Information Sciences 606 (2022) 453–468

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Gender-based multi-aspect sentiment detection using multilabel learning

J. Ashok Kumar [a], Tina Esther Trueman [a], Erik Cambria [b,*]

[a] Department of Information Science and Technology, Anna University, Chennai, India
[b] School of Computer Science and Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 26 April 2021
Received in revised form 10 May 2022
Accepted 15 May 2022
Available online 21 May 2022

Keywords:
Aspect-based sentiment analysis
Multilabel learning
Gender-based sentiment
Feature representation
Deep learning

ABSTRACT

Sentiment analysis is an important task in the field of natural language processing that aims to gauge and predict people's opinions from large amounts of data. In particular, gender-based sentiment analysis can influence stakeholders and drug developers in real-world markets. In this work, we present a gender-based multi-aspect sentiment detection model using multilabel learning algorithms. We divide Abilify and Celebrex datasets into three groups based on gender information, namely: male, female, and mixed. We then represent bag-of-words (BoW), term frequency-inverse document frequency (TF-IDF), and global vectors for word representation (GloVe) based features for each group. Next, we apply problem transformation approaches and multichannel recurrent neural networks with attention mechanism. Results show that traditional multilabel transformation methods achieve better performance for small amounts of data and long-range sequence in terms of samples and labels, and that deep learning models achieve better performance in terms of mean test accuracy, AUC Score, RL, and average precision using GloVe word embedding features in both datasets.

1. Introduction

In recent years, sentiment analysis has become increasingly popular for processing social media data on online communities, blogs, wikis, microblogging platforms, and other online collaborative media [7]. Besides traditional domains like business intelligence [27,3], sentiment analysis applications also include many other areas like financial forecasting [48], healthcare [6], cyber-harassment prevention [24], political forecasting [20], and dialogue systems [29].

While most works approach it as a simple categorization problem, sentiment analysis is actually a complex research problem that requires tackling many NLP tasks [8] tasks such as microtext normalization [36], to decode informal text, anaphora resolution [40], to link pronouns with the entities of a sentence, personality recognition [30], for distinguishing between different personality types of the users, and more. Even though the subjectivity and polarity-based sentiment analysis provide a block box solution to stakeholders such as the classification of negative, neutral, and positive reviews. They are not finding many interesting results in detail like the sentiment on entities based on gender [45,31,5].

Aspect-based sentiment analysis (ABSA) solves this issue by detecting sentiment on different entities of a product or service [47,25,28]. The term aspect defines a component or an attribute of a thing in a review. Overall, ABSA involves two tasks, namely, aspect identification and sentiment classification with respect to the aspect. A document or review may be

[*] Corresponding author.

https://doi.org/10.1016/j.ins.2022.05.057

associated with a single aspect or multiple aspects. The categorization of these aspects based on gender and age provides the most important information to stakeholders [26,33].

In this study, we use a real-life online health web forum dataset, called Abilify or aripiprazole oral user reviews and Celebrex oral user reviews. Abilify is used to treat mental disorders such as depression, anxiety, mania, and negative symptoms [35]. It improves concentration and decreases the hallucinations of the user. Celebrex is used to treat acute pain, inflammation, arthritis, and menstrual pain. The reaction of the drug also varies from one individual to another. Therefore, Abilify users and Celebrex users express their opinions based on their reaction in terms of comments or reviews. These reviews were associated with three important aspects such as satisfaction, ease of use, and effectiveness and their corresponding sentiments [23]. As these reviews were associated with more than one aspect (label), the problem leads to multilabel learning.

Traditionally, this multilabel learning problem can be solved using the problem transformation method, adapted method, and an ensemble method. The problem transformation method uses a binary relevance, classifier chain, and label powerset techniques to solve the multilabel problems. Recently, deep learning models have shown promising results in the field of opinion mining and sentiment analysis due to their long-term dependencies [41,10]. In particular, gender-based multi-aspect sentiment identification becomes another challenging problem such as lack of data availability based on gender and variation in age-groups. However, Abilify and Celebrex users and caregivers are providing drug's reaction and their effectiveness, satisfaction, and ease of use in online health forums with gender information. For instance, a male patient in the age-group of 19–24 expressed in a review "After taking a lot of medications and no luck, I tried this one it works well. After two days I noticed a difference. Well worth a try" and a female patient in the age-group of 45–54 stated "It's starting to work and I'm trusting it will continue to do so" with positive sentiment on effectiveness, satisfaction, and ease of use. Therefore, we propose a gender-based multi-aspect sentiment detection model using multilabel learning. Specifically, we use both traditional learning and deep learning models to identify multi-aspects. This paper mainly contributes to the following:

- Identifies multi-aspect sentiment based on genders such as male, female, or both male and female.
- Explores different feature representations such as BoW, TF-IDF, and GloVe.
- Compares the problem transformation approach, adapted algorithms, and attention-based recurrent neural networks (RNN).

The rest of this paper is arranged as follows: Section 2 describes related works in gender-based sentiment analysis; Section 3 presents gender-based multi-aspect sentiment detection using multilabel learning algorithms; Section 4 explains multilabel evaluation metrics; Section 5 presents experimental results; finally, Section 6 offers concluding remarks and discusses future directions.

2. Related works

The gender-based model has attracted researchers in the fields of natural language processing (NLP), linguistics [12,19], psychology [41], image processing [42], abusive language detection [32], e-commerce [50], science and engineering education [9,46,2,43], and healthcare [33,16,15]. Specifically, a gender-based sentiment analysis model allows stakeholders to analyze their business using the social network and linguistic data such as text, voice, image, and video. Therefore, we describe the recent developments in sentiment analysis based on gender. Ertam [42] and Kacamarga et al. [19] predicted gender from voice data using a deeper long short-term memory (LSTM) network and multi-layer perceptron (MLP), respectively. The authors used 20 acoustic features such as mean frequency, standard deviation, median frequency, etc. Kiritchenko et al. [21] examined gender and bias in sentiment analysis. Their study indicated that several systems have shown significant bias with a little higher sentiment intensity for one gender or one race. Park et al. [33] measured gender biases in abusive language detection with three methods, namely, word embeddings, data augmentation, and model fine-tuning. The authors found that these methods reduce 90–98% of gender biases. Moreover, Carver et al. [9] and Treceñe [46] reported the sociotechnical issues about gender and sentiment in software engineering education and students' learning diaries, respectively. Amarasekara and Grant [2] explored the gender gap sentiment analysis using a Univariate Analysis of Variance (ANOVA) in STEM (Science, Technology, Engineering, and Mathematics) YouTube channels. Their study indicated that female hosted channels accumulate more comments. Tannenbaum et al. [43] studied sex and gender analysis to improve efficiency and social equality in scientific discovery. This study provided a roadmap to researchers, journals, funding agencies, and universities to implement sex and gender analysis.

Social media contents such as reviews, posts, comments, and tweets become a relevant source for sentiment analysis tasks. Hussein et al. [17] addressed the gender identification problem for Arabic dialects on Twitter. The authors proposed a mixed feature vector, in addition to N-gram vectors. Moreover, they applied ensemble weighted average to the Random Forest algorithm with the mixed feature vector and the logistic regression algorithm with N-gram vectors. Their results achieve 87.6% accuracy for gender identification. Zheng et al. [50] proposed an MLP model with sentiment representation for gender classification. This model trains e-commerce reviews using LSTM and transfers their sentiment representation to the MLP model. The authors improved 5.53% for the gender classification task. Diaz et al. [11] studied the age-related bias on 15 sentiment analysis algorithms and 10 GloVe word embeddings. Their study finds that the age-related bias is significantly reduced using a custom-built classifier. Their result indicates that significant age bias encoded in sentiment analysis models and word embeddings reduces bias. Sboev et al. [37] evaluated the deep learning framework for the task of text categorization based on the author's gender and text sentiment identification. They performed different machine learning techniques such as a simple model only with three neurons, stacked LSTM, and stacked GRU (Gated recurrent units) on Russian language texts for both tasks. Branz et al. [4] analyzed emotional content in developer comments in terms of gender for open source projects. The authors used decision tree classifiers to categorize commenters' gender such as male or female based on their sentiment. They used the oversampling method to balance the imbalanced gender ratio in developer comments. Furthermore, Gallego and Corchuelo [13] presented a system called Torri, which collects reviews, reports aspects, and their sentiment by considering conditions in the sentence. They employed an encoder-decoder model with GRU and BiGRU (Bidirectional Gated Recurrent Units), respectively. Suman et al. [41] presented a system to recognize age, gender, and dialects in a tweet. They used the LSTM network with hand-crafted features to recognize the author's profile. Teso et al. [44] presented the application of text mining techniques to reveal the gender discourse and their preferences in a specific product category. The authors used a set of frequent keywords to predict users' gender. Kumar et al. [26] explored the impact of gender and age in sentiment analysis to market the product based on demographics. They collected book reviews based on age and gender from Facebook users in questionnaires. The authors used support vector machine (SVM), maximum entropy, CNN, and LSTM networks to study the impact of gender and age.

Researchers have also focused on gender-based sentiment analysis in healthcare applications apart from reviews, posts, comments, tweets, or dialects. Park and Woo [33] presented a gender classification model using sentiment analysis and machine learning techniques in a health web forum. They investigated that senti-words generates better accuracy than sentiment features with SVM. Also, they suggested that there is a high misclassification rate in the female category using traditional machine learning techniques. But, the authors overcome this issue using a CNN model with over 90% accuracy. Haimson [16] used social media data to investigate gender transition sentiment patterns in mental health disparities. The author suggested that sentiment increases on gender transition with supportive responses to mass disclosure and transgender identity disclosure. Garcia-Rudolph et al. [15] studied stroke survivors' tweets using the frequency of words and compared their posts by gender for basic emotions such as anger, anticipation, joy, etc. Their study investigated that women's topic scores a higher level of happiness. The existing researchers are not much focused on gender-based sentiment analysis with multi-aspect category identification in drugs and medication reviews. In this paper, we propose the problem transformation approaches, adapted algorithms, and recurrent neural networks for the task of gender-based multi-aspect sentiment detection.

3. The proposed model

In this section, we present the gender-based multi-aspect sentiment detection model using multilabel learning algorithms. The proposed model architecture is shown in Fig. 1. It consists of four main parts: input data, data preprocessing, feature representation such as Bag of Words (BoW), TF-IDF, and word embeddings, multilabel learning algorithms in traditional machine learning and deep learning techniques. The following sections explain each of the parts in a multilabel environment.

3.1. Input data

We use the Abilify Oral and Celebrex Oral user reviews dataset [23], which are collected from an online health service provider (WebMD). The Abilify dataset consists of 1722 and Celebrex Oral consists of 2438 instances. Each instance includes a condition of the patient, age, gender, type (patient or caregiver), duration of the treatment, drug users comments, and

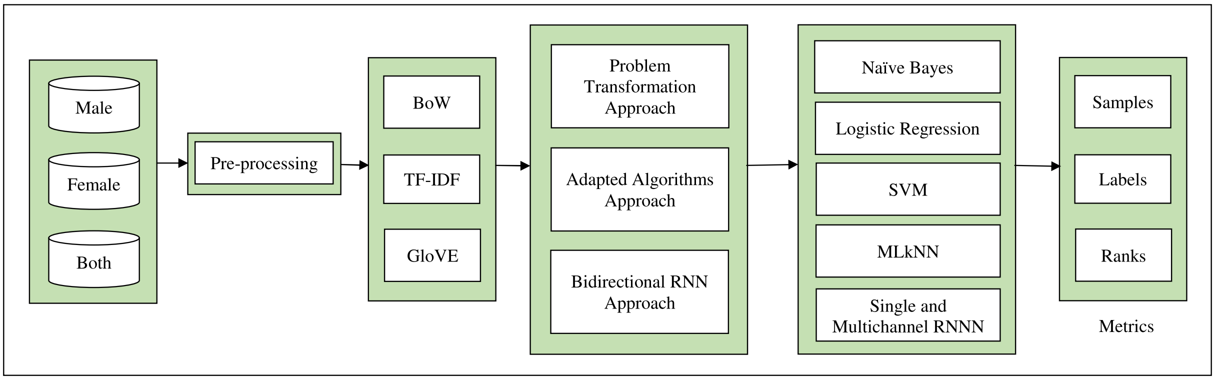
Male
Female
Both
Pre-processing
BoW
TF-IDF
GloVE
Problem Transformation Approach
Adapted Algorithms Approach
Bidirectional RNN Approach
Naïve Bayes
Logistic Regression
SVM
MLkNN
Single and Multichannel RNNN
Samples
Labels
Ranks
Metrics

Fig. 1. Proposed model.

drug's aspect sentiment ratings namely, satisfaction, ease of use, and effectiveness. In this paper, we use only 1615 gender-based instances (397 male and 1218 female) from the Abilify dataset and 2438 gender-based instances (533 male and 1905 female) from the Celebrex dataset. We have omitted some of the instances from the Abilify dataset which have not explicitly define their gender information.

3.2. Feature representation

3.2.1. Bag of words

The Bag of Words (BoW) model is a simple model used for text representation [26]. This model creates a list of words in the documents or sentences and forms a matrix based on word frequencies or the number of occurrences. However, the BoW model ignores document or sentence structure, grammar, and semantic meaning between words.

3.2.2. TF-IDF (Term Frequency-Inverse Document Frequency)

Similar to the BoW model, term frequency (TF) counts the number of occurrences of a term in each document and assigns them into a feature space. Sparck Jones [18] introduced inverse document frequency (IDF) in conjunction with term frequency to reduce the effect of common words in a document collection or corpus. The IDF measures the importance of a term in a document with a higher weight. Therefore, the combination of the TF-IDF method statistically measures the importance of a word in a corpus as follows in Eq. (1).

$$TF - IDF(t,d) = TF(t,d) * \log\left(\frac{N}{df(t)}\right) \tag{1}$$

where N refers to the number of documents and $df(t)$ represents the number of documents that contain the term t in a document collection. In particular, the TF-IDF solves the problem of common word occurrences in the document. But, it fails to capture the semantic meaning between words [22,50].

3.2.3. Word embeddings

To capture the semantic meaning between words, a word embedding mechanism is introduced in NLP. In this model, we tokenized and encoded the preprocessed data into an input sequence with a fixed length. We then generate the semantic meaning of each word with a fixed dimension using the pre-trained global vectors for word representation (GloVe). The GloVe method uses word-word co-occurrence statistics of a corpus to showcase the word vector space.

3.3. Multilabel learning algorithms

In traditional machine learning, the multilabel classification problem is solved using the problem transformation approach, adapted learning, and ensemble learning. First, the problem transformation approach converts the multilabel classification problem into single-label classification problems. Specifically, the problem transformation problems can be solved with three different techniques [38,14], namely, binary relevance (BR), classifiers chain (CC), and label powerset (LP). The BR method transforms the multilabel problem into single label problems and it learns using their class information. In the CC method, the multilabel problem is modified into a chain of the single-label classification problem. The modification extends the single label input space along the chain sequence. The LP method transforms the multilabel classification problem into a single multi-class classification problem. This method trains all unique label combinations in the input data. Second, the adapted learning approach directly solves the multilabel classification problem rather than dividing it into many subproblems. Third, ensemble learning combines multiple learning algorithms to provide better predictive results [23].

In deep learning, the multilabel classification problems can be solved directly with various techniques [26,37] such as CNNs and RNNs. These methods automatically find features in a context and perform very well. In this paper, we focus on the problem transformation approach, adapted algorithms, and RNNs for the task of multi-aspect category detection. Particularly, we use Naïve Bayes (NB), logistic regression (LR), and SVM in the problem transformation approach, multilabel k-nearest neighbor (MLkNN) in the adapted learning, and in deep learning, we use bidirectional RNNs such as bidirectional long short-term memory (BiLSTM) and BiGRU with attention mechanism in a single and multichannel setup. The multichannel extends multiple versions of the standard attention-based bidirectional models. Moreover, we describe the above-mentioned algorithms for the task of multi-aspect category detection.

3.3.1. Multinomial Naïve Bayes

NB is a simple probabilistic or generative model that applies Bayes' theorem with the strong assumptions of independence between features. The NB computes the probability of a given document (or instance) that belongs to a certain class [26]. Let X be a given document with a feature vector $(x_1, x_2, \ldots, x_n)$ and y be a class label. Then, the conditional probability of a document is computed based on Bayes's rule with the strong feature independence as in Eq. (2) and (3).

$$P(y/X) = \frac{P(y)P(X/y)}{P(X)} = \frac{P(y)\prod_{i=1}^{n} P(x_i/y)}{P(X)} \tag{2}$$

where the $P(X)$ is a constant value for a given document and it is classified using the following rule.

$$\hat{y} = \text{argmax}_x P(y)\prod_{i=1}^{n} P(x_i/y) \tag{3}$$

In the NB model, the posterior probability commonly estimates the parameters based on the frequency of document with class y including $P(y)$ and $P(x_i/y)$. In this paper, we use the multinomial Naïve Bayes (MNB) classification. This model performs well with fast and accurate results in various classification problems such as text classification, sentiment analysis, and spam filtering. The MNB uses frequency distribution data as word count vectors and TF-IDF vectors with conditional independence. Let $\theta_y = (\theta_{y1}, \theta_{y2}, \ldots, \theta_{yn})$ be the parameterized vectors for each class y, n is the number of text features, and θ_{yi} be the probability of feature i that belongs to class y. Let T be the training data, $N_{yi} = \sum_{x \in T} x_i$be the number of times feature i occurs in a document of class y, and $\sum_{i=1-n} N_{yi}$ is the total count for all features of class y. Then, the parameters θ_y is computed with a relative frequency counting (smoothed maximum likelihood estimation) as follows in Eq. (4).

$$\hat{\theta_{yi}} = \frac{N_{yi} + \alpha}{N_i + \alpha n} \tag{4}$$

where α refers to the smoothing priors, if $\alpha \geqslant 0$, then, the features are not present in the learning documents and it prevents zero probability values in further computations. Also, the smoothing priors can be set $\alpha = 1$ for Laplace smoothing and $\alpha < 1$ for Lidstone smoothing.

3.3.2. Logistic regression

Logistic regression (LR) is one of the simplest and earliest discriminative models. The LR is also known as logit regression and it models both linear and non-linear data with binary responses such as success (1) or failure (0). In particular, the LR model uses a logistic curve or logistic function to compute the probabilities of an event or a certain class as in Eq. (5). It is implemented with the distributional assumptions of the maximum-likelihood estimation of class in the observation.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

where $f(x)$ refers to the probability estimate values between 0 and 1, x denotes the input values to the function, and e denotes the natural logarithm. In this model, a decision boundary is used to map discrete values with a threshold point such as $p \geqslant 0.5$ for success and $p < 0$ for failure. In the case of multiple classes, the logistic regression selects the highest predicted probability value. Moreover, the LR model is used the Broyden–Fletcher–Goldfarb–Shanno optimization algorithm with L2 regularization to improve numerical stability.

3.3.3. Support vector machine

SVM is a supervised machine learning algorithm that analyzes data in both regression and classification problems. It is widely used in text classification, handwritten digit recognition, and object recognition. Let $x_1, x_2, \ldots, x_n$ be the training data in the vector space $X \subseteq R^d$. Let $y_1, y_2, \ldots, y_n$ be the given labels where $y_i \in \{1, -1\}$. The SVM method separates the training data points in a hyperplane (decision surface) with a maximal margin (Eq. 6). The data points belong to one side of the decision surface is labeled as 1 and the other side is labeled as -1. The data points closest to the decision surface are called support vectors in a vector space. Specifically, the SVM method uses a kernel operator to project the given training data into a higher dimensional representation.

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) \tag{6}$$

Here, K refers to the Mercer's condition that can be written as $K(a, b) = \Phi(a) \cdot \Phi(b)$. where Φ denotes the $X \to F$ and "$\cdot$" represents an inner product. The classifier can be written as in Eq. (7).

$$f(x) = w \cdot \Phi(x), \text{ where } w = \sum_{i=1}^{n} \alpha_i \Phi(x_i) \tag{7}$$

Moreover, the SVM computes the support vectors ($\alpha_i s$) concerning the decision surface in F. In this paper, we use the radial basis function (RBF) kernel to induce the training data with decision surface boundaries using Gaussian weights.

3.3.4. MLkNN

MLkNN is an extension of the kNN (k-nearest neighbor algorithm) to directly handle multilabel data. Let x be a document (or instance) and $Y \subseteq y$ be the associated label set. The category vector $\left(\vec{y}_x\right)$ takes the value of 1 if $l \in y$ for the $l - th$ component of document x and 0 for otherwise. Let $N(x)$ be the set of kNNs identified for a document x in the training data. Then, a membership counting vector (MCV) is defined based on the identified kNNs as follows in Eq. (8).

$$\overrightarrow{C_x}(l) = \sum_{a \in N(x)} \overrightarrow{y}_a(l),\ l \in y \tag{8}$$

where $\overrightarrow{C_x}(l)$ refers to the number of neighbors counted for a document x that belongs to the $l - th$ class. For each test document t, MLkNN identifies the set of kNNs in the training data. Let A_1^l be the event that has label l and A_0^l be the event that has no label l. Moreover, let $E_j^l(j \in \{0, 1, 2, \ldots, K\})$ be the event that indicates j instances exactly have label l. Then, the category vector $\left(\overrightarrow{y}_t\right)$ is calculated based on $\overrightarrow{C_t}$ using MAP (maximum a posteriori) as in Eq. (9). The prior probabilities and posterior probabilities are directly determined from the training data based on count frequency.

$$\overrightarrow{y}_t(l) = \text{argmax}_{b \in \{0,1\}} P\left(A_b^l | E^l_{\overrightarrow{C}_{t(l)}}\right),\ l \in y \tag{9}$$

The above equation is rewritten using the Bayesian rule as follows in Eq. (10) and (11).

$$\overrightarrow{y}_t(l) = \text{argmax}_{b \in \{0,1\}} \frac{P\left(A_b^l\right) P\left(E^l_{\overrightarrow{C}_{t(l)}} | A_b^l\right)}{P\left(E^l_{\overrightarrow{C}_{t(l)}}\right)} \tag{10}$$

$$\overrightarrow{y}_t(l) = \text{argmax}_{b \in \{0,1\}} P\left(A_b^l\right) P\left(E^l_{\overrightarrow{C}_{t(l)}} | A_b^l\right) \tag{11}$$

3.3.5. Multichannel bidirectional RNN with an attention mechanism

A RNN is defined from FNN (feed-forward neural networks) with internal memory to process input sequences. It is used in many tasks such as speech recognition and handwriting recognition. However, the RNN fails to solve the gradient vanishing and exploding problems and to capture very long input sequences. To overcome these issues, the variant of RNN architecture is introduced namely, LSTM network and GRU. In particular, we construct the multichannel RNN with attention mechanism for two input channels. Each input channel represents the same window size. These networks are explained as follows.

Long Short-Term Memory Network: An LSTM network explicitly handles long-term dependencies. Its default behavior is to remember information for long periods. The LSTM memory cell is implemented with three gates, namely, input gate (i_t), output gate (o_t), and forget gate (f_t). The input gate decides what new information to be stored in the memory cell. Next, the output gate decides what information should be moved to the next hidden state. The forget gate decides what information to be removed from the memory cell. Specifically, these gates update the current memory cell (c_t) and the current hidden state (h_t) as follows in Eqs. (12)–(16).

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{12}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{13}$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{14}$$

$$C_t = f_t \otimes C_{t-1} + i_t \otimes (\tanh(W_C \cdot [h_{t-1}, x_t] + b_C)) \tag{15}$$

$$h_t = o_t \otimes \tanh(C_t) \tag{16}$$

where σ refers to the logistic sigmoid function, where it returns all real number values in the range 0 to 1. tanh represents the hyperbolic tangent function that returns real values in the range -1 to 1. The LSTM network preserves only the past information unidirectionally and fails to preserve future information. Therefore, bidirectional RNN was introduced to preserve the past and future information for a specific time frame. In this work, the BiLSTM contains the forward LSTM and backward LSTM as follows in Eq. (17).

$$h_t = \left[\overrightarrow{h}_t, \overleftarrow{h}_t\right] \tag{17}$$

Gated Recurrent Units: The GRU model is a simpler version of the LSTM network. The GRU reduces the gating signals from three gates to two gates, namely, a reset gate (r_t) and an update gate (z_t) and. The reset gate adjusts the inclusion of new information with the previous activation, and the update gate controls the previous activation for preserving the information. Mathematically, the hidden state (h_t) of GRU is computed as follows in Eqs. (18)–(20).

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \tag{18}$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{19}$$

$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes \tanh(W_C \cdot [x_t, r_t \otimes h_{t-1}] + b_c) \tag{20}$$

where σ refers to the logistic sigmoid function, W^* denotes the weights applied to the input sequence and recurrent hidden state. Specifically, there is no separate memory block in the GRU like LSTM. The update gate combines the input gate and the forget gate to balance the previous activation and update activation. The forward process and backward process concatenated for bidirectional GRU as follows in Eq. (21).

$$h_t = \left[\overrightarrow{h}_t, \overleftarrow{h}_t\right] \tag{21}$$

Attention Mechanism (AM): In recent years, the attention mechanism has shown breakthrough results in machine translation, NLP, speech recognition, image caption, and question answering [49]. Each word of a document or sentence contributes to unequal representation in various NLP tasks such as sentiment analysis, text classification, document classification, POS tagging, and relation classification. In particular, the attention mechanism pays selective weights to the important words of the document. The attention weight for each word is computed in the context (c_t) as follows in Eqs. (22)–(25).

$$M_{t,t'} = \tanh\left(W_t \cdot \left[x_t^T, x_{t'}^T\right] + b_t\right) \tag{22}$$

$$e_{t,t'} = \sigma\left(W_a M_{t,t'} + b_a\right) \tag{23}$$

$$a_t = softmax(e_t) \tag{24}$$

$$c_t = \sum_{t'} a_{t,t'} x_{t'} \tag{25}$$

Output Layer: We use an output layer with three neurons, a sigmoid activation function [39], and a binary cross-entropy loss. Each neuron represents one of the three aspects. The sigmoid function converts each of these aspects scores between 0 and 1 independently. Moreover, the computation of loss function for a certain class is not influenced by other classes.

4. Multilabel evaluation metrics

In traditional classification problems, we use accuracy, precision, recall, and F1-measure to evaluate single label multiclass problems. In multilabel learning, we use different measures, namely, sample-based measures, label-based measures, and rank-based measures for evaluating the performance of the model [1,14,38]. Each of these measures is described as follows. Let $T = \left(x_{(i)}, Y_{(i)}\right), i = 1, 2, \ldots, n.$ $x_{(i)} \in X, Y_{(i)} \in y = \{0, 1\}^l$, be the set of training instances with a set of actual labels $(L), |L| = l$. Let H be a multilabel learning algorithm and $Z_i = H\left(x_{(i)}\right)$ be the set of predicted labels.

4.1. Sample-based measures

Sample-based measures calculate the average difference between the actual labels and the predicted labels for each test sample and then averaged over all samples in the test set. Specifically, we use exam match ratio (EM), zero-one loss (ZOL), Hamming Loss (HL), accuracy (A), precision (P), recall (R), and F1-measure (F1) to predict the performance of the model based on samples. The EMR (also called subset accuracy) identifies the average of correctly classified labels over all samples (Eq. 26). Also, it ignores partially identified correct labels. HL computes the average of misclassified labels over all samples (Eq. 27). The zero value of HL indicates the best. Accuracy measures the average proportion between predicted correct labels and both predicted and actual labels (Eq. 28). Precision and recall calculates the average proportion between predicted correct labels and the actual labels and predicted labels, respectively (Eqs. (29)–(31).

$$ExactMatchRatio(EM) = \frac{1}{S}\sum_{i=1}^{S} I\left(Y_{(i)} = Z_{(i)}\right) \tag{26}$$

$$Zero - OneLoss(ZOL) = \frac{1}{S}\sum_{i=1}^{S} I\left(Y_{(i)} \neq Z_{(i)}\right) \tag{27}$$

$$Hamming - loss(HL) = \frac{1}{S}\sum_{i=1}^{S} \frac{Y_{(i)} \Delta Z_{(i)}}{L} \tag{28}$$

$$Accuracy(A) = \frac{1}{S}\sum_{i=1}^{S} \frac{Y_{(i)} \cap Z_{(i)}}{Y_{(i)} \cup Z_{(i)}} \tag{29}$$

$$Precision(P) = \frac{1}{S}\sum_{i=1}^{S} \frac{Y_{(i)} \cap Z_{(i)}}{Z_{(i)}} \tag{30}$$

$$Recall(R) = \frac{1}{S}\sum_{i=1}^{S} \frac{Y_{(i)} \cap Z_{(i)}}{Y_{(i)}} \tag{31}$$

$$F1 - measure(F1) = \frac{1}{S}\sum_{i=1}^{S} \frac{Y_{(i)} \cap Z_{(i)}}{|Y_{(i)}| + |Z_{(i)}|} \tag{32}$$

where I(false)=0 and I(true)=1 in Eq. (26), I(false)=1 and I(true)=0 in Eq. (27), and Δ defines the symmetric difference between actual and predicted label sets.

4.2. Label-based measures

Label-based measures compute precision, recall, F1-measure, and their macro, micro, weighted average, and AUC scores based on the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) as in Eqs. (33)–(44).

$$Precision(P) = \frac{TP}{TP + FP} \tag{33}$$

$$Recall(R) = \frac{TP}{TP + FN} \tag{34}$$

$$F1 - measure(F1) = 2 \times \frac{P \times R}{P + R} \tag{35}$$

$$Macro_{Precision} = \frac{1}{L}\sum_{i=1}^{L}\text{Precisionofi} \tag{36}$$

$$Macro_{Recall} = \frac{1}{L}\sum_{i=1}^{L}\text{Recallofi} \tag{37}$$

$$Macro_{F1-measure} = \frac{1}{L}\sum_{i=1}^{L}2 \times \frac{Macro_{Precision} \times Macro_{Recall}}{Macro_{Precision} + Macro_{Recall}} \tag{38}$$

$$Micro_{Precision} = \frac{\sum_{i=1}^{L}\text{TPofi}}{\sum_{i=1}^{L}(\text{TPofi} + \text{FPofi})} \tag{39}$$

$$Micro_{Recall} = \frac{\sum_{i=1}^{L}\text{TPofi}}{\sum_{i=1}^{L}(\text{TPofi} + \text{FNofi})} \tag{40}$$

$$Micro_{F1-measure} = 2 \times \frac{Micro_{Precision} \times Micro_{Recall}}{Micro_{Precision} + Micro_{Recall}} \tag{41}$$

$$Weighted_{Precision} = \sum_{i=1}^{L}\text{Precisionofi} \times \text{Weightofi} \tag{42}$$

$$Weighted_{Recall} = \sum_{i=1}^{L}\text{Recallofi} \times \text{Weightofi} \tag{43}$$

$$Weighted_{F1-measure} = \sum_{i=1}^{L}\text{F1} - \text{measureofi} \times \text{Weightofi} \tag{44}$$

4.3. Rank-based measures

In multilabel learning, the rank-based measures compute the rank of the predicted labels associated with each sample. It aims to give better ranks with higher scores for the predicted labels. Specifically, we use the ranking loss and label average precision metrics to evaluate model performance.

4.3.1. Ranking loss

Ranking loss (RL) computes the average number of times an incorrect label is ranked against the correct labels (Eq. 45). A value of zero indicates the best-ranking performance.

$$RL(Y, Z) = \frac{1}{S}\sum_{i=1}^{S}\frac{1}{||Y_{(i)}||_0||\overline{Y}_{(i)}||_0}|\{(a, b) : Z_{(ia)} \leqslant Z_{(ib)}, Y_{(ia)} = 1, \overline{Y}_{(ib)} = 0\}| \tag{45}$$

where $\overline{Y}_{(i)}$ refers to the complementary set of $Y_{(i)}$ with respect to the set of all labels, $|\cdot|$ calculates the cardinality of the set, and $||\cdot||_0$ calculates the nonzero elements.

Table 1
Aspects and sentiment polarities.

Dataset	Aspects	Male		Female		Male and Female	
		Positive	Negative	Positive	Negative	Positive	Negative
Abilify oral	Satisfaction	234	163	680	538	914	701
	Ease of use	336	61	1036	182	1372	243
	Effectiveness	259	138	833	385	1092	523
Celebrex oral	Satisfaction	395	138	1510	395	1905	533
	Ease of use	475	58	1737	168	2212	226
	Effectiveness	350	183	1313	592	1663	775

4.3.2. Average precision

Ranking average precision (RAP) calculates the average fraction of correct labels that are ranked on all samples as in Eq. 46. The larger value indicates better performance.

$$RAP(Y,Z) = \frac{1}{S}\sum_{i=1}^{S}\frac{1}{||Y_{(i)}||_0}\sum_{j:Y_{(ij)}=1}\frac{|L_{(ij)}|}{rank_{(ij)}} \tag{46}$$

where $L_{(ij)} = \{a : Y_{(ia)} = 1, Z_{(ia)} \geqslant Z_{(ij)}\}, rank_{(ij)} = |\{a : Z_{(ia)} \geqslant Z_{(ij)}\}|, |\cdot|$ calculates the cardinality of the set, and $||\cdot||_0$ calculates the nonzero elements.

Table 2
Model performances for the Abilify oral male group with 5-fold.

Models	Features	EM	HL	ZOL	AUC	RL	AP	Macro	Micro	WA	S	F1-score per label		
								F1	F1	F1	F1	0	1	2
LSTM_F	GloVe	41.82	0.342	0.582	47.48	0.057	96.41	76.31	77.72	77.97	68.9	72.39	90.15	66.38
BiLSTM	GloVe	32.75	0.401	0.672	48.87	0.106	93.97	67.30	70.20	68.77	60.58	59.90	78.75	63.26
GRU_F	GloVe	38.09	0.356	0.619	50.93	0.078	94.88	73.85	75.32	75.32	65.02	73.96	84.83	62.77
BiGRU	GloVe	29.20	0.403	0.708	50.16	0.113	92.53	66.99	70.53	69.10	59.74	64.80	83.57	52.60
MCLSTM_F	GloVe	43.36	0.335	0.566	51.63	0.083	95.11	76.71	77.96	78.16	68.66	71.73	88.85	69.54
MCBiLSTM	GloVe	40.79	0.347	0.592	**53.92**	0.135	90.85	75.80	77.06	77.10	65.94	72.41	86.61	68.39
MCGRU_F	GloVe	37.25	0.341	0.628	48.57	**0.044**	**97.76**	73.63	76.80	75.80	68.66	73.27	90.61	57.00
MCBiGRU	GloVe	38.51	0.369	0.615	50.02	0.112	93.62	73.60	75.19	75.13	66.29	66.25	87.57	66.97
BR_NB	GloVe	**56.22**	0.303	**0.438**	50.13	0.302	83.23	81.51	82.02	82.78	74.33	78.83	**91.66**	74.06
BR_LR	GloVe	46.60	0.314	0.534	51.15	0.260	84.65	79.39	80.54	80.87	72.56	77.69	**91.66**	68.83
BR_SVM	GloVe	49.63	0.306	0.504	51.57	0.265	84.59	80.20	81.21	81.57	73.32	**78.85**	**91.66**	70.09
CC_NB	GloVe	**56.22**	0.304	**0.438**	50.00	0.302	83.23	81.47	81.98	82.74	74.33	78.83	**91.66**	73.94
CC_LR	GloVe	55.71	0.304	0.443	50.72	0.294	83.56	81.21	81.77	82.48	73.70	77.69	91.64	**74.31**
CC_SVM	GloVe	**56.22**	0.302	**0.438**	50.56	0.297	83.56	81.48	82.01	82.74	74.33	**78.85**	**91.66**	73.94
LP_NB	GloVe	**56.22**	0.304	**0.438**	50.00	0.302	83.23	81.47	81.98	82.74	74.33	78.83	**91.66**	73.94
LP_LR	GloVe	55.71	0.307	0.443	50.16	0.301	83.27	81.15	81.68	82.44	73.80	78.22	91.64	73.58
LP_SVM	GloVe	**56.22**	0.304	**0.438**	50.00	0.302	83.23	81.47	81.98	82.74	74.33	78.83	**91.66**	73.94
AA_MLkNN	GloVe	41.01	0.369	0.589	50.11	0.240	86.60	72.93	74.80	74.84	64.99	69.18	89.05	60.54
BR_NB	BoW	46.39	0.331	0.536	51.54	0.254	85.50	77.37	78.53	78.89	69.05	73.87	90.26	67.99
BR_LR	BoW	45.34	0.332	0.547	51.73	0.229	86.85	76.81	78.15	78.32	69.19	73.43	89.71	67.29
BR_SVM	BoW	52.67	**0.299**	0.473	51.58	0.272	84.21	81.10	81.84	82.39	73.97	78.83	**91.66**	72.81
CC_NB	BoW	49.66	0.325	0.503	51.87	0.267	84.99	78.27	79.15	79.61	69.33	73.87	89.94	71.00
CC_LR	BoW	49.14	0.336	0.509	51.18	0.254	86.05	77.11	78.15	78.50	68.19	73.43	88.77	69.13
CC_SVM	BoW	53.43	0.306	0.467	50.27	0.292	83.56	81.03	81.66	82.34	73.95	78.83	**91.66**	72.59
LP_NB	BoW	55.21	0.304	0.448	51.62	0.288	84.02	80.87	81.50	82.14	72.84	77.57	91.32	73.73
LP_LR	BoW	50.91	0.324	0.491	52.33	0.262	85.58	78.18	79.08	79.46	68.71	74.09	89.03	71.42
LP_SVM	BoW	**56.22**	0.304	**0.438**	50.00	0.302	83.23	81.47	81.98	82.74	74.33	78.83	**91.66**	73.94
AA_MLkNN	BoW	36.53	0.373	0.635	49.41	0.245	85.21	72.83	74.76	74.47	64.65	73.55	84.98	59.95
BR_NB	TF-IDF	**56.22**	0.304	**0.438**	50.00	0.302	83.23	81.47	81.98	82.74	74.33	78.83	**91.66**	73.94
BR_LR	TF-IDF	55.96	0.302	0.440	50.38	0.298	83.35	81.52	82.04	82.78	74.33	78.83	**91.66**	74.06
BR_SVM	TF-IDF	**56.22**	0.302	**0.438**	50.25	0.301	83.27	**81.56**	**82.06**	**82.81**	**74.38**	78.83	91.66	74.18
CC_NB	TF-IDF	**56.22**	0.304	**0.438**	50.00	0.302	83.23	81.47	81.98	82.74	74.33	78.83	**91.66**	73.94
CC_LR	TF-IDF	**56.22**	0.304	**0.438**	50.00	0.302	83.23	81.47	81.98	82.74	74.33	78.83	**91.66**	73.94
CC_SVM	TF-IDF	**56.22**	0.304	**0.438**	50.00	0.302	83.23	81.47	81.98	82.74	74.33	78.83	**91.66**	73.94
LP_NB	TF-IDF	**56.22**	0.304	**0.438**	50.00	0.302	83.23	81.47	81.98	82.74	74.33	78.83	**91.66**	73.94
LP_LR	TF-IDF	**56.22**	0.304	**0.438**	50.00	0.302	83.23	81.47	81.98	82.74	74.33	78.83	**91.66**	73.94
LP_SVM	TF-IDF	**56.22**	0.304	**0.438**	50.00	0.302	83.23	81.47	81.98	82.74	74.33	78.83	**91.66**	73.94
AA_MLkNN	TF-IDF	34.53	0.407	0.654	52.79	0.268	84.84	68.70	69.74	69.75	54.55	70.26	76.46	59.39

Table 3
Model performances for the Abilify oral female group with 5-fold.

Models	Features	EM	HL	ZOL	AUC	RL	AP	Macro	Micro	WA	S	F1-score per label		
								F1	F1	F1	F1	0	1	2
LSTM_F	GloVe	43.44	0.322	0.566	57.33	0.111	92.65	76.19	78.14	78.00	68.24	74.65	89.60	64.32
BiLSTM	GloVe	40.47	0.334	0.595	**58.71**	0.083	94.95	75.43	76.94	77.01	65.59	73.94	87.54	64.82
GRU_F	GloVe	44.99	0.331	0.550	55.52	0.121	92.42	76.56	77.77	78.10	67.10	76.04	87.71	65.93
BiGRU	GloVe	43.18	0.337	0.568	55.40	0.083	95.16	75.20	77.11	77.16	67.19	73.37	89.72	62.5
MCLSTM_F	GloVe	44.83	0.321	0.552	58.44	0.084	94.77	76.67	78.42	78.43	68.19	75.45	89.64	64.93
MCBiLSTM	GloVe	33.74	0.359	0.663	56.13	**0.055**	**96.77**	69.53	73.74	72.17	63.66	71.07	87.33	50.19
MCGRU_F	GloVe	39.89	0.353	0.601	57.08	0.099	93.73	73.47	75.44	75.31	65.02	73.34	86.32	60.75
MCBiGRU	GloVe	35.21	0.357	0.648	56.96	0.102	93.22	71.85	74.19	73.50	63.47	70.31	84.26	60.98
BR_NB	GloVe	**53.53**	0.302	**0.465**	50.00	0.341	82.16	81.54	82.15	83.02	**74.95**	81.17	**91.90**	71.56
BR_LR	GloVe	45.97	0.304	0.540	50.89	0.283	84.07	79.84	81.18	81.57	73.92	80.66	**91.90**	66.95
BR_SVM	GloVe	46.46	0.307	0.535	50.51	0.298	83.55	79.96	81.22	81.70	74.03	81.17	**91.90**	66.79
CC_NB	GloVe	**53.53**	0.302	**0.465**	50.00	0.341	82.16	81.54	82.15	83.02	**74.95**	81.17	**91.90**	71.56
CC_LR	GloVe	52.87	0.306	0.471	49.75	0.340	82.21	81.21	81.85	82.72	74.58	80.66	**91.90**	71.07
CC_SVM	GloVe	**53.53**	0.302	**0.465**	50.00	0.341	82.16	81.54	82.15	83.02	**74.95**	81.17	**91.90**	71.56
LP_NB	GloVe	**53.53**	0.302	**0.465**	50.00	0.341	82.16	81.54	82.15	83.02	**74.95**	81.17	**91.90**	71.56
LP_LR	GloVe	52.79	0.306	0.472	49.79	0.339	82.25	81.19	81.84	82.69	74.55	80.55	**91.90**	71.13
LP_SVM	GloVe	**53.53**	0.302	**0.465**	50.00	0.341	82.16	81.54	82.15	83.02	**74.95**	81.17	**91.90**	71.56
AA_MLkNN	GloVe	40.31	0.336	0.597	51.40	0.258	85.59	75.47	77.47	77.49	68.36	74.47	90.13	61.81
BR_NB	BoW	45.73	0.302	0.543	53.48	0.268	85.41	79.20	80.66	80.94	72.47	79.13	91.52	66.95
BR_LR	BoW	42.69	0.320	0.573	53.36	0.242	86.51	76.89	78.77	78.86	70.35	76.73	90.89	63.04
BR_SVM	BoW	47.62	**0.299**	0.524	51.17	0.284	84.02	80.39	81.63	82.06	74.37	81.16	**91.90**	68.11
CC_NB	BoW	48.19	0.301	0.518	53.23	0.280	85.02	79.61	80.86	81.26	72.47	79.13	91.59	68.10
CC_LR	BoW	47.45	0.320	0.525	53.82	0.278	85.32	77.80	79.09	79.49	69.31	76.73	90.27	66.40
CC_SVM	BoW	52.95	0.300	0.470	50.45	0.328	82.57	81.53	**82.20**	82.98	74.92	81.16	**91.90**	71.52
LP_NB	BoW	52.79	0.300	0.472	51.29	0.323	83.07	81.24	81.97	82.72	74.42	80.52	91.98	71.22
LP_LR	BoW	48.27	0.318	0.517	53.02	0.282	85.00	78.22	79.39	79.80	69.61	77.36	89.81	67.49
LP_SVM	BoW	**53.53**	0.302	**0.465**	50.00	0.341	82.16	**81.59**	82.18	**83.03**	**74.95**	**81.22**	**91.90**	**71.64**
AA_MLkNN	BoW	40.56	0.320	0.594	51.60	0.245	86.00	76.54	79.00	78.65	71.18	77.76	91.29	60.56
BR_NB	TF-IDF	**53.53**	0.302	**0.465**	50.00	0.341	82.16	**81.59**	82.18	**83.03**	**74.95**	**81.22**	**91.90**	**71.64**
BR_LR	TF-IDF	50.99	**0.299**	0.490	50.65	0.321	82.80	81.30	82.09	82.79	74.78	**81.22**	**91.90**	70.76
BR_SVM	TF-IDF	52.79	0.302	0.472	50.11	0.335	82.35	81.45	82.10	82.92	74.86	**81.22**	**91.90**	71.22
CC_NB	TF-IDF	**53.53**	0.302	**0.465**	50.00	0.341	82.16	**81.59**	82.18	**83.03**	**74.9**5	**81.22**	**91.90**	**71.64**
CC_LR	TF-IDF	**53.53**	0.302	**0.465**	50.00	0.341	82.16	**81.59**	82.18	**83.03**	**74.95**	**81.22**	**91.90**	**71.64**
CC_SVM	TF-IDF	**53.53**	0.302	**0.465**	50.00	0.341	82.16	**81.59**	82.18	**83.03**	**74.95**	**81.22**	**91.90**	**71.64**
LP_NB	TF-IDF	**53.53**	0.302	**0.465**	50.00	0.341	82.16	**81.59**	82.18	**83.03**	**74.95**	**81.22**	**91.90**	**71.64**
LP_LR	TF-IDF	**53.53**	0.302	**0.465**	50.00	0.341	82.16	**81.59**	82.18	**83.03**	**74.95**	**81.22**	**91.90**	**71.64**
LP_SVM	TF-IDF	**53.53**	0.302	**0.465**	50.00	0.341	82.16	81.59	82.18	83.03	74.95	81.22	91.90	71.64
AA_MLkNN	TF-IDF	35.70	0.332	0.643	51.98	0.225	86.30	74.57	77.42	76.85	69.15	74.93	90.93	57.84

5. Results and discussion

We implement the proposed gender-based multi-aspect sentiment detection task on Abilify Oral dataset and Celebrex Oral dataset. The Abilify dataset includes 1615 gender-based instances (397 male and 1218 female) with three predefined multi-aspects, namely, satisfaction, ease of use, and effectiveness. Similarly, the Celebrex dataset includes 2438 instances (533 male and 1905 female) with three multi-aspects. The sentiment polarity on these aspects is shown in Table 1. Specifically, we used Anaconda software with python libraries [34] on a Windows 10 machine with 8 GB memory and an i7 processor at 3.60 GHz. We experiment on three groups separately such as male, female, and both male and female. To obtain the quality of data, we apply various preprocessing techniques such as upper case to lower case conversion, removing stopwords and punctuations, and tokenization. Next, the features are represented with the BoWs, TF-IDF, and GloVe models. For the multilabel classification task, we employ both traditional machine learning and deep learning algorithms. In particular, we used the MNB, LR, SVM, unidirectional and bidirectional RNNs, and multichannel unidirectional and bidirectional RNNs with attention mechanism on the above feature representations. The traditional machine learning employs the problem transformation approaches (BR, CC, and LP) and adapted algorithms (MLkNN) for the multi-aspect sentiment detection task. We set the deep learning model parameters based on random searches such as 400 input sequence length, 25 epochs, 10000 maximum words, 100 GloVe embedding dimension, 0.25 dropout rate, 32 unit for dense layers, sigmoid activation for both attention and output function, 32 Units in both LSTM and GRU, binary loss cross-entropy, Adam optimizer, 0.0001 learning rate, and single and double input channels. Moreover, we used the Laplace smoothing value with 1 for learning class prior probabilities for the NB model. In the logistic regression model, we used the inverse regularization strength and intercept scaling value with 1, 0.0001 tolerance stopping criteria, and Broyden–Fletcher–Goldfarb–Shanno optimization algorithm with L2 penalty. The SVM method applied with regularization parameter value as 1 and radial basis kernel. In the MLkNN

Table 4
Model performances for the Abilify oral male and female group with 5-fold.

Models	Features	EM	HL	ZOL	AUC	RL	AP	Macro	Micro	WA	S	F1-score per label		
								F1	F1	F1	F1	0	1	2
LSTM_F	GloVe	37.65	0.360	0.624	56.82	0.098	94.14	72.71	74.47	74.14	63.29	66.83	85.52	65.79
BiLSTM	GloVe	34.24	0.365	0.658	57.92	0.101	93.32	67.73	71.57	69.80	61.64	73.93	80.19	49.07
GRU_F	GloVe	42.72	0.332	0.573	57.08	0.096	94.19	73.52	76.92	75.64	68.56	69.68	90.57	60.33
BiGRU	GloVe	40.00	0.357	0.600	57.39	0.098	94.29	73.39	75.01	74.67	64.75	66.67	85.34	68.17
MCLSTM_F	GloVe	35.23	0.371	0.648	57.56	0.080	94.89	67.08	71.29	69.38	60.95	66.95	83.96	50.32
MCBiLSTM	GloVe	43.41	0.336	0.566	55.42	0.093	94.15	73.70	76.98	75.64	67.44	76.31	87.12	57.66
MCGRU_F	GloVe	32.69	0.355	0.673	**57.99**	**0.046**	**97.41**	68.21	73.11	70.93	63.05	66.16	88.96	49.52
MCBiGRU	GloVe	51.02	0.304	0.490	55.40	0.088	94.88	80.39	81.21	81.76	72.90	79.09	90.95	71.11
BR_NB	GloVe	54.12	0.303	0.458	50.00	0.331	82.41	81.56	82.13	82.95	74.79	80.64	91.85	72.19
BR_LR	GloVe	47.37	0.304	0.526	50.99	0.280	84.09	80.15	81.34	81.75	73.97	80.45	91.85	68.14
BR_SVM	GloVe	48.30	0.308	0.517	50.34	0.298	83.48	80.26	81.30	81.86	73.98	80.64	91.85	68.29
CC_NB	GloVe	**54.18**	0.303	**0.458**	50.00	0.331	82.41	81.56	82.13	82.95	74.79	80.64	91.85	72.19
CC_LR	GloVe	53.99	0.304	0.460	49.94	0.330	82.47	81.45	82.03	82.85	74.69	80.45	91.85	72.05
CC_SVM	GloVe	**54.18**	0.303	**0.458**	50.00	0.331	82.41	81.56	82.13	82.95	74.79	80.64	91.85	72.19
LP_NB	GloVe	**54.18**	0.303	**0.458**	50.00	0.331	82.41	81.56	82.13	82.95	74.79	80.64	91.85	72.19
LP_LR	GloVe	53.93	0.304	0.461	49.98	0.330	82.48	81.45	82.03	82.84	74.67	80.44	91.85	72.06
LP_SVM	GloVe	**54.18**	0.303	**0.458**	50.00	0.331	82.41	81.56	82.13	82.95	74.79	80.64	91.85	72.19
AA_MLkNN	GloVe	39.57	0.343	0.604	50.28	0.257	85.59	75.24	77.25	77.27	68.46	74.19	90.46	61.05
BR_NB	BoW	47.68	0.301	0.523	53.51	0.264	85.42	79.56	80.79	81.10	72.21	78.60	91.18	68.89
BR_LR	BoW	42.04	0.323	0.560	53.24	0.243	86.28	76.78	78.50	78.57	69.46	75.58	90.39	64.36
BR_SVM	BoW	48.11	**0.294**	0.519	51.94	0.267	84.55	80.66	81.85	82.19	74.41	80.65	**91.86**	69.48
CC_NB	BoW	49.66	0.303	0.503	53.08	0.277	84.96	79.78	80.82	81.25	72.11	78.60	90.99	69.73
CC_LR	BoW	46.07	0.333	0.539	52.68	0.280	85.12	76.86	78.09	78.46	67.66	75.58	89.10	65.91
CC_SVM	BoW	53.62	0.299	0.464	50.62	0.318	82.89	**81.61**	**82.25**	**82.96**	**74.85**	80.65	**91.86**	**72.33**
LP_NB	BoW	53.44	0.304	0.466	50.89	0.319	83.12	81.05	81.71	82.42	73.85	79.63	91.69	71.82
LP_LR	BoW	48.24	0.321	0.518	53.46	0.276	85.23	77.84	78.97	79.34	68.57	76.28	89.41	67.83
LP_SVM	BoW	**54.18**	0.303	**0.458**	50.00	0.331	82.41	81.59	82.15	82.95	74.79	**80.66**	**91.86**	72.26
AA_MLkNN	BoW	46.69	0.311	0.533	52.43	0.265	85.38	78.82	80.10	80.42	71.71	77.22	91.21	68.02
BR_NB	TF-IDF	**54.18**	0.303	**0.458**	50.00	0.331	82.41	81.59	82.15	82.95	74.79	**80.66**	**91.86**	72.26
BR_LR	TF-IDF	52.14	0.298	0.479	50.85	0.310	83.11	81.40	82.14	82.79	74.67	**80.66**	**91.86**	71.68
BR_SVM	TF-IDF	53.50	0.302	0.465	50.24	0.326	82.58	81.53	82.14	82.89	74.70	**80.66**	**91.86**	72.06
CC_NB	TF-IDF	**54.18**	0.303	**0.458**	50.00	0.331	82.41	81.59	82.15	82.95	74.79	**80.66**	**91.86**	72.26
CC_LR	TF-IDF	**54.18**	0.303	**0.458**	50.00	0.331	82.41	81.59	82.15	82.95	74.79	**80.66**	**91.86**	72.26
CC_SVM	TF-IDF	**54.18**	0.303	**0.458**	50.00	0.331	82.41	81.59	82.15	82.95	74.79	**80.66**	**91.86**	72.26
LP_NB	TF-IDF	**54.18**	0.303	**0.458**	50.00	0.331	82.41	81.59	82.15	82.95	74.79	**80.66**	**91.86**	72.26
LP_LR	TF-IDF	**54.18**	0.303	**0.458**	50.00	0.331	82.41	81.59	82.15	82.95	74.79	**80.66**	**91.86**	72.26
LP_SVM	TF-IDF	**54.18**	0.303	**0.458**	50.00	0.331	82.41	81.59	82.15	82.95	74.79	**80.66**	**91.86**	72.26
AA_MLkNN	TF-IDF	33.19	0.342	0.668	51.82	0.222	86.52	73.89	76.52	76.02	67.54	75.45	88.85	57.36

method, we randomly selected k-nearest neighbors as 3 using the Minkowski distance and smoothing parameter values as 1.

We conducted 5-fold cross-validation on the above models with three feature representations for each group on both datasets. These models are evaluated using multilabel evaluation metrics in terms of samples, labels, and ranks. Specifically, we used exact match (EM), Hamming-Loss (HL), Zero-one Loss (ZOL), and F1-measures in sample-based metrics, AUC score, Macro-F1, Micro-F1, Weighted-F1, and F1-measures per label in label-based metrics, and the ranking loss (RL) and average precision (AP) in rank-based metrics. Tables 2–4 show the gender-based (male, female, and both) evaluation of machine learning and deep learning models for the Abilify dataset. In the sample-based metrics, the EM (56.22%, 53.53%, and 54.18%) and ZOL (0.438, 0.465, and 0.458) achieves the better results on each group with CC_NB + GloVe, CC_SVM + GloVe, LP_NB + GloVe, LP_SVM + GloVe, LP_SVM + BoW, BR_NB + TF-IDF, CC_NB + TF-IDF, and CC_LR + TF-IDF, and CC_SVM + TF-IDF, LP_NB + TF-IDF, LP_LR + TF-IDF, and LP_SVM + TF-IDF models. The HL (0.299, 0.299, and 0.294) achieves better results on each group with the BR_SVM + BoW model. Moreover, the sample F1-measure achieves 74.38%, 74.95%, and 74.85% on each group. In the label-based metrics, the macro F1, micro F1, and weighted average F1 achieves 81.56%, 82.06%, and 82.81% with BR_SVM + BoW for male group, 81.59%, 82.20%, and 83.03% with CC_SVM + BoW, LP_SVM + BoW, BR_NB + TF-IDF, CC_NB + TF-IDF, CC_LR + TF-IDF, CC_SVM + TF-IDF, LP_NB + TF-IDF, LP_LR + TF-IDF, & LP_SVM + TF-IDF for female group, and 81.61%, 82.25%, and 82.96% with CC_SVM + BoW for both male and female groups. The AUC score achieves 53.92% with MCBiLSTM + GloVe, 58.71% with BiLSTM + GloVe, and 57.99% with MCGRU_F + GloVe for each group, respectively. Particularly, the traditional multilabel learning models achieve a better F1-score per label on all groups. In the ranked-based measures, the RL and AP achieve 0.044 and 97.76% scores with MCGRU_F + GloVe for the male group, and 0.055 and 96.77% scores with MCBiLSTM + GloVe for the female group, and 0.046 and 97.4% scores with MCGRU_F + GloVe for both male and female.

Table 5
Model performances for the Celebrex oral male group with 5-fold.

Models	Features	EM	HL	ZOL	AUC	RL	AP	Macro	Micro	WA	S	F1-score per label		
								F1	F1	F1	F1	0	1	2
LSTM_F	GloVe	54.07	0.281	0.459	61.73	0.107	93.28	81.21	81.90	81.99	72.11	79.53	89.34	74.74
BiLSTM	GloVe	54.59	0.268	0.454	65.46	0.102	93.28	81.76	82.82	82.66	73.13	79.57	90.84	74.87
GRU_F	GloVe	42.00	0.343	0.580	64.56	0.077	95.14	72.07	74.93	73.38	63.58	68.83	86.28	61.13
BiGRU	GloVe	44.62	0.276	0.554	66.05	0.090	94.49	78.08	81.81	79.41	72.50	79.82	91.42	62.99
MCLSTM_F	GloVe	45.57	0.294	0.544	66.70	0.105	93.16	79.19	80.07	79.84	68.54	79.16	85.21	73.19
MCBiLSTM	GloVe	52.15	0.270	0.478	66.73	0.096	93.71	81.55	82.27	82.24	70.52	80.23	89.02	75.40
MCGRU_F	GloVe	51.42	0.269	0.486	68.40	0.093	94.05	81.12	82.49	82.17	73.33	79.31	92.11	71.95
MCBiGRU	GloVe	48.97	0.291	0.510	**70.91**	**0.068**	**95.40**	79.31	80.07	79.70	66.92	78.66	83.57	75.69
BR_NB	GloVe	64.55	0.237	0.354	50.00	0.264	85.31	86.19	86.54	87.00	80.42	85.10	**94.24**	79.24
BR_LR	GloVe	61.17	0.240	0.388	50.91	0.243	86.07	85.59	86.11	86.48	79.92	84.89	**94.24**	77.65
BR_SVM	GloVe	63.24	0.239	0.368	50.22	0.258	85.51	85.94	86.35	86.78	80.20	85.10	**94.24**	78.48
CC_NB	GloVe	64.55	0.237	0.354	50.00	0.264	85.31	86.19	86.54	87.00	**80.42**	85.10	**94.24**	79.24
CC_LR	GloVe	64.37	0.237	0.356	50.26	0.262	85.46	86.12	86.49	86.93	80.30	84.89	**94.24**	79.22
CC_SVM	GloVe	64.55	0.237	0.354	50.00	0.264	85.31	86.19	86.54	87.00	**80.42**	85.10	**94.24**	79.24
LP_NB	GloVe	64.55	0.237	0.354	50.00	0.264	85.31	86.19	86.54	87.00	**80.42**	85.10	**94.24**	79.24
LP_LR	GloVe	64.36	0.239	0.356	49.98	0.263	85.37	86.04	86.40	86.85	80.18	84.71	94.12	79.27
LP_SVM	GloVe	64.55	0.237	0.354	50.00	0.264	85.31	86.19	86.54	87.00	**80.42**	85.10	**94.24**	79.24
AA_MLkNN	GloVe	52.17	0.237	0.478	51.77	0.224	87.01	82.19	83.17	83.25	75.31	80.75	92.94	72.88
BR_NB	BoW	59.48	0.228	0.405	55.86	0.211	87.76	85.69	86.29	86.49	79.02	84.32	93.80	78.95
BR_LR	BoW	60.61	0.206	0.394	59.86	0.181	89.34	86.70	87.32	87.37	79.58	85.29	93.71	81.10
BR_SVM	BoW	64.36	0.235	0.356	50.40	0.262	85.39	86.25	86.61	87.06	**80.42**	85.10	**94.24**	79.42
CC_NB	BoW	61.54	0.226	0.385	55.63	0.217	87.58	85.97	86.49	86.73	79.24	84.32	93.80	79.78
CC_LR	BoW	65.49	0.209	0.345	59.43	0.205	88.78	86.87	87.39	87.57	79.75	85.29	94.09	81.22
CC_SVM	BoW	64.55	0.237	0.354	50.00	0.264	85.31	86.19	86.54	87.00	**80.42**	85.10	**94.24**	79.24
LP_NB	BoW	63.61	0.242	0.364	51.28	0.242	86.68	85.49	85.96	86.35	79.81	83.86	**94.24**	78.36
LP_LR	BoW	**66.25**	**0.204**	**0.338**	60.39	0.216	88.25	**87.27**	**87.70**	**87.91**	79.49	**86.39**	93.64	**81.78**
LP_SVM	BoW	64.55	0.237	0.354	50.00	0.264	85.31	86.19	86.54	87.00	**80.42**	85.10	**94.24**	79.24
AA_MLkNN	BoW	47.85	0.288	0.522	56.46	0.193	88.88	80.04	81.14	81.05	71.32	80.47	89.78	69.87
BR_NB	TF-IDF	64.55	0.237	0.354	50.00	0.264	85.31	86.19	86.54	87.00	**80.42**	85.10	**94.24**	79.24
BR_LR	TF-IDF	64.55	0.238	0.354	49.95	0.264	85.29	86.15	86.50	86.96	80.36	85.10	**94.24**	79.10
BR_SVM	TF-IDF	64.36	0.238	0.356	49.91	0.264	85.29	86.10	86.46	86.92	80.32	85.10	**94.24**	78.97
CC_NB	TF-IDF	64.55	0.237	0.354	50.00	0.264	85.31	86.19	86.54	87.00	**80.42**	85.10	**94.24**	79.24
CC_LR	TF-IDF	64.55	0.237	0.354	50.00	0.264	85.31	86.19	86.54	87.00	**80.42**	85.10	**94.24**	79.24
CC_SVM	TF-IDF	64.55	0.237	0.354	50.00	0.264	85.31	86.19	86.54	87.00	**80.42**	85.10	**94.24**	79.24
LP_NB	TF-IDF	64.55	0.237	0.354	50.00	0.264	85.31	86.19	86.54	87.00	**80.42**	85.10	**94.24**	79.24
LP_LR	TF-IDF	64.55	0.237	0.354	50.00	0.264	85.31	86.19	86.54	87.00	**80.42**	85.10	**94.24**	79.24
LP_SVM	TF-IDF	64.55	0.237	0.354	50.00	0.264	85.31	86.19	86.54	87.00	**80.42**	85.10	**94.24**	79.24
AA_MLkNN	TF-IDF	39.17	0.300	0.608	52.73	0.174	89.32	78.19	80.61	79.69	73.68	80.48	92.15	61.92

Similarly, Tables 5–7 show the gender-based (male, female, and both) evaluation of machine learning and deep learning models for the Celebrex dataset. In the sample-based metrics, the EM (66.25%, 67.61%, and 67.48%) and ZOL (0.338, 0.324, and 0.328) achieve better scores with LP_LR + BoW, CC_SVM + BoW, CC_LR + BoW for male, female, and both groups, respectively. The HL (0.204, 0.181, and 0.178) achieves better results on each group with LP_LR + BoW and BR_LR + BoW models. Moreover, the sample F1-measure achieves 80.42%, 84.02%, and 83.42% on each group. In the label-based metrics, the macro F1, micro F1, and weighted average F1 achieve 87.27%, 87.70%, and 87.91% with LP_LR + BoW for the male group, 89.19%, 89.54%, and 89.76% scores with BR_SVM + BoW for the female group, and 88.91%, 89.29%, and 89.49% with BR_SVM + BoW for both male and female groups. The AUC score achieves 70.91% with MCBiGRU + GloVe, 76.21% with MCBiLSTM + GloVe, and 75.57% with LSTM_F + GloVe for each group, respectively. Particularly, the traditional multilabel learning models achieve a better F1-score per label on all groups. In the ranked-based measures, the RL and AP achieve 0.068 and 95.40% scores with MCBiGRU_F + GloVe for the male group, and 0.023 and 98.59% scores with MCBiGRU + GloVe for the female group, and 0.025 and 98.51% scores with MCBiLSTM_F + GloVe for both male and female (See Table 8).

The mean test accuracy of the deep learning models with the 5-fold is shown in Table 5 and Fig. 2. In this table, the female group achieves better test accuracy in Abilify and Celebrex datasets. We fixed 0.5 as the threshold value with deep learning models to output the sentiment that belongs to a particular aspect. In both datasets, the Overall, the proposed gender-based multi-aspect sentiment detection task achieves better results with traditional multilabel transformation methods using GloVe, BoW, and TF-IDF features in terms of samples and labels such as EM, HL, ZOL, Macro-F1, Micro-F1, Weighted-F1, and F1-measures for small datasets. The deep learning-based multi-aspect sentiment models perform better results in terms of mean test accuracy, AUC Score, RL, and average precision in both Abilify and Celebrex datasets. Specifically, the deep learn-

Table 6
Model performances for the Celebrex oral female group with 5-fold

Models	Features	EM	HL	ZOL	AUC	RL	AP	Macro	Micro	WA	S	F1-score per label		
								F1	F1	F1	F1	0	1	2
LSTM_F	GloVe	62.68	0.190	0.373	75.21	0.036	97.71	87.68	88.37	88.33	80.56	87.21	94.63	81.22
BiLSTM	GloVe	60.05	0.204	0.399	73.97	0.037	97.63	86.56	87.43	87.33	79.90	85.98	94.90	78.81
GRU_F	GloVe	59.63	0.204	0.404	74.33	0.028	98.32	86.43	87.43	87.22	80.30	86.19	95.07	78.02
BiGRU	GloVe	64.93	0.179	0.351	75.55	0.032	98.17	88.66	89.24	89.27	82.42	88.05	95.16	82.77
MCLSTM_F	GloVe	59.90	0.210	0.401	75.94	0.053	96.63	85.94	86.94	86.76	79.35	85.79	94.82	77.22
MCBiLSTM	GloVe	63.73	0.191	0.363	**76.21**	0.051	96.95	87.87	88.48	88.53	81.14	87.57	94.84	81.21
MCGRU_F	GloVe	53.96	0.219	0.460	75.70	0.027	98.24	84.40	86.11	85.41	79.14	86.17	94.84	72.18
MCBiGRU	GloVe	59.74	0.206	0.403	76.18	**0.023**	**98.59**	86.23	87.36	87.12	80.10	87.08	94.61	76.99
BR_NB	GloVe	67.40	0.202	0.326	50.00	0.258	86.61	88.45	88.75	89.12	83.71	88.42	**95.38**	81.56
BR_LR	GloVe	65.77	0.196	0.342	51.97	0.234	87.55	88.45	88.87	89.11	83.63	88.30	**95.38**	81.67
BR_SVM	GloVe	66.82	0.200	0.332	50.49	0.252	86.79	88.48	88.81	89.14	83.70	88.42	**95.38**	81.65
CC_NB	GloVe	67.40	0.202	0.326	50.00	0.258	86.61	88.45	88.75	89.12	83.71	88.42	**95.38**	81.56
CC_LR	GloVe	67.24	0.202	0.328	50.67	0.255	86.78	88.37	88.67	89.02	83.31	88.30	95.15	81.68
CC_SVM	GloVe	67.40	0.202	0.326	50.00	0.258	86.61	88.45	88.75	89.12	83.71	88.42	**95.38**	81.56
LP_NB	GloVe	67.40	0.202	0.326	50.00	0.258	86.61	88.45	88.75	89.12	83.71	88.42	**95.38**	81.56
LP_LR	GloVe	67.40	0.200	0.326	50.81	0.252	86.89	88.47	88.77	89.11	83.52	88.37	95.23	81.80
LP_SVM	GloVe	67.40	0.202	0.326	50.00	0.258	86.61	88.45	88.75	89.12	83.71	88.42	**95.38**	81.56
AA_MLkNN	GloVe	58.69	0.230	0.413	52.20	0.224	88.07	85.78	86.46	86.63	80.22	85.66	94.64	77.04
BR_NB	BoW	65.56	0.191	0.344	57.82	0.207	88.88	88.19	88.64	88.76	81.66	87.48	94.36	82.75
BR_LR	BoW	63.83	0.183	0.362	62.37	0.188	89.91	88.41	88.94	89.00	81.35	88.61	94.50	82.13
BR_SVM	BoW	66.77	0.184	0.332	53.68	0.230	87.61	**89.19**	**89.54**	**89.76**	**84.02**	88.57	**95.38**	**83.61**
CC_NB	BoW	66.35	0.194	0.336	57.29	0.212	88.72	88.12	88.53	88.69	81.54	87.48	94.32	82.58
CC_LR	BoW	67.03	0.182	0.330	62.32	0.207	89.25	88.72	89.12	89.25	81.06	88.61	94.30	83.27
CC_SVM	BoW	**67.61**	0.199	**0.324**	50.77	0.255	86.80	88.61	88.91	89.26	83.75	88.57	**95.38**	81.87
LP_NB	BoW	67.35	0.197	0.327	53.20	0.239	87.62	88.42	88.77	89.04	82.92	88.06	95.07	82.14
LP_LR	BoW	67.03	**0.181**	0.330	63.41	0.209	89.23	88.73	89.12	89.26	80.61	**88.81**	94.18	83.21
LP_SVM	BoW	67.40	0.202	0.326	50.00	0.258	86.61	88.45	88.75	89.12	83.71	88.42	**95.38**	81.56
AA_MLkNN	BoW	59.37	0.233	0.406	57.04	0.202	89.59	84.97	85.77	85.79	78.20	83.24	94.35	77.33
BR_NB	TF-IDF	67.40	0.202	0.326	50.00	0.258	86.61	88.45	88.75	89.12	83.71	88.42	**95.38**	81.56
BR_LR	TF-IDF	67.03	0.199	0.330	50.56	0.254	86.73	88.55	88.86	89.20	83.73	88.39	**95.38**	81.88
BR_SVM	TF-IDF	67.09	0.198	0.329	50.76	0.254	86.73	88.62	88.93	89.27	83.76	88.42	**95.38**	82.07
CC_NB	TF-IDF	67.40	0.202	0.326	50.00	0.258	86.61	88.45	88.75	89.12	83.71	88.42	**95.38**	81.56
CC_LR	TF-IDF	67.35	0.202	0.327	49.98	0.258	86.61	88.43	88.73	89.10	83.68	88.39	**95.38**	81.52
CC_SVM	TF-IDF	67.40	0.202	0.326	50.00	0.258	86.61	88.45	88.75	89.12	83.71	88.42	**95.38**	81.56
LP_NB	TF-IDF	67.40	0.202	0.326	50.00	0.258	86.61	88.45	88.75	89.12	83.71	88.42	**95.38**	81.56
LP_LR	TF-IDF	67.40	0.202	0.326	50.07	0.258	86.61	88.47	88.76	89.13	83.71	88.45	**95.38**	81.58
LP_SVM	TF-IDF	67.40	0.202	0.326	50.00	0.258	86.61	88.45	88.75	89.12	83.71	88.42	**95.38**	81.56
AA_MLkNN	TF-IDF	45.30	0.269	0.547	51.14	0.187	89.17	81.32	83.44	82.69	77.34	84.10	94.52	65.34

ing methods train the data for 25 epochs. Therefore, they perform better than classical machine learning methods with respect to test accuracy, AUC Score, RL, and average precision.

6. Conclusion

In this paper, we studied the gender-based multi-aspect sentiment detection task using multilabel learning algorithms. Also, we proposed multichannel-based unidirectional and bidirectional recurrent neural networks with attention mechanism for this task. We divided the Abilify Oral and Celebrex Oral dataset into three groups based on gender information, namely, male, female, and both male and female. We then represented these datasets using BoW, TF-IDF, and GloVe word embedding features. Next, we applied multilabel learning methods such as the problem transformation approach and adapted algorithm and bidirectional and multichannel bidirectional recurrent neural networks on these feature representations. The performance of the model was evaluated based on samples, labels, and rank measures. The results showed that traditional multilabel transformation methods perform better on small datasets in terms of samples and labels using GloVe, BoW, and TF-IDF features. Deep learning-based multi-aspect sentiment models perform better results in terms of mean test accuracy, AUC Score, RL, and average precision in both Abilify and Celebrex datasets. As future work, we intend to study the multi-aspect sentiment detection task on different age groups and their age impact with a large amount of data in a distributed environment.

CRediT authorship contribution statement

Ashok Kumar J: Conceptualization, Data curation, Writing – original draft, Software. **Tina Esther Trueman:** Investigation. **Erik Cambria:** Supervision, Methodology, Writing – review & editing.

Table 7
Model performances for the Celebrex oral male and female group with 5-fold

Models	Features	EM	HL	ZOL	AUC	RL	AP	Macro	Micro	WA	S	F1-score per label		
								F1	F1	F1	F1	0	1	2
LSTM_F	GloVe	61.03	0.213	0.390	**75.57**	0.037	97.87	85.87	86.52	86.51	76.84	85.00	92.89	79.91
BiLSTM	GloVe	59.51	0.221	0.405	71.23	0.028	98.43	85.28	86.18	86.11	77.62	84.30	94.32	77.23
GRU_F	GloVe	65.75	0.190	0.342	71.18	0.026	98.31	88.67	89.06	89.29	82.75	88.04	95.11	82.86
BiGRU	GloVe	63.66	0.194	0.363	75.47	0.046	97.26	87.81	88.34	88.43	80.89	87.41	94.39	81.65
MCLSTM_F	GloVe	61.33	0.202	0.387	75.19	0.049	96.56	86.82	87.37	87.38	77.98	86.87	92.82	80.78
MCBiLSTM	GloVe	56.32	0.227	0.437	74.97	**0.025**	**98.51**	84.22	85.67	85.21	78.33	83.81	94.94	73.89
MCGRU_F	GloVe	65.51	0.192	0.345	75.53	0.036	97.84	88.27	88.68	88.85	81.15	87.38	94.51	82.92
MCBiGRU	GloVe	63.95	0.194	0.361	75.47	0.066	96.11	87.84	88.40	88.50	81.25	87.34	94.83	81.36
BR_NB	GloVe	66.78	0.210	0.332	50.00	0.259	86.33	87.96	88.27	88.66	82.99	87.71	**95.13**	81.05
BR_LR	GloVe	65.14	0.204	0.349	51.91	0.234	87.26	87.94	88.36	88.63	82.88	87.54	**95.13**	81.14
BR_SVM	GloVe	66.33	0.208	0.337	50.52	0.253	86.52	88.01	88.34	88.69	82.99	87.71	**95.13**	81.19
CC_NB	GloVe	66.78	0.210	0.332	50.00	0.259	86.33	87.96	88.27	88.66	82.99	87.71	**95.13**	81.05
CC_LR	GloVe	66.53	0.210	0.334	50.48	0.257	86.43	87.87	88.18	88.56	82.62	87.54	94.95	81.12
CC_SVM	GloVe	66.78	0.210	0.332	50.00	0.259	86.33	87.96	88.27	88.66	82.99	87.71	**95.13**	81.05
LP_NB	GloVe	66.78	0.210	0.332	50.00	0.259	86.33	87.96	88.27	88.66	82.99	87.71	**95.13**	81.05
LP_LR	GloVe	66.57	0.209	0.334	50.61	0.255	86.53	87.93	88.25	88.61	82.78	87.53	95.06	81.18
LP_SVM	GloVe	66.78	0.210	0.332	50.00	0.259	86.33	87.96	88.27	88.66	82.99	87.71	**95.13**	81.05
AA_MLkNN	GloVe	58.04	0.235	0.410	52.50	0.223	87.89	85.41	86.10	86.23	79.46	85.09	94.28	76.86
BR_NB	BoW	64.98	0.197	0.350	57.17	0.212	88.43	87.91	88.34	88.50	81.30	87.11	94.18	82.44
BR_LR	BoW	64.52	**0.178**	0.355	64.19	0.178	90.30	88.63	89.18	89.23	81.60	**88.36**	94.85	82.67
BR_SVM	BoW	66.08	0.187	0.339	54.76	0.225	87.62	**88.91**	**89.29**	**89.49**	**83.42**	88.05	**95.13**	83.56
CC_NB	BoW	65.75	0.200	0.342	56.49	0.219	88.19	87.85	88.24	88.43	81.17	87.11	94.09	82.34
CC_LR	BoW	**67.48**	0.179	**0.325**	64.44	0.202	89.41	88.80	89.21	89.34	80.79	**88.36**	94.43	**83.62**
CC_SVM	BoW	67.19	0.203	0.328	51.60	0.253	86.68	88.29	88.6	88.95	83.11	88.05	**95.13**	81.69
LP_NB	BoW	66.74	0.203	0.333	52.91	0.240	87.39	88.05	88.42	88.71	82.67	87.62	94.99	81.53
LP_LR	BoW	66.65	0.184	0.333	64.43	0.207	89.09	88.48	88.86	88.98	79.70	88.05	93.81	83.59
LP_SVM	BoW	66.78	0.210	0.332	50.00	0.259	86.33	87.96	88.27	88.66	82.99	87.71	**95.13**	81.05
AA_MLkNN	BoW	56.27	0.252	0.437	56.06	0.202	89.29	83.50	84.39	84.4	76.35	81.71	93.59	75.19
BR_NB	TF-IDF	66.78	0.210	0.332	50.00	0.259	86.33	87.96	88.27	88.66	82.99	87.71	**95.13**	81.05
BR_LR	TF-IDF	66.61	0.203	0.334	51.36	0.250	86.67	88.26	88.59	88.91	83.12	87.73	**95.13**	81.92
BR_SVM	TF-IDF	66.65	0.200	0.333	51.84	0.247	86.74	88.41	88.74	89.04	83.18	87.69	**95.13**	82.40
CC_NB	TF-IDF	66.78	0.210	0.332	50.00	0.259	86.33	87.96	88.27	88.66	82.99	87.71	**95.13**	81.05
CC_LR	TF-IDF	66.82	0.209	0.332	50.23	0.258	86.40	87.99	88.30	88.68	82.98	87.73	**95.13**	81.11
CC_SVM	TF-IDF	66.78	0.209	0.332	50.14	0.259	86.37	87.97	88.27	88.66	82.96	87.69	**95.13**	81.07
LP_NB	TF-IDF	66.78	0.210	0.332	50.00	0.259	86.33	87.96	88.27	88.66	82.99	87.71	**95.13**	81.05
LP_LR	TF-IDF	66.82	0.209	0.332	50.23	0.258	86.37	87.99	88.30	88.68	83.00	87.74	**95.13**	81.08
LP_SVM	TF-IDF	66.78	0.210	0.332	50.00	0.259	86.33	87.96	88.27	88.66	82.99	87.71	**95.13**	81.05
AA_MLkNN	TF-IDF	48.07	0.257	0.519	52.06	0.197	88.47	82.82	84.38	83.99	77.99	85.11	93.85	69.50

Table 8
Mean validation and test accuracy for deep learning models with 5-fold

Dataset	Models	Features	Male		Female		Male and Female	
			Train	Test	Train	Test	Train	Test
Abilify oral	LSTM_F	GloVe	64.11	61.42	69.61	67.13	66.78	**66.19**
	BiLSTM	GloVe	63.52	57.99	70.32	66.61	66.18	63.03
	GRU_F	GloVe	63.47	**61.83**	68.60	64.00	66.79	66.44
	BiGRU	GloVe	63.87	57.41	69.44	65.56	66.75	63.41
	MCLSTM_F	GloVe	62.31	59.85	70.74	**67.37**	66.77	62.34
	MCBiLSTM	GloVe	63.89	57.00	**70.96**	66.10	67.15	62.20
	MCGRU_F	GloVe	63.76	57.26	70.33	63.02	**67.45**	62.77
	MCBiGRU	GloVe	**64.16**	54.78	69.99	63.11	66.04	63.14
Celebrex oral	LSTM_F	GloVe	72.27	**72.07**	81.30	78.39	81.84	78.31
	BiLSTM	GloVe	73.29	71.61	82.28	77.44	81.99	77.44
	GRU_F	GloVe	72.12	70.77	81.19	78.23	81.77	**78.62**
	BiGRU	GloVe	73.08	69.93	81.73	**79.16**	82.66	77.60
	MCLSTM_F	GloVe	73.08	69.51	82.39	76.99	**83.58**	77.82
	MCBiLSTM	GloVe	**74.93**	67.46	**83.28**	77.68	83.45	77.12
	MCGRU_F	GloVe	74.52	69.32	82.16	78.56	82.25	78.39
	MCBiGRU	GloVe	72.52	66.94	82.07	77.35	83.35	77.17

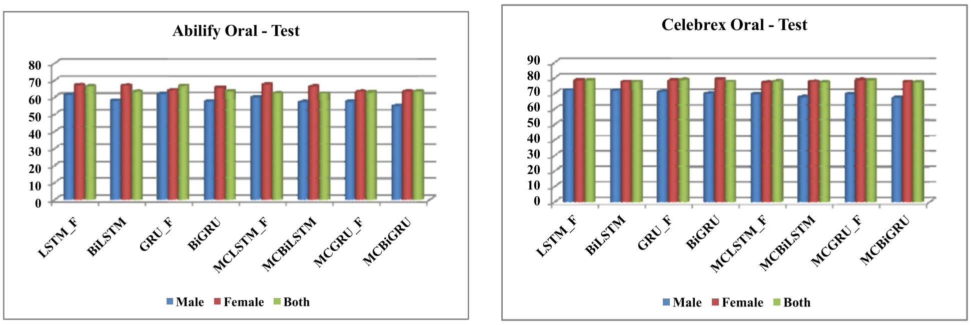

Fig. 2. Training and test accuracy of the deep learning models with 5-fold.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the University Grants Commission (UGC), Government of India under the National Doctoral Fellowship.

References

[1] M. Afzaal, M. Usman, A.C. Fong, S. Fong, Multiaspect-based opinion classification model for tourist reviews, Expert Syst. 36 (2) (2019) e12371.
[2] I. Amarasekara, W.J. Grant, Exploring the youtube science communication gender gap: A sentiment analysis, Public Understand. Sci. 28 (1) (2019) 68–84.
[3] J.-W. Bi, Y. Liu, Z.-P. Fan, E. Cambria, Modelling customer satisfaction from online reviews using ensemble neural network and effect-based kano model, Int. J. Prod. Res. 57 (22) (2019) 7068–7088.
[4] L. Branz, L.P. Reina, J. Richter, B. Waizmann, P. Brockmann, Sentiment analysis of male and female developer comments, ICSE (2019).
[5] A. Bukeer, G. Roffo, A. Vinciarelli, Type like a man! inferring gender from keystroke dynamics in live-chats, IEEE Intell. Syst. 34 (6) (2019) 53–59.
[6] E. Cambria, A. Hussain, T. Durrani, C. Havasi, C. Eckl, and J. Munro. Sentic computing for patient centered applications. In IEEE ICSP, pages 1279–1282, 2010..
[7] E. Cambria, B. Schuller, B. Liu, H. Wang, C. Havasi, Statistical approaches to concept-level sentiment analysis, IEEE Intell. Syst. 28 (3) (2013) 6–9.
[8] E. Cambria, S. Poria, F. Bisio, R. Bajpai, and I. Chaturvedi. The CLSA model: A novel framework for concept-level sentiment analysis. In LNCS, volume 9042, pages 3–22. Springer, 2015..
[9] J. Carver, R. Capilla, B. Penzenstadler, A. Serebrenik, A. Valdezate, Gender, sentiment and emotions, and safety-critical systems, IEEE Softw. 35 (6) (2018) 16–19.
[10] G. Chen, D. Ye, E. Cambria, J. Chen, and Z. Xing. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In IJCNN, pages 2377–2383, 2017..
[11] M. Díaz, I. Johnson, A. Lazar, A.M. Piper, D. Gergle, Addressing age-related bias in sentiment analysis, in: Proceedings of the 2018 chi conference on human factors in computing systems, 2018, pp. 1–14.
[12] F. Ertam, An effective gender recognition approach using voice data via deeper lstm networks, Appl. Acoust. 156 (2019) 351–358.
[13] F.O. Gallego, R. Corchuelo, Torii: An aspect-based sentiment analysis system that can mine conditions, Software: Practice and Experience 50 (1) (2020) 47–64.
[14] D. Ganda, R. Buch, A survey on multi label classification, Recent Trends Programming Lang. 5 (1) (2018) 19–23.
[15] A. Garcia-Rudolph, S. Laxe, J. Saurí, M.B. Guitart, Stroke survivors on twitter: sentiment and topic analysis from a gender perspective, J. Med. Internet Res. 21 (8) (2019) e14077.
[16] O.L. Haimson, Mapping gender transition sentiment patterns via social media data: toward decreasing transgender mental health disparities, J. Am. Med. Inform. Assoc. 26 (8–9) (2019) 749–758.
[17] S. Hussein, M. Farouk, E. Hemayed, Gender identification of egyptian dialect in twitter, Egypt. Inf. J. 20 (2) (2019) 109–116.
[18] K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, J. Documentation (1972).
[19] M.F. Kacamarga, T.W. Cenggoro, A. Budiarto, R. Rahutomo, B. Pardamean, Analysis of acoustic features in gender identification model for english and bahasa indonesia telephone speeches, Proc. Comput. Sci. 157 (2019) 199–204.
[20] A. Khatua, A. Khatua, E. Cambria, Predicting political sentiments of voters from twitter in multi-party contexts, Appl. Soft Comput. 97 (106743) (2020).
[21] S. Kiritchenko and S.M. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. arXiv preprint arXiv:1805.04508, 2018..
[22] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown. Text classification algorithms: a survey. information 10 (4): 150. arXiv preprint arXiv:1904.08067, 2019..
[23] J.A. Kumar, S. Abirami, T.E. Trueman, Multilabel aspect-based sentiment classification for abilify drug user review, in: 2019 11th International Conference on Advanced Computing (ICoAC), IEEE, 2019, pp. 376–380.
[24] J.A. Kumar, S. Abirami, T.E. Trueman, E. Cambria, Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit, Neurocomputing 441 (2021) 272–278.
[25] J.A. Kumar, T.E. Trueman, E. Cambria, A convolutional stacked bidirectional lstm with a multiplicative attention mechanism for aspect category and sentiment detection, Cognitive Computation 13 (6) (2021) 1423–1432.
[26] S. Kumar, M. Gahalawat, P.P. Roy, D.P. Dogra, B.-G. Kim, Exploring impact of age and gender on sentiment analysis using machine learning, Electronics 9 (2) (2020) 374.
[27] Y. Li, S. Wang, Y. Ma, Q. Pan, E. Cambria, Popularity prediction on vacation rental websites, Neurocomputing 412 (2020) 372–380.

[28] B. Liang, H. Su, L. Gui, E. Cambria, R. Xu, Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks, Knowl.-Based Syst. 235 (107643) (2022).
[29] Y. Ma, K.L. Nguyen, F. Xing, E. Cambria, A survey on empathetic dialogue systems, Information Fusion 64 (2020) 50–70.
[30] Y. Mehta, N. Majumder, A. Gelbukh, E. Cambria, Recent trends in deep learning based personality detection, Artif. Intell. Rev. 53 (2020) 2313–2339.
[31] R. Mihalcea, A. Garimella, What men say, what women hear: Finding gender-specific meaning shades, IEEE Intell. Syst. 31 (4) (2016) 62–67.
[32] J.H. Park, J. Shin, and P. Fung. Reducing gender bias in abusive language detection. arXiv preprint arXiv:1808.07231, 2018..
[33] S. Park, J. Woo, Gender classification using sentiment analysis and deep learning in a health web forum, Appl. Sci. 9 (6) (2019) 1249.
[34] F. Pedregosa, G. Varoquaux, et al. Scikit-learn-machine learning in python, 2011..
[35] M.P. Pondé, A.C.C. Freire, Increased anxiety, akathisia, and suicidal thoughts in patients with mood disorder on aripiprazole and lamotrigine, Case Rep. Psychiatry (2015.).
[36] R. Satapathy, E. Cambria, A. Nanetti, A. Hussain, A review of shorthand systems: From brachygraphy to microtext and beyond, Cogn. Comput. 12 (4) (2020) 778–792.
[37] A. Sboev, T. Litvinova, I. Voronina, D. Gudovskikh, R. Rybka, Deep learning network models to categorize texts according to author's gender and to identify text sentiment, in: 2016 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, 2016, pp. 1101–1106.
[38] M.S. Sorower, A literature survey on algorithms for multi-label learning, Oregon State University, Corvallis 18 (2010) 1–25.
[39] R.C. Staudemeyer and E.R. Morris. Understanding lstm–a tutorial into long short-term memory recurrent neural networks. arXiv preprint arXiv:1909.09586, 2019..
[40] R. Sukthanker, S. Poria, E. Cambria, R. Thirunavukarasu, Anaphora and coreference resolution: A review, Inf. Fusion 59 (2020) 139–162.
[41] C. Suman, P. Kumar, S. Saha, and P. Bhattacharyya. Gender age and dialect recognition using tweets in a deep learning framework-notebook for fire 2019. In FIRE (Working Notes), pages 160–166, 2019..
[42] A. Swaminathan, M. Chaba, D.K. Sharma, Y. Chaba, Gender classification using facial embeddings: A novel approach, Proc. Comput. Sci. 167 (2020) 2634–2642.
[43] C. Tannenbaum, R.P. Ellis, F. Eyssel, J. Zou, L. Schiebinger, Sex and gender analysis improves science and engineering, Nature 575 (7781) (2019) 137–146.
[44] E. Teso, M. Olmedilla, M. Martínez-Torres, S. Toral, Application of text mining techniques to the analysis of discourse in ewom communications from a gender perspective, Technol. Forecast. Soc. Chang. 129 (2018) 131–142.
[45] M. Thelwall, Gender bias in sentiment analysis, Online Information Review (2018).
[46] J.K. Treceñe, Delving the sentiments to track emotions in gender issues: A plutchik-based sentiment analysis in students' learning diaries, Int. J. Scientific Technol. Res. 8 (12) (2019) 1134–1139.
[47] Y. Xia, E. Cambria, A. Hussain, Aspnet: Aspect extraction by bootstrapping generalization and propagation using an aspect network, Cogn. Comput. 7 (2) (2015) 241–253.
[48] F. Xing, E. Cambria, R. Welsch, Intelligent asset allocation via market sentiment views, IEEE Comput. Intell. Mag. 13 (4) (2018) 25–34.
[49] J. Zheng, F. Cai, T. Shao, H. Chen, Self-interaction attention mechanism-based text representation for document classification, Appl. Sci. 8 (4) (2018) 613.
[50] Y. Zheng, L. Li, J. Zhang, Q. Xie, L. Zhong, Using sentiment representation learning to enhance gender classification for user profiling, in: Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data, Springer, 2019, pp. 3–11.