Avnesh Kumar JOSHI, PhD Candidate (corresponding author)
avneshjoshi24@gmail.com
Jaypee University of Engineering and Technology, Guna, Madhya Pradesh, India

Ajay KUMAR, PhD
ajaymits@gmail.com
Jaypee University of Engineering and Technology, Guna, Madhya Pradesh, India

Nilesh Kumar PATEL, PhD
nilesh.juet@gmail.com
Nirma University, Ahmedabad, Gujarat, India

Leveraging Social Media and Multitask Natural Language Processing to Detect Financial Consumer Complaints

Abstract. *The widespread reach and one-to-one interactions with businesses through social media now allow people to speak their minds easily. It enhances customer services by making it easier for users to raise complaints and hold the businesses accountable, while also allowing companies to respond promptly in real time. Public visibility can also support others with similar issues and influence other potential customers. This benefits both customers and businesses, creating a feedback loop that can lead to improvements in products or services. In natural language processing, classifying complaints based on severity level is a key challenge. Also, in human conversations, a single sentence can carry multiple meanings, making it a complex task to classify these meanings effectively using a multitasking model. In this work, we proposed a multitasking model applied to a financial dataset containing tweets from bank customers. The proposed multitasking framework classifies four tasks: complaint label, severity label, emotion, and sentiment for each tweet. Our model outperformed existing models, achieving an accuracy of 97.93% and an F1 score of 97.95% for the complaint identification task, and an accuracy of 84.49% with an F1 score of 84.02% for the complaint severity task.*

Keywords: *Twitter, transformers, sentiment analysis, deep learning, Natural Language Processing.*

JEL Classification: C10, D12, G20.

Received: 11 June 2025	Revised: 14 June 2025	Accepted: 10 September 2025

1. Introduction

Social media serves as a platform for sharing ideas, emotions, and experiences. It enables users to connect and engage with a wide range of perspectives on trending issues. When topics go viral, they often reflect the collective thoughts and feelings of society, creating a dynamic dialogue that can influence public opinion and drive change (Dwivedi et al., 2021; Sandiwarno, Niu, and Nyamawe, 2024).

DOI: 10.24818/18423264/59.3.25.09

Artificial intelligence (AI) integration in social media is transforming how users and businesses interact and engage with content. With the use of AI tools, users and businesses can enhance interactions on social media and engage with its content for Content personalization, Trend analysis, Sentiment analysis, Moderation, Influencer identifier, Engagement optimization and Crisis management through personalized content recommendations and sentiment analysis (Joshi, Kumar, & Patel, 2025). Overall, the ability of AI technology to process and analyze massive amounts of data enhances our understanding about social media and helps to find valuable and useful insights from it (Ghosh et al., 2023).

Organizations which are working in the finance sectors are flooded with customer complaints and feedback coming from various channels such as social media, emails, and customer support platforms. To enhance customer satisfaction and streamline operations, it's crucial to manage and categorize these complaints efficiently. Traditional models that focus on single-task classification like classifying a message into complaint or non-complaint, fails in capturing additional critical information like the sentiment behind the complaint, the emotions involved, and the severity of the issue (Hake, Rehse, & Fettke, 2021).

Natural language processing (NLP) for complaint classification helps businesses, particularly in the financial sector, automate and speed up how they handle customer complaints. Another important aspect is severity classification, which organizes complaints based on their impact and urgency—like critical, high, moderate, and low. By adopting a solid severity classification system, NLP enables organizations to improve customer service, streamline their operations, and respond more effectively to what customers need (Hake, Rehse, & Fettke, 2021), (Coussement & den Poel, 2008), (Ilias, Mouzakitis, & Askounis, 2024).

A multitask learning framework that offers a fuller understanding of customer complaints, enhancing both the accuracy and efficiency of complaint classification. By addressing sentiment, emotion, and severity alongside complaint identification, organizations can gain a deeper understanding of customer concerns, enabling them to respond more quickly and effectively (Oyewola, Omotehinwa, & Dada, 2023). Previous studies show that emotion and sentiment are closely related in understanding how customers think (Singh, Bhatia, & Saha, 2023) . By classifying severity, emotion, and sentiment together, models can better grasp the nature of complaints and help businesses prioritize their responses more effectively (Vairetti et al., 2004; Jin & Aletras, 2021). In this work, we introduce a multitasking framework for classifying complaints, specifically targeting issues between consumers and financial organizations. Our model handles four tasks at once and outperforms previous baseline models using deep neural networks.

The primary contributions of our proposed model are as follows:

1. We developed a multitasking, transformer-based model that classifies complaints into two categories (Complaint and Non-Complaint) and sentiment into three categories (Positive, Negative, and Neutral).
2. Furthermore, the theory groups six emotions—disgust, anger, fear, happiness, sadness and surprise—according to Akman's work.
3. Accusation, blame, disapproval and no direct method are the other gravities that the model uses to set them apart.
4. Our BERT-based model achieves better performance than previous approaches, with accuracy of 97.93% and an F1 score of 97.95% for the complaint identification task, and an accuracy of 84.49% with an F1 score of 84.02% for the complaint severity task.
5. The author of the RoBERTa-based (Singh, Bhatia, & Saha, 2023) model compared results for the quad task. However, we have shown how our proposed model performs across single-task, dual-task, tri-task, and quad-task scenarios.

The second part of the study presents current research topics and covers a review of relevant publications. Details of the approach used in the investigation can be found in Section 3. The results of the experiment and their explanation are explained in Section 4. Section 5 ends and provides advice for carrying out more research.

2. Literature review

Recent research on complaint classification in linguistics has attracted a lot of attention because it benefits both the consumers and the organizations.

Vairetti et al. (2004), by using state-of-the-art models like BERT and RoBERTa, presented a framework that effectively classifies complaints into urgent and non-urgent categories. This approach improves response times and resource allocation, showing significant improvements over traditional methods.

In (Jin & Aletras, 2021), using technology and important linguistic elements, the authors recommend a process to categorize the seriousness of complaints shared on social media. Complaints are sorted into four groups depending on their level of seriousness: blame, accusation, disapproval and not taking a direct approach. The study highlights how the emotional nuances in complaints can help to improve customer service and other NLP applications.

Recently, Singh et al. (2023) presented a multitask learning framework using RoBERTa that significantly outperforms existing models in classifying complaints and their associated attributes. Their work highlights the importance of integrating emotion and sentiment analysis in complaint detection.

In references (Jin & Aletras, 2021), and (Islam et al., 2023), the Transformer model was introduced by the authors and it uses self-attention instead of the usual recurrent or convolutional layers. As opposed to the earlier models, this is simpler to scale up, trainer much more rapidly and uses fewer resources. Further, Fadel et al.

(2024) introduce the MTL-AraBERT model, a multi-task learning framework that simultaneously addresses aspect term extraction (ATE) and aspect category detection (ACD) (Dagur et al., 2024) in Arabic sentiment analysis. This model shows significant performance improvements compared to traditional single-task models. It utilizes the AraBERT architecture and deep learning techniques, achieving better results on benchmark datasets.

Wang et al. (2021) developed a new sentiment analysis model, the Fuzzy-Deep Neural Network (Fuzzy-DNN). This model uses information from text, audio, and images to better capture and understand emotions, making the sentiment analysis more accurate and detailed. Because BERT-based models require a lot of expensive computing power, Gong et al. (2022) introduced a new way to improve sentiment analysis. They boost Transformer-based model results by using both text improvement and knowledge distillation approaches when the available data is not heavily labelled.

Gupta et al. (2021), developed an advanced Attention-based Bi-Directional LSTM model using Fast Text Embedding to classify complaints in the Indian Railways' complaint management system. The goal is to improve service quality and make railway related complaint resolution more efficient.

In (CARUANA, 1997), the author presents a comprehensive exploration of the Multitask Learning (MTL) mechanism with a case-based method applicable to many domains.

Bensoltane and Zaki (2024), introduce a new end-to-end model for Arabic aspect-based sentiment analysis (ABSA), Both extraction of aspect phrases and the determination of their emotions is achieved using a neural multi-task technique. It surpasses the problems with previous research that have mainly relied on standard or rule-based machine learning.

Tan et al. (2023) suggest a way to learn both understanding sarcasm and analysing sentiments at the same time. It displays that the use of sarcasm detection helps substantially improve the accuracy of sentiment analysis tools.

Dong et al. (2022) find out the roles text-based techniques and natural language processing (NLP) could play in the railway sector, especially to raise productivity, improve safety and lower costs.

In Zhao et al. (2023), BERT and RGAT are used by the author to create a new multitask learning model that brings together APC and ATE. Over many sets of data, this way of collecting features makes sentiment analysis much better.

Sandiwarno, Niu, and Nyamawe (2024), highlighted that user satisfaction is important to improve platforms of e-learning. The existing methods are often not formulated to integrate sentiment, emotion, and semantic relations completely. However, such features are not learned and combined well by the proposed SES-Net, a multi-task deep neural model that combines Bi-LSTM and CNN, which achieves only a 90.59% F1-score.

The literature review shows that there is a strong connection between emotion and sentiment that encourages us to take both into account when analysing

complaints. Since sentiment and emotion are key aspects of human interactions, they play an important role in creating more effective and adaptable AI-based systems.

3. Model specification

Here, we explain what is being compared and list the current methods being use. Additionally, we provide an overview of the methodology used in the proposed model, highlighting the improvements it offers over the existing approaches.

3.1 Problem Definition

In multitask complaint classification problem (Singh, Bhatia, & Saha, 2023), It is the aim to train a model that can separate different parts of a complaint together, such as sentiment (positive, neutral, negative), emotions (e.g., anger, sadness), severity of the complaint, and whether the content qualifies as a formal complaint. This is achieved by feeding a shared input (text data) into a model like BERT, which generates contextual embeddings for each task.
Mathematically, each task T_i (for sentiment, emotion, severity, and complaint detection) has its own output head, and the total loss L_{total} is the total of the individual losses from each task defined in Equation (1).

$$L_{total} = \sum_{i=1}^{N} L(T_i) \quad (1)$$

Where, $L(T_i)$ represents the categorical cross-entropy loss for the i^{th} task. By adding these losses together, the model optimizes all tasks at once, which helps the learning of shared representations that work well across various tasks. In this fashion, the model works better and embeds the connections between features, resulting in better maneuvering while assigning categories to complaints.

3.2 Complaint classification using MTL-RoBERTa

In the recent work of multitask complaint classification using MTL-RoBERTa (Singh, Bhatia, & Saha, 2023), the model is trained to predict multiple outputs for each input instance (such as a customer complaint or feedback). The input data consists of a set of text instances $X = \{x_1, x_2, \ldots\ x_n\}$,, where each x_i represents a text sequence. This model performs the following tasks:

1. *Complaint Classification*: A kind of task where the model figures out if the text is a complaint or not $(y_1 \in \{0, 1\})$, where 0 denotes "non-complaint" and 1 denotes "complaint".
2. *Sentiment Classification*: A multiclass classification task where the model predicts the sentiment of the text $(y_2 \in \{0, 1, 2\})$, where, “negative,” “neutral,” and “positive” are assigned the values 0, 1 and 2, respectively.

3. *Emotion Recognition*: Another multiclass classification task where the model predicts the emotion conveyed in the text$(y_3 \in \{e_1, e_2, e_3\})$, with different emotions such as anger, sadness, or happiness.
4. *Severity Classification*: A multiclass classification task that predicts the severity of the complaint $(y_4 \in \{s_1, s_2, s_3, s_4\})$, where s_1, s_2, s_3, s_4 show varying levels of severity such as minor, moderate, or severe.

The objective of MTL-RoBERTa is to learn a function $f(X, \theta)$, parameterized by θ, which predicts the outputs for each task: $Y = \{y_1, y_2, y_3, y_4\}$ (2)
For each input x_i. The model uses shared parameters that are learned from all tasks, as well as task-specific parameters that fine-tune the performance for each task.

3.3 Proposed Methodology (Multitask BERT Model)

In transformer-based NLP, BERT is a widely used pre-trained model for a variety of tasks. There are also popular variants like Electra and RoBERTa that are often applied in NLP tasks (Casola, Lauriola, & Lavelli, 2022). For example, in (Singh, Bhatia, & Saha, 2023), MTL-RoBERTa is used to improve binary complaint classification by factoring in severity, emotion, and sentiment. It uses a shared encoder for all tasks and separate branches for each specific task, which helps the model classify complaints more effectively.

In the baseline model (MTL-RoBERTa), which we used for comparison, a variant of the BERT transformer is used. In contrast, our proposed framework directly uses the BERT model (shown in Figure 1), leading to better performance. We also applied the AdamW optimization algorithm, which enhances performance by better handling weight decay and improving model generalization.

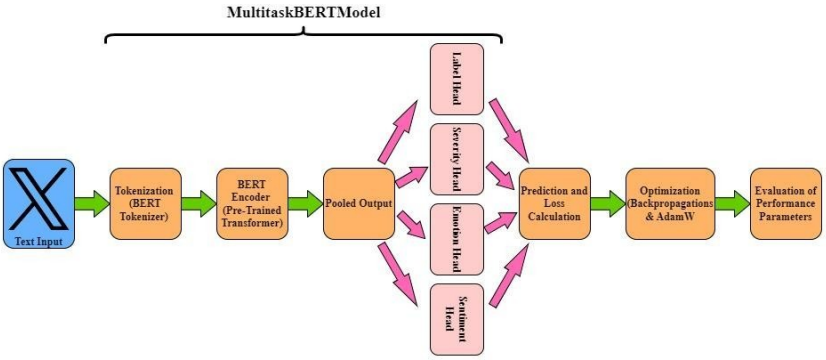

Figure 1. Proposed Model
Source: Authors' own creation.

3.3.1 Tokenization (BERT Tokenization)

Getting the data needed by our model requires special consideration, we used the BERT (Bi-directional Encoder Representations from Transformers) Tokenizer, which converts text into numerical values. We applied this tokenization to FINCROP dataset to make the process more efficient. We also created an attention mask to distinguish between meaningful tokens and padding tokens, ensuring the data works well with the pre-trained BERT model.

In the BERT tokenization process, each word in a sentence is broken down into smaller tokens, and embeddings are created for each of them (Gong, Ying, Zhong, & Gong, 2022). Token, segment and positional embeddings are combined to form the final embedding, as shown in Figure 2.

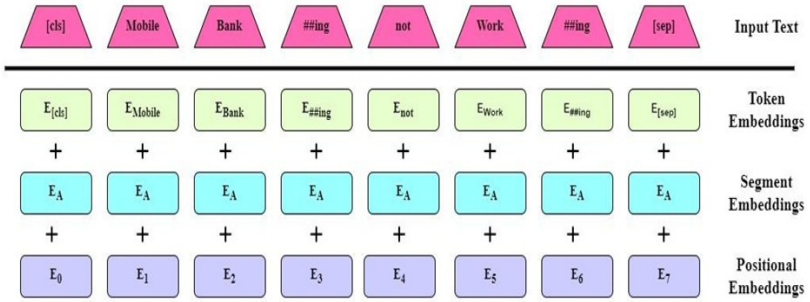

Figure 2. BERT Tokenization
Source: Authors' own creation.

3.3.2 BERT Encoder

To input data into the BERT[1] encoder, we provide tokenized text along with an attention mask. The BERT model processes this input and generates a contextualized representation for each token. The model is built on a self-attention mechanism and is trained using two key tasks: predicting masked tokens and predicting the next sentence (Cesario, Comito, & Zumpano, 2024).

Figure 3 shows the architectural representation of proposed Multitask BERT Model.

The overall framework is divided into 4 layers, and it is explained below.

[1] "https://en.wikipedia.org/wiki/BERT_(language_model)"

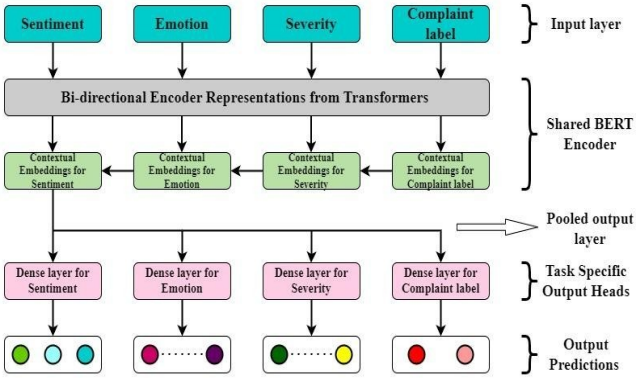

Figure 3. Multitask BERT Model Architecture
Source: Authors' own creation.

***Input layer*:** Here, we provide inputs for the different tasks the model needs to perform. For example, the text data is used to predict sentiment (positive, neutral, negative), emotions (For example, people feel rage, disgust, fear, joy, sorrow and surprise), severity levels (accusation, blame, disapproval, no explicit reproach), and complaint label (complaint, non-complaint).

***Shared BERT Encoder layer*:** This is the main part of the proposed model. We use BERT, a transformer model that is very effective and was trained by combining next sentence prediction (NSP) and masked language modelling (MLM) (CARUANA, 1997). NSP is especially helpful when it's important to understand the relationship between sentences in a document.
With every task, BERT works together to produce contextualized word representations, allowing the model to understand the text better in its whole context.

***Pooled output layer*:** This layer offers a summarized representation of the input text sequence. It includes the outputs from each BERT encoder layer and sends the input sequence to the specific output heads for each task.

***Task specific output heads layer*:** Once the contextual embeddings for the entire sentence are created, they are sent to task-specific output heads. Each head is a fully connected (dense) layer designed to make predictions for a particular task (CARUANA, 1997). This scenario shows four dense layers set up for the multitask scenarios of sentiment, emotion, severity, and complaint labelling.

***Output prediction layer*:** This is the final configuration of the Multitask BERT model framework, which is designed to generate predictions for multitask processing. The output from this layer is organized in the same way as the input text, categorizing it into different classes. The different colours in each head represent the various classes or labels that each task may predict.

3.4 Prediction and Loss calculation

In the proposed model, we have calculated categorical cross entropy loss[2] of all individual tasks and sum up all these losses for total loss calculation during model execution.

3.4.1 Task Specific loss calculation

We have assumed, there are *T* tasks and for each task *t* (here *t* =4 for complaint label, severity, emotion, and sentiment).

We calculate task specific loss denoted asL_t, for a single task t, categorical cross entropy loss calculation can be calculated as:

$$L_t = -\frac{1}{N_t}\sum_{i=1}^{N_t}\sum_{c=1}^{C_t} y_{i,c}^{(t)} \log\left(\hat{y}_{i,c}^{(t)}\right) \quad (3)$$

Where,

N_t, total number of samples for task t,

C_t, total number of classes for task t,

$y_{i,c}^{(t)}$, the exact group and sample assigned to class c and sample i for task t,

$\hat{y}_{i,c}^{(t)}$, The predicted chance that sample i is in group c when it is supposed to be in task t

3.4.2 Combined loss calculation

For calculating total loss L_{total} of Multitask BERT Model, we have combined all losses from tasks. This is calculated using Equation (4).

$$L_{total} = \sum_{t=1}^{T} L_t \quad (4)$$

3.5 Optimization

Here we have used AdamW [3] optimization algorithm that merges the functionality of Adam for enhancing generalization of outcomes. AdamW uses these gradients to update the parameters, putting together the strengths of weight decay regularization and adaptive learning rates.

3.5.1 Gradients Calculation (backward pass)

With the chain system, we have determined the gradient of the overall loss for every model parameter shows in Equation (5).

$$\nabla_t L_{total} = \frac{\partial L_{total}}{\partial \theta} = \sum_{t=1}^{T} \lambda_t \nabla_\theta L_t \quad (5)$$

Afterward, the model learns by changing its parameters according to the gradients.

[2] "https://www.tensorflow.org/api_docs/python/tf/keras/losses/CategoricalCrossentropy"

[3] "https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html"

3.5.2 Parameter update procedure

for each parameter θ_i at the time stamp t, the steps to update rule in AdamW is:

a) Compute gradients: $g_{t,i} = \nabla_{\theta_i} L_{total}$ (6)

b) Now moving average of gradients:

$$m_{t,i} = \beta_1 m_{t-1,i} + (1 - \beta_{1)} g_{t,i} \quad (7)$$

Where, β_1 is estimate of the exponential decay rate for the first moment.

c) Now moving average of squared gradients:

$$v_{t,i} = \beta_2 v_{t-1,i} + (1 - \beta_2) g_{t,i}^2 \quad (8)$$

Where, β_2 is the rate of exponential decrease in the second moment.

d) Now, correct the bias values in the first and second moment estimate:

$$\widehat{m}_{t,i} = \frac{m_{t,i}}{1-\beta_1^t} \quad (9)$$

$$\hat{v}_{t,i} = \frac{v_{t,i}}{1-\beta_2^t} \quad (10)$$

e) Now calculate the update for the parameter with weight decay:

$$\theta_{t,i} = \theta_{t-1,i} - \eta \left(\frac{\widehat{m}_{t,i}}{\sqrt{\widehat{v_{t,i}}}} + \lambda \theta_{t-1,i} \right) \quad (11)$$

Were,

η is the learning rate,

$\in$ is the tiny constant that keeps division by zero from happening,

λ is the coefficient of weight decay.

4. Results and discussion

4.1 Experimental setup

We have used Google colab platform including scikit-learn[4] ,TensorFlow[5] ,Pytorch package[6] .The pre-trained BERT Model was taken from open source huggingface transformer library[7]. We contrasted the accuracy and F1 score of our suggested model with those of earlier models in order to assess its performance. We applied categorical cross entropy loss[8] for training and ran the model for 7 epochs. To optimize the number of epochs, we used early stopping, which halted training if no improvement was seen for 3 consecutive epochs.

The training used the Adam optimizer with a learning rate of 1e-5. Adam makes the training process faster and steadier by updating the learning rate for every parameter with reference to previous gradients. It calculates two important values: the mean of previous gradients (to smooth out the updates) and the

[4] "https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html"

[5] "https://www.tensorflow.org/"

[6] "https://pytorch.org/get-started/locally/"

[7] "https://huggingface.co/docs/transformers/en/model_doc/bert"

[8] "https://www.tensorflow.org/api_docs/python/tf/keras/losses/CategoricalCrossentropy"

mean of squared gradients (to reduce volatility). These corrections help Adam make larger updates for stable parameters and smaller updates for more unpredictable ones, ensuring the model converges faster and more reliable.

Performance Parameters

To check the model performance, we have used performance parameters accuracy, precision, recall, and f1 score.

[1] Accuracy

It describes the percentage of accurately predicted samples, out of all the samples. (He, Wang, Zou, Dai, & Jin, 2021).

For a given task t: Accuracy = $\frac{\sum_{i=1}^{N_t} 1(\hat{y}_i^{(t)}=y_i^{(t)})}{N_t}$ (12)

Where, N_t is the total number of samples for class T,

$\hat{y}_i^{(t)}$, the predicted label, $y_i^{(t)}$, the actual label for the i-th sample.

[2] Precision

The precision for a model check, how many of the positive predictions turn out to be correct?

For a specific class c in task t: $Precision_{t,c} = \frac{TP_{t,c}}{TP_{t,c}+FP_{t,c}}$ (13)

Where, $TP_{t,c}$ is the number of true positive, $FP_{t,c}$, the number of false positive for class c in task t.

[3] Recall

The proportion of actual positive predictions that are spotted by the model calculates recall.

For a specific class c in task t: $Recall_{t,c} = \frac{TP_{t,c}}{TP_{t,c}+FN_{t,c}}$ (14)

Where, $FN_{t,c}$ is the number of false negative for class c in task t.

[4] F1-score

It uses a mathematical technique to weigh precision and recall,

For a specific class c in task t: $F1\_score_{t,c} = 2 * \frac{Precision_{t,c} * Recall_{t,c}}{Precision_{t,c}+ Reacll_{t,c}}$ (15)

Data Set

We utilized the *FINCORP* dataset as described in (Singh, Bhatia, & Saha, 2023), which consists of customer com- plaints related to banks across various

domains. The dataset includes several attributes such as complaint labels, severity levels, and classifications for emotion and sentiment. According to the authors, these labels were manually annotated into different categories. A snippet of this data is given in Table 1.

The FINCORP dataset was created to support research in complaint identification and severity prediction within the financial sector. Unlike other datasets that only include bi- nary complaints, FINCORP features a balanced collection of 6,282 tweets, almost evenly split between complaints (50.13%) and non-complaints (49.87%). It involves looking at complaints that people post about financial organizations on Twitter. In addition to identifying complaints, the dataset categorizes tweets by severity, sentiment, and emotion, providing a more in-depth look at financial grievances.

In FINCORP data set, most complaint tweets reflect negative sentiment, with sadness and anger being the most common emotions. Frequent complaint keywords include terms like "response," "refund," "RBI," and "fraud".

Table 1. Sample Dataset

Domain	Complaint/Opinion	Complaint label	Severity label	Sentiment	Emotion
Netbanking	@lClClBankcare2 @ICICIBank_Care @ICICIBank Why is netbanking not working? Problem persists since 2 days now. Mobilebanking is not working for my wife too.Why are you not looking into it even after reporting multiple times? I see lot of people reporting this issue.@TOIBreakingNews @RBI	1	Blame	Negative	Sadness
Mortgage	@BarclaysUKHelp Hi I need to request a Certificate of Interest for my BTL mortgage but I can't login and the phone wait times are 30 minutes+, is there any other way I can get this request in?	1	No explicit reproach	Negative	Sadness
Loan	@ICICIBank_Care I am a Customer with a 20 uear relationship with ICICI bank. Took a personal loan. Now trying to foreclose it. The Loan department is not able to provide even the exact nos from last 2 months. Paid twice, still balance keeps coming up	1	Blame	Negative	Anger

Domain	Complaint/Opinion	Complaint label	Severity label	Sentiment	Emotion
Customer Service	@TheOfficialSBI @SBICard_Connect Ap no.2435808056451 Very poor response from verification guy he asked money from me for verification.I email to sbi for this but no response from sbi card too.I called that guy many times but no response.Got a vip approval but sbi sucksThnks https://t.co/eO5N229Vta	1	Blame	Negative	Sadness

Source: (Singh, Bhatia, & Saha, 2023).

4.2 Results and discussion

Compared with the earlier models, our model does a better job of detecting complaints and their severity. This is significantly boosted by the next sentence prediction of BERT that rocks context between linked sentences.

Table 2 compares our proposed model (Multitask BERT Model) with others based on accuracy and F1 score for the Complaint Identification (CI) task across different multitasking setups: single-task, two-task, three-task, and four-task scenarios. In the single-task case, our model achieved the highest accuracy (97.93%) and F1 score (97.90%) compared to other models. In the two-task setup, it continued to perform better, with accuracy between 87.95% and 91.60%, and F1 scores ranging from 88.15% to 91.37%. For the three-task scenario, the model showed further improvement, reaching 92.05% accuracy and a 92.25% F1 score. Lastly, in the four-task setup, our model once again outperformed others, hitting 97.93% accuracy and a 97.95% F1 score, surpassing the MTL-RoBERTa model. It proves that the model performs well and is reliable when handling multitask learning.

Table 2. Results of the proposed model (MultitaskBERTModel) and other models with respect to Accuracy and F1 score for Complaint identification (CI) task

Multitasking Scenario	Model	Accuracy (%)	F1 Score (%)
Single task	STLCI	86.26 ±0.03	85.25 ±0.05
	$Baseline_2$	83.58 ±1.02	82.03 ±1.05
	ProposedModelCI	**97.93±0.50**	**97.90 ±0.55**
Dual-Task	MTL_{CI+ER}	88.69 ±1.07	87.64 ±1.04
	MTL_{CI+SA}	87.58 ±1.08	87.21 ±1.06
	MTL_{CI+CS}	89.87 ±1.04	87.92 ±1.05
	$\boldsymbol{ProposedModel_{CI+ER}}$	**90.71 ±0.55**	**88.15 ±0.55**

Multitasking Scenario	Model	Accuracy (%)	F1 Score (%)
Tri-task	$\boldsymbol{ProposedModel_{CI+SA}}$	**87.95 ±1.05**	**88.45 ±0.65**
	$\boldsymbol{ProposedModel_{CI+CS}}$	**91.60 ±0.85**	**91.37 ±0.50**
	$MTL_{CI+ER+SA}$	90.44 ±1.07	90.13 ±1.03
	$MTL_{CI+ER+CS}$	91.28 ±1.15	90.04 ±1.02
	$MTL_{CI+SA+CS}$	88.39 ±1.12	87.58 ±1.06
	$\boldsymbol{ProposedModel_{CI+ER+SA}}$	**91.05 ±0.55**	**91.25 ±0.60**
	$\boldsymbol{ProposedModel_{CI+ER+CS}}$	**92.05 ±0.45**	**92.25 ±0.50**
	$\boldsymbol{ProposedModel_{CI+SA+CS}}$	**90.35 ±0.55**	**90.50 ±0.40**
	MTL-RoBERTa	94.38 ±0.13	93.78 ± 0.08
Quad-task	**Proposed Model (MultitaskBERTModel)**	**97.93 ± 1.10**	**97.95 ± 0.80**

Source: (Singh, Bhatia, & Saha, 2023).

Table 3 compares the proposed model (MultitaskBERT Model) performance to other models for the Complaint Severity (CS) task across different setups: single-task, two-task, three-task, and four-task scenarios. In the single-task setup, the proposed model clearly out- performs the existing model, achieving 84.73% accuracy and an 84.10% F1 score, compared to just 67.09% and 66.29% for the baseline.

In two-task scenarios, especially when paired with Complaint Identification (CI), the model performs exceptionally well, with accuracy reaching 91.29% and an F1 score of 91.30%. In the three-task setup, the model continues to deliver strong results, particularly in the CS+ER+CI combination, achieving 87.03% accuracy and an 86.32% F1 score. Finally, in the four-task setup, the proposed model significantly outperforms MTL-RoBERTa, achieving 84.49% accuracy and an 84.02% F1 score, showing how effective it is at handling multiple tasks at once.

Figure 4 contrasts the training and validation loss in epochs with the proposed model. Both losses reduce at the beginning, which means successful learning and adequate generalization. Training losses decrease monotonically, and there is no overfitting discovered at the beginning. The validation loss increases late, which is a good indication that the model treats unexpected examples fairly.

Table 3. Results of the proposed model (MultitaskBERTModel) and other models with respect to Accuracy and F1 score for Complaint Severity (CS) task

Multitasking Scenario	Model	Accuracy (%)	F1 Score (%)
Single task	STL_{Cs}	67.09 ±0.12	66.29 ±1.02
	ProposedModel$_{CS}$	**84.73 ±0.45**	**84.10 ± 0.50**
Dual-Task	MTL_{CS+CI}	70.08 ±1.04	68.55 ±1.03
	MTL_{CS+ER}	68.57 ±1.02	67.18 ±1.06
	MTL_{CS+SA}	68.38 ±1.11	67.07 ±1.13
	ProposedModel$_{CS+CI}$	**91.29 ±0.55**	**91.30 ±0.45**
	ProposedModel$_{CS+ER}$	**73.31 ±0.60**	**72.56 ±0.75**
	ProposedModel$_{CS+SA}$	**80.27 ±0.40**	**80.05 ±0.65**
Tri-task	$MTL_{CS+ER+SA}$	70.14 ±1.05	68.74 ±1.03
	$MTL_{CS+ER+CI}$	73.78 ±1.07	72.89 ±1.08
	$MTL_{CS+SA+CI}$	72.49 ±1.12	70.19±1.04
	ProposedModel$_{CS+ER+SA}$	**74.38 ±0.50**	**73.49±0.55**
	ProposedModel$_{CS+ER+CI}$	**81.25±0.40**	**80.03±0.50**
	ProposedModel$_{CS+SA+CI}$	**87.03 ±0.55**	**86.32 ±0.65**
Quad-task	MTL-RoBERTa	75.44 ±0.11	75.43 ± 0.08
	ProposedModel$_{CI+CS+ER+SA}$	**84.49 ± 1.25**	**84.02 ± 0.95**

Source: (Singh, Bhatia, & Saha, 2023).

Figure 5 shows the accuracy across four tasks during the quad-task model across seven epochs: Complaint Label, Severity, Emotion, and Sentiment. All tasks get more accurate with training. The Label task achieves almost perfect accuracy whereas the Emotion begins low but gradually climbs up. Sentiment and Severity also improve steadily.

Figure 6 shows the F1 scores over seven epochs for the quad tasks: Label, Severity, Emotion, and Sentiment in the proposed model. The F1 score for the Label task is consistently the highest, nearing 100%, indicating strong performance. The Severity and Sentiment tasks also demonstrate stable improvements, approaching high F1 scores as training progresses. Overall, the F1 scores improve consistently, reflecting the model's increasing capability in all tasks over time.

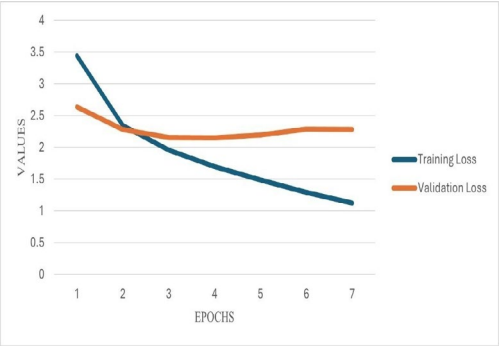

Figure 4. Training vs Validation plot of proposed model
Source: Authors' own creation.

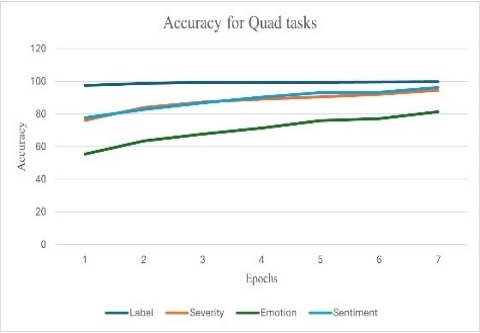

Figure 5. Accuracy of proposed model
Source: Authors' own creation.

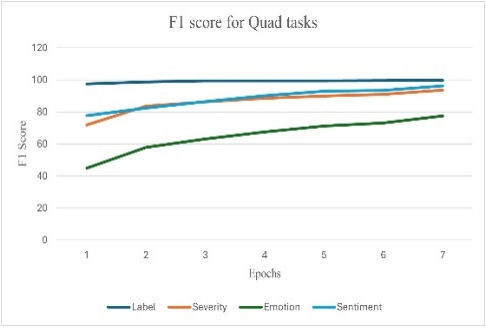

Figure 6. f1 score of proposed model
Source: Authors' own creation.

5. Conclusions

A key challenge in natural language processing (NLP) is classifying complaints, especially when trying to determine their severity. Human conversations are often complex, with a single sentence carrying multiple layers of meaning. Because of this situation, standard ways of finding complaints are no longer effective. To address this, we propose a multitasking model designed to handle this problem. Our model was tested on a financial dataset of tweets from bank customers, previously used in recent research work, and was tasked with classifying complaint labels, severity, emotion, and sentiment for each tweet.

We got very promising results, as the model reached 97.93% accuracy and 97.95% F1 score on complaint identification, surpassing the state of the art. It also obtained 84.49% accuracy and 84.02% F1 score on the more complicated severity classification. These results demonstrate the promise of our multitasking framework in finance customer complaint processing. The methodology is applicable to other fields such as railways, education and e-commerce where various insights are required based on user feedback.

References

[1] Bensoltane, R., Zaki, T. (2024), *Neural multi-task learning for end-to end Arabic aspect-based sentiment analysis. Computer Speech & Language*, 1-13, https://doi.org/10.1016/j.csl.2024.101683.

[2] Caruana, R. (1997), *Multitask Learning. Machine Learning,* 28, 41-75, DOI:10.1023/A:1007379606734

[3] Casola, S., Lauriola, I., Lavelli, A. (2022), *Pre-trained transformers: an empirical comparison. MachineLearningwithApplications*, 9, https://doi.org/10.1016/j.mlwa. 2022.100334.

[4] Cesario, E., Comito, C., Zumpano, E. (2024), *A survey of the recent trends in deep learning for literature based discovery in the biomedical domain. Neurocomputing,* 568, https://doi.org/10.1016/j.neucom.2023.127079.

[5] Coussement, K., den Poel, D. (2008), *Improving customer complaint management by automatic email classification using linguistic style features as predictors.DecisionSupportSystems*, 44(4), 870-882, retrieved from https://www.infona.pl/resource/bwmeta1.element.elsevier-468e675d-9418-39ed-b73d-e98c42ce8903.

[6] Dong, K., Romanov, I., McLellan, C., Esen, A. (2022), *Recent text-based research and applications in railways: A critical review and future trends. Engineering Applications of Artificial Intelligence.* ttps://doi.org/10.1016/j.engappai.2022.105435.

[7] Dwivedi, Y., Ismagilova, E., Hughes, D., Carlson, J., Filieri, R. (2021), *Setting the future of digital and social media marketing research: Perspectives and research propositions. International Journalof Information Management,* 59, https://doi.org/10.1016/j.ijinfomgt.2020.102168.

[8] Fadel, A., Saleh, M., Salama, R., Abulnaja, O. (2024), *MTL-AraBERT: An Enhanced Multi-Task Learning Model for Arabic Aspect-Based Sentiment Analysis*. *Computers,* 13(4), https://doi.org/10.3390/computers13040098.

[9] Ghosh, S., Priyankar, A., Ekbal, A., Bhattacharyya, P. (2023), *Multitasking of sentiment detection and emotion recognition in code-mixed Hinglish data*. *Knowledge-Based Systems,* 260(C), https://doi.org/10.1016/j.knosys.2022.11018.

[10] Gong, X., Ying, W., Zhong, S., Gong, S. (2022), *Text Sentiment Analysis Basedon Transformer and Augmentation. Frontiers in Psychology,* 13, DOI:10.3389/fpsyg.2022.906061.

[11] Goyal, P., Verma, D.K., Shishir, K. (2025), *Detection and monitoring for enhanced prevention of grain plant disease using classification-based deep ensemble neural networks in smart agriculture*. *International Journal of Remote Sensing,* 46(5), 1992-2022, DOI: 10.1080/01431161.2024.2443618.

[12] Gupta, M., Singh, A., Jain, R., Saxena, A., Ahmed, S. (2021), *Multi-class railway complaints categorization using Neural Networks: RailNeural*. *Journal of Rail Transport Planning & Management*, 1-14, https://doi.org/10.1016/j.jrtpm. 2021.100265.

[13] Hake, P., Rehse, J.-R., Fettke, P. (2021), T*oward Automated Support of Complaint Handling Processes: An Application in the Medical Technology Industry*. *Journal on Data Semantics,* 10, 41-56, retrieved from https://link.springer.com/article/10.1007/ s13740-021-00124-z.

[14] He, M., Wang, X., Zou, C., Dai, B., Jin, L. (2021), *A Commodity Classification Framework Based on Machine Learning for Analysis of Trade Declaration*. *Symmetry,* 13(6), https://doi.org/10.3390/sym13060964.

[15] Ilias, L., Mouzakitis, S., Askounis, D. (2024), *Calibration of Transformer-Based Models for Identifying Stress and Depression in Social Media*. *IEEE transactions on computational social systems,* 11(2), DOI:10.1109/TCSS.2023.3283009.

[16] Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N. (2023), *A comprehensive survey on applications of transformers for deep learning tasks. Expert Systems With Applications*, 1-48, https://doi.org/10.1016/j.eswa.2023.122666.

[17] Jin, M., Aletras, N. (2021), *Modeling the Severity of Complaints in Social Media. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* NAACL: Association for Computational Linguistics, DOI:10.18653/v1/2021.naacl-main.180.

[18] Joshi, A.K., Kumar, A., Patel, N. (2025), *Investigating the dissemination of deepfake conversations on social media via tweeting.* In *Communications in computer and information science* (pp. 107-117), https://doi.org/10.1007/978-3-031-87511-3_8.

[19] Oyewola, D., Omotehinwa, T., Dada, E. (2023), *Consumer complaints of consumer financial protection bureau via two-stage residual one-dimensional convolutionalneural network (TSR1DCNN)*. *Data and Information Management, 7*(4), https://doi.org/10.1016/j.dim.2023.100046.

[20] Sandiwarno, S., Niu, Z., Nyamawe, A. (2024), *SES-Net: A Novel Multi-Task Deep Neural Network Model for Analyzing E-learning Users' Satisfaction via Sentiment, Emotion, and Semantic. International Journal of Human–Computer Interaction*, https://doi.org/10.1080/10447318.2024.2356356.

[21] Singh, A., Bhatia, R., Saha, S. (2023), *Complaint and Severity Identification From Online Financial Content. IEEE transactions on computational social systems*, 1-11, DOI:10.1109/TCSS.2022.3215528.

[22] Tan, Y., Chow, C.-O., Kanesan, J., Chuah, J., Lim, Y. (2023), *Sentiment Analysis and Sarcasm Detection using Deep Multi-Task Learning.* 129(3), 2213-2237, DOI:10.1007/s11277-023-10235-4.

[23] Vairetti, C., Aránguiz, I., Maldonado, S., Karmy, J.P., Leal, A. (2004), *Analytics-driven complaint prioritisation via deep learning and multicriteria decision-making. European Journal of Operational Research,* 312(3), 1108-1118, DOI:10.1016/j.ejor.2023.08.027.

[24] Wang, X., Kim, B.-G., Parameshachari, B., Li, Q. (2021), *Exploring Multimodal Multiscale Features for Sentiment Analysis Using Fuzzy-Deep Neural Network Learning. IEEE Transactions on Fuzzy System*, 1-15, DOI: 10.1109/TFUZZ.2024.3419140.

[25] Zhao, G., Luo, Y., Chen, Q., Qian, X. (2023), *Aspect-based sentiment analysis via multitask learning for online reviews. Knowledge-Based Systems*, 1-12, https://doi.org/10.1016/j.knosys.2023.110326.