Chen, L.; He, X.; Cai, D.; Wang, B.; Wang, Q. Weakly-supervised deep embedding for product review sentiment analysis. *IEEE Trans. Knowl. Data Eng.* **2017**, *30*, 185–197. [\[CrossRef\]](#)
15. Sindhura, S.; Praveen, S.P.; Safali, M.A.; Rao, N. Sentiment analysis for product reviews based on weakly-supervised deep embedding. In Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2–4 September 2021; pp. 999–1004. [\[CrossRef\]](#)
16. Sadhana, S.; SaiRamesh, L.; Sabena, S.; Ganapathy, S.; Kannan, A. Mining target opinions from online reviews using semi-supervised word alignment model. In Proceedings of the 2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), Tindivanam, India, 3–4 February 2017; pp. 196–200. [\[CrossRef\]](#)
17. Nasim, Z.; Rajput, Q.; Haider, S. Sentiment analysis of student feedback using machine learning and lexicon based approaches. In Proceedings of the 2017 International Conference on Research and Innovation in Information Systems (ICRIIS), Langkawi, Malaysia, 16–17 July 2017; pp. 1–6. [\[CrossRef\]](#)
18. Yao, D.; Chatterjee, I.; Zhou, M. Conditioning Customers' Product Reviews for Accurate Classification Performance. In Proceedings of the 2022 IEEE International Conference on Networking, Sensing and Control (ICNSC), Shanghai, China, 15–18 December 2022; pp. 1–5. [\[CrossRef\]](#)
19. Wahyudi, E.; Kusumaningrum, R. Aspect based sentiment analysis in E-commerce user reviews using Latent Dirichlet Allocation (LDA) and Sentiment Lexicon. In Proceedings of the 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, 29–30 October 2019; pp. 1–6. [\[CrossRef\]](#)
20. Suryadi, D.; Kim, H.M. A data-driven approach to product usage context identification from online customer reviews. *J. Mech. Des.* **2019**, *141*, 121104. [\[CrossRef\]](#)

21. Amirifar, T.; Lahmiri, S.; Zanjani, M.K. An NLP-Deep Learning Approach for Product Rating Prediction Based on Online Reviews and Product Features. *IEEE Trans. Comput. Soc. Syst.* **2024**, *11*, 8156–8168. [\[CrossRef\]](#)
22. Hayatin, N.; Alias, S.; Hung, L.P. Trends and challenges in sentiment summarization: A systematic review of aspect extraction techniques. *Knowl. Inf. Syst.* **2024**, *66*, 3671–3717. [\[CrossRef\]](#)
23. Krichen, M.; Mihoub, A. Long Short-Term Memory Networks: A Comprehensive Survey. *AI* **2025**, *6*, 215. [\[CrossRef\]](#)
24. Xie, X.; Qin, B.; Wan, Z.; Nie, W. Text Aspect-Level Sentiment Analysis based on Multi-Task Joint Learning. In Proceedings of the 2021 2nd International Symposium on Computer Engineering and Intelligent Communications (ISCEIC), Nanjing, China, 6–8 August 2021; pp. 127–131. [\[CrossRef\]](#)
25. Al-Garadi, M.A.; Yang, Y.C.; Cai, H.; Ruan, Y.; O'Connor, K.; Gonzalez-Hernandez, G.; Perrone, J.; Sarker, A. Text classification models for the automatic detection of nonmedical prescription medication use from social media. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 27. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Ali, H.; Hashmi, E.; Yayilgan Yildirim, S.; Shaikh, S. Analyzing Amazon Products Sentiment: A Comparative Study of Machine and Deep Learning, and Transformer-Based Techniques. *Electronics* **2024**, *13*, 1305. [\[CrossRef\]](#)
27. Xiaoyan, L.; Raga, R.C.; Xuemei, S. GloVe-CNN-BiLSTM Model for Sentiment Analysis on Text Reviews. *J. Sens.* **2022**, *2022*, 7212366. [\[CrossRef\]](#)
28. Pimpalkar, A.; Raj R, J.R. MBiLSTM GloVe: Embedding GloVe knowledge into the corpus using multi-layer BiLSTM deep learning model for social media sentiment analysis. *Expert Syst. Appl.* **2022**, *203*, 117581. [\[CrossRef\]](#)
29. Bellar, O.; Baina, A.; Ballafkih, M. Sentiment Analysis: Predicting Product Reviews for E-Commerce Recommendations Using Deep Learning and Transformers. *Mathematics* **2024**, *12*, 2403. [\[CrossRef\]](#)
30. Alghaligah, A.; Alotaibi, A.; Abbas, Q.; Alhumoud, S. Optimized Hybrid Deep Learning for Enhanced Spam Review Detection in E-Commerce Platforms. *Int. J. Adv. Comput. Sci. Appl.* **2025**, *16*, 348–357. [\[CrossRef\]](#)
31. Arthamevia, N.P.; Purbolaksono, M.D. Aspect-Based Sentiment Analysis in Beauty Product Reviews Using TF-IDF and SVM Algorithm. In Proceedings of the 2021 9th International Conference on Information and Communication Technology (ICoICT), Yogyakarta, Indonesia, 3–5 August 2021; pp. 197–201. [\[CrossRef\]](#)
32. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543. [\[CrossRef\]](#)
33. Hasan, M.M.; Zaman, S.M.; Talukdar, M.A.; Siddika, A.; Alam, M.G.R. An Analysis of Machine Learning Algorithms and Deep Neural Networks for Email Spam Classification using Natural Language Processing. In Proceedings of the 2021 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), Singapore, 11–12 December 2021; pp. 1–6. [\[CrossRef\]](#)
34. Shrivastava, P.; Sharma, D.K. Fake Content Identification Using Pre-Trained Glove-Embedding. In Proceedings of the 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 22–23 October 2021; pp. 1–6. [\[CrossRef\]](#)
35. Venkatesh.; Hegde, S.U.; Zaiba, A.; Nagaraju, Y. Hybrid CNN-LSTM Model with GloVe Word Vector for Sentiment Analysis on Football Specific Tweets. In Proceedings of the 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 19–20 February 2021; pp. 1–8. [\[CrossRef\]](#)
36. Rudkowsky, E.; Haselmayer, M.; Wastian, M.; Jenny, M.; Emrich, Š.; Sedlmair, M. More than bags of words: Sentiment analysis with word embeddings. *Commun. Methods Meas.* **2018**, *12*, 140–157. [\[CrossRef\]](#)
37. Ahmed, F.; Akther, N.; Hasan, M.; Chowdhury, K.; Mukta, M.S.H. Word Embedding Based News Classification by Using CNN. In Proceedings of the 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), Pekan, Malaysia, 24–26 August 2021; pp. 609–613. [\[CrossRef\]](#)
38. Shelar, A.; Moharir, M. Predicting Outcomes of Court Judgments—A Machine Learning Approach. In Proceedings of the 2021 International Conference on Intelligent Technologies (CONIT), Hubli, India, 25–27 June 2021; pp. 1–6. [\[CrossRef\]](#)
39. Kalaivani, K.; Uma, S.; Kanimozhiselvi, C. Comparison of Deep Learning Approaches for Sentiment Classification. In Proceedings of the 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 20–22 January 2021; pp. 1043–1047. [\[CrossRef\]](#)
40. Zhang, J.; Li, Y.; Tian, J.; Li, T. LSTM-CNN hybrid model for text classification. In Proceedings of the 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 12–14 October 2018; pp. 1675–1680. [\[CrossRef\]](#)
41. Rasool, A.; Jiang, Q.; Qu, Q.; Ji, C. WRS: A novel word-embedding method for real-time sentiment with integrated LSTM-CNN model. In Proceedings of the 2021 IEEE International Conference on Real-Time Computing and Robotics (RCAR), Xining, China, 15–19 July 2021; pp. 590–595. [\[CrossRef\]](#)

42. Zhang, Y. Research on Text Classification Method Based on LSTM Neural Network Model. In Proceedings of the 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 14–16 April 2021; pp. 1019–1022. [\[CrossRef\]](#)
43. Bai, X. Text classification based on LSTM and attention. In Proceedings of the 2018 Thirteenth International Conference on Digital Information Management (ICDIM), Berlin, Germany, 24–26 September 2018; pp. 29–32. [\[CrossRef\]](#)
44. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108. [\[CrossRef\]](#)
45. Parwez, M.A.; Abulaish, M.; Jahiruddin. Multi-label classification of microblogging texts using convolution neural network. *IEEE Access* **2019**, *7*, 68678–68691. [\[CrossRef\]](#)
46. Suryana, N.; Basari, A. Generate Contextual Insight of Product Review Using Deep LSTM and Word Embedding. *J. Phys. Conf. Ser.* **2020**, *1577*, 012006. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.