Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 02 (2026)

# An NLP-Based Approach for Automated Task Identification in Unstructured Chat Conversations

**Divisha Upadhyay, Megha Kuliha**
*Department of Information Technology, Shri Govindram Seksaria Institute of Technology and Science, Indore, Madhya Pradesh, India*

***Abstract:-*** The task of identifying relevant actions from short workplace messages is difficult because of the lack of contextual information, use of informal language, and code-mixing. Although transformer models have demonstrated remarkable success in minimal-context settings, their performance in such settings has not been adequately explored. This paper presents the application of binary task identification in a single-utterance classification paradigm, where messages are assessed independently without any conversational context. Three different methods are employed for the task: a traditional lexical approach using TF-IDF with logistic regression, a multilingual transformer model with XLM-RoBERTa-base, and an optimized transformer model. Empirical evaluation on 1,190 real-world workplace messages with class imbalance and Hinglish code-mixing demonstrates that lexical approaches are robust baselines, and optimized transformer models perform best. The results show that increased model complexity does not necessarily translate to improved performance and the need for empirical analysis in designing task identification systems for minimal-context workplace communication.

***Keywords****: Task Identification, Short Text Classification, Action Item Detection, TF-IDFLogistic Regression, XLM-RoBERTa, Transformer Models, Minimal-Context Classification, Multilingual NLP, Workplace Communication*

## 1. Introduction

The modern workplace communication system is characterized by the use of short digital messages that facilitate fast coordination but at the same time pose a risk of missing important tasks, especially when the message is conveyed in an informal and concise manner. The automatic identification of actionable tasks from everyday communication using natural language processing techniques has consequently attracted significant research attention. However, the majority of existing studies on task or action-item detection focus on text containing sufficient contextual information, such as meeting minutes or structured business communications [1]. Recent research has demonstrated that transformer-based models effectively identify action items in context-rich environments by leveraging longer utterances and surrounding discourse information [1]. Nevertheless, short workplace messages differ substantially in nature, often consisting of only a few tokens and appearing in isolation, thereby providing very limited linguistic cues. This lack of context complicates task detection, as many features employed in previous approaches do not adapt well to such settings. At the same time, the state-of-the-art in short-text intent classification has shown that transformer models are capable of picking up on subtle linguistic variations within individual sentences [2]. For example, XLM-RoBERTa, a multilingual transformer model trained on large diverse datasets, has shown robustness to informal and noisy language typical of natural communication [3]. However, empirical studies directly contrasting classical lexical approaches with modern transformer-based models for detecting actionable content in short, isolated workplace messages remain limited. This gap partly stems from the prevailing perception that machine learning methods are difficult to deploy and interpret in real workplace scenarios. This paper addresses the gap by comparing the TF-IDF with the logistic regression baseline model to the XLM-RoBERTa-base model for binary classification tasks in a minimal context.

Our contributions are (1) a comparison of lexical and transformer models for task classification in short isolated messages, (2) evidence that classical models are still valid in a minimal context, and (3) insights for choosing models in task automation systems.

## 2. Literature Review

A.Task Extraction and Action Item Detection

Current research developments in task extraction from textual material concentrate on three specific areas: action item extraction, short-text intent analysis, and deep learning based sentence classification. The researchers established task extraction as a sentence level binary classification task that tested meeting transcript sentences and demonstrated that transformer models outperformed traditional machine learning methods when researchers provided adequate meeting context to the study participants. [1]. The study demonstrated that task extraction fulfills its function as a supervised learning problem that operates through structured conversational interfaces. Cohen and his colleagues demonstrated their research results through automated task item paraphrasing, which proved that even indirect and polite task statements could be extracted accurately. [2]. This research showed that task extraction requires researchers to study both imperative and semantic meanings of tasks. Golia and Kalita developed a task-driven summarization method, which they used to demonstrate that task relationships are essential for generating summaries of extensive meeting recordings that contain multiple spoken parts. [4]. Asthana et al. provided essential advancements to this field through their development of a system that automatically summarizes content and extracts important information and meeting tasks by utilizing large language models (LLMs) together with extensive meeting data. [5].

B.Short Text Classification

Research on short text understanding has recently recognized the challenges involved in determining the intent of messages in a non-contextual manner. Wu et al. developed the Quartet Logic: A Four-Step Reasoning (QLFR) system, which solves Short Text Classification (STC) challenges through its three-step process that first determines main concepts before evaluating knowledge questions and rewriting text for final assessment. [19]. The QLFR method shows that brief textual content struggles to understand both semantic and syntactic complexities that exist in standard pre-trained language models. The QLFR method demonstrates that structured reasoning can apply LLMs' inherent knowledge to solve STC problems, yet its effectiveness in identifying tasks during brief work contexts remains untested. Larson et al. established a baseline for message intent classification and proved that short messages are difficult to understand because they lack linguistic cues, which created difficulties for word-level feature analysis in their research. [6]. Zhang and Shafiq answered the Larson et al. challenge through their development of an advanced text classification system, which enables users to analyze messages because it can handle difficult statements through its superior sentence analysis capabilities while it handles context information limitations [7]. Shu et al. conducted an experiment to test weakly supervised learning methods for email message classification, and their results showed that message intent determination works under conditions of restricted annotation data, which enables its application in message interpretation tasks. [8].

C.Transformer-Based Approaches

researchers conducted their study by using surveys to evaluate the current advancements in text classification methods. Minaee et al. covered the deep learning based text classification approaches and concluded that the transformer-based approaches are more effective as superior approaches, although with mixed results based on applications [9]. On the other hand, Cunha et al. compared classical machine learning-based approaches, neural networks, and transformers, concluding that although transformers are superior in performance, TF-IDF with linear classifiers still has use in structured text classification to some extent [10]. Zhang and Shafiq reviewed transformers and ensemble learning for NLP tasks, covering how architectures have a significant role in the accuracy of text classification, thus requiring comparisons based on empirical experiments among various architectures for accurate results [7]. Said and Ismail reviewed the current trends in text classification and concluded that, although transformers are greatly preferred, simple approaches are also equally effective in low context scenarios [12]. Taha et al. covered text classification in an overall survey study and concluded that there

are no superior or inferior text classification approaches overall, which are completely different based on data types [13]. Recent emphasis has been on short text classification in low-context settings. Zhu et al. introduced soft prompt-tuning techniques designed particularly for short text classification and observed improvements in performance without modifying model architectures [14]. Hong et al. investigated intent classification using natural language description in the zero-shot paradigm, achieving impressive cross-domain transfer using pretrained language models [15]. Li et al. introduced a hybrid BERT architecture with attention and sequence modules tailored for short text classification problems [16], while Li et al. studied word representation enhancement techniques to solve the problem of missing contextual word representation [17]. Yang et al. proposed a dual stream transformer model specifically designed for short text classification of sentiment [18]. In addition to accuracy enhancement, recent studies have focused on the aspects of robustness, interpretability, and domain adaptation. Mustafa and Hama Saeed discussed the ideas associated with explainable AI in text classification tasks, in which the role of explainability in implementation was emphasized [20]. Rostam and Kertesz conducted a systematic review on domain-specific pretrained language models, proving the effect of domain adaptation on text classification accuracy [21]. Chen et al. proposed knowledge-aided attention models to improve semantic understanding using external knowledge [22]. Han et al. explored few-shot text classification models via transformers and verified the effectiveness of good pretraining for promoting generalization performance even when available labeled data are limited [23].

D.Research Gap

Despite these advances, the state-of-the-art literature is mainly populated by methods that rely on access to conversational context, target general or generic intent categories, or assess individual models rather than directly comparing different approaches. Therefore, very little existing work affords a systematic comparison between traditional lexically based approaches and state-of-the-art transformer-based models on binary task identification within short, independent workplace communications with very low contextual information and potentially implicit tasks. Furthermore, despite multilingual transformer models such as XLM-RoBERTa proving robust and effective against noisy and informal text corpora [3], their utility on realworld chat communication with minimal contextual information remains an underexplored area. This calls forth the current investigation and comparison between TF-IDF with logistic regression and XLM-RoBERTa-base on task identification within minimal-context settings. The QLFR framework's emphasis on the challenges of short text classification—particularly the difficulty in grasping semantic and syntactic intricacies with limited context—provides theoretical motivation for our empirical investigation of whether increased model complexity (transformers) necessarily outperforms simpler lexical approaches (TF-IDF + LR) in minimal-context task detection scenarios.

## 3. Methods

A.Problem Formulation

This work treats the task detection problem as a binary text classification problem in the context of workplace messaging. Given a short message $m$ , the goal is to classify the message as containing a task request or not. This approach contrasts with traditional conversational methods that rely on the flow of messages. It treats each incoming message as an independent instance under a Single Utterance Classification (SUC) framework. Formally, let $m$ be an individual workplace message and y $\in$ {0, 1} be the corresponding class label for the message. Here, $y$=0 & 1 represent messages containing and not containing a task request, respectively. The goal here is to learn a classification function $f$ that maps the input message to a probability distribution over the two classes:

$$f{:}m{\rightarrow}P(y|m)$$

The classifier should be able to make the correct classification using only the information in the current message, without access to future messages or extensive conversation histories. This is because, in real-world notification-based scenarios, users need to be able to recognize messages quickly, based only on the limited information provided in the message preview. Hence, the problem is to develop a model that is capable of classifying short workplace messages as belonging to tasks and non tasks under limited context.

B.System Architecture

The design of the system architecture uses two pipelines to compare lexical approaches with transformer approaches. Figure 1 below shows the system architecture of the entire system. This shows the flow from raw input to preprocessing and then to the two pipelines running in parallel.

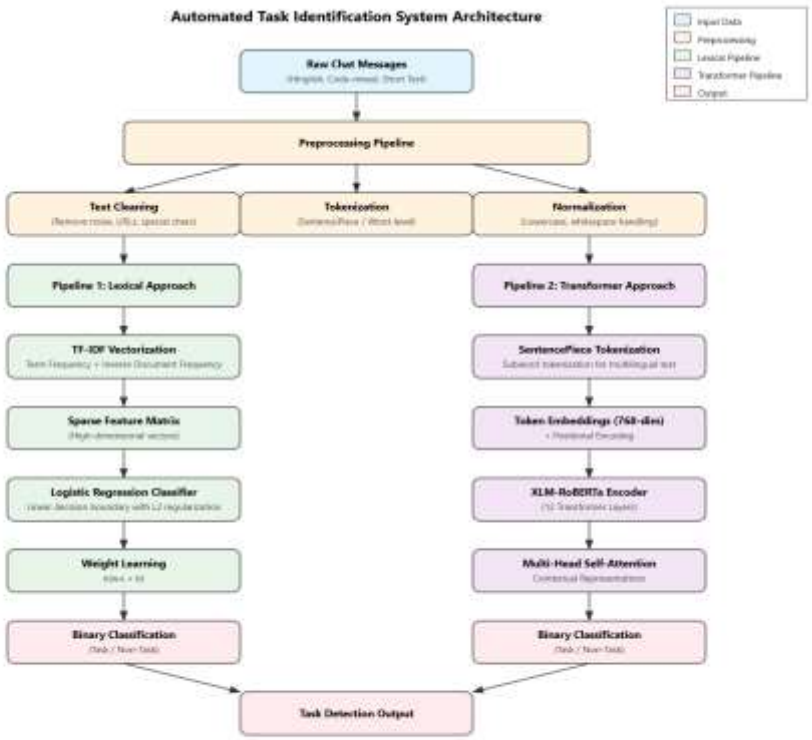


Fig. 1. System architecture showing dual classification pipelines operating on pre-processed workplace messages.

C.Dataset

The dataset comprises 1,190 labeled messages that were gathered from real work settings. The dataset has three key features that make it capable of illustrating real-life scenarios requiring only basic context knowledge.

| Characteristic | Description | Count/Percentage |
|---|---|---|
| Total Messages | Complete dataset size | 1,190 |
| Task Messages (y=1) | Messages containing actionable tasks | 238 (20%) |

| | | |
|---|---|---|
| Non-Task Messages (y=0) | Informational/conversational messages | 952 (80%) |
| Average Message Length | Mean number of tokens per message | 8.3 tokens |
| Hinglish Messages | Code-mixed Hindi-English messages | ~35% |
| Pure English Messages | English-only messages | ~60% |
| Pure Hindi Messages | Hindi-only messages | ~5% |

Table 1. Dataset characteristics and distribution

The dataset shows a severe class imbalance because direct task instructions (is_task=1) occur less frequently than updates and discussions. (is_task=0). The typical workplace chat environment shows an imbalance ratio of 1:4 because most messages contain informational content, which exceeds the amount of directive content.

D.Preprocessing

The preprocessing pipeline was designed to remove background noise while it preserved essential evidence needed for intent detection.

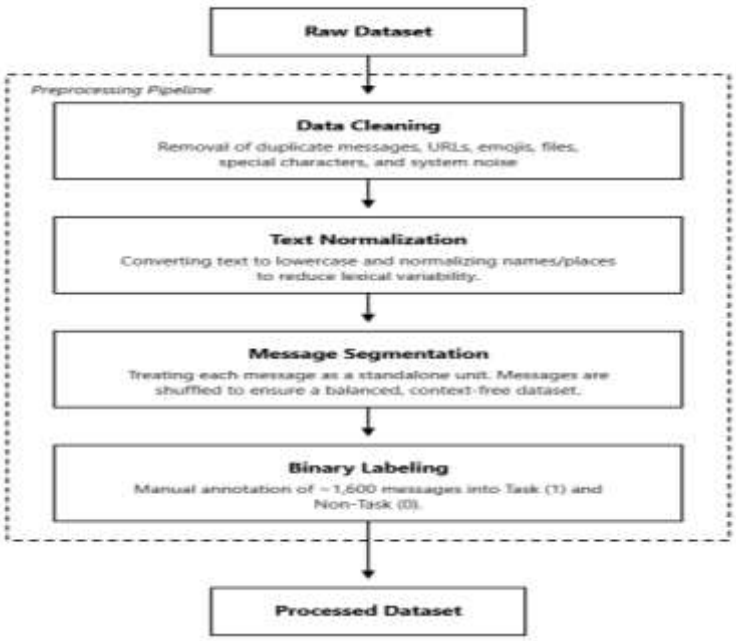

Fig. 2 shows four stages of preprocessing pipeline

The pipeline has four major stages:

Stage 1: Data Cleaning

- The process eliminates duplicate messages in the dataset.
- The process entails the removal of URLs, emojis, and file attachments.
- Special character normalization eliminates system noise and platform-specific formatting characters.
- The system normalizes the treatment of whitespace and irrelevant metadata.

Stage 2: Text Normalization

- The process entails the conversion of all text to lowercase to eliminate lexical variability.
- Names and geographic locations are normalized to ensure consistency in the dataset.
- The process eliminates the effects of surface-level variability.

Stage 3: Message Segmentation

- Each message is considered a standalone, context-free unit.
- Messages are rearranged to ensure a balanced distribution in the dataset. The process eliminates the effects of ordering biasin the training of the model.

Stage 4: Binary Labeling

- Approximately 1,600 messages are labelled manually.
- Each message is labeled with one of two labels:
- Task (1) or Non-Task (0).
- The labeled dataset is used for supervised classification.
- The preprocessing method preserves essential linguistic characteristics, which help determine the intended task of the text because it focuses on maintaining these features.
- The system identifies text functions through four linguistic elements, which include imperative verbs, temporal expressions, and responsibility markers.

E.Algorithm 1: TF-IDF + Logistic Regression

The first pipeline implements TF-IDF together with logistic regression for its operational framework. The lexical baseline pipeline uses Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction and logistic regression for classification.

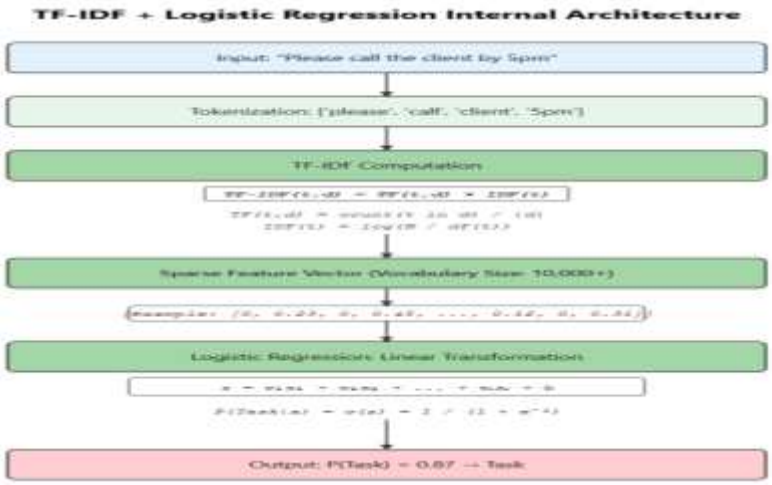


Figure 3: Internal architectures of TF-IDF + logistic regression with sparse feature extraction and linear decision boundary.

**TF-IDF Feature Extraction :**

The text data gets converted into a sparse vector representation through TF-IDF, which uses multiple dimensions that correspond to every term in the vocabulary. The technique consists of two steps:

Term Frequency (TF): The term frequency for a term t in document d gets calculated through this process

$$TF(t, d) = count(t \text{ in } d) / |d|$$

The formula defines |d| as the complete termcount present in document d.

Inverse Document Frequency (IDF): This step computes the rarity or commonness of a term t in the whole corpus.

$$IDF(t) = \log(N / df(t))$$

The formula calculates N as the total document count, while df(t) represents the number of documents which contain the term t.

Combined TF-IDF Score:

$$TF\text{-}IDF(t, d) = TF(t, d) \times IDF(t)$$

The scoring function gives higher scores to terms that appear more often in a document while appearing less frequently throughout the entire corpus.

Example: For the message "Please call the client by 5pm."

Tokenization: ["please", "call", "client", "5pm"

Each term has TF = 1/4 = 0.25 (assuming each term appears once) TF computation.

IDF Computation:

"please" may have an IDF of 2.3 (in 10% of the docs)

"call" may have an IDF of 3.1 (in 5% of the docs)

"client" may have IDF = 2.8 (in 7% of the docs)

"5 pm" may have an IDF of 3.5 (in 3% of the docs)

The TF-IDF vector contains elements that represent 0.575, 0.775, 0.70, and 0.875 as its elements.

The TF-IDF vector becomes sparse because most of its elements contain zero values, while it maintains high dimensionality because its vocabulary range extends between 5,000 and 10,000 terms.

**Logistic Regression Classification**:

The logistic regression classifier uses a linear function to transform the TF-IDF feature vector and applies the sigmoid activation function to compute class probabilities. The linear equation

$$z = w_1x_1 + w_2x_2 + ... + w_nx_n + b$$

describes a linear relationship that uses trainingderived weight values for w and the TF-IDF feature vector x and the bias value b.

The sigmoid activation function

$$P(\text{Task} \mid x) = \sigma(z) = 1 / (1 + e^{-z})$$

transforms the linear combination z into a probability result that exists between the values of 0 and 1.

Training Objective: The model achieves its training objective through two loss functions, which include binary cross-entropy loss and L2 regularization:

$$\text{Loss} = -[y \log(p) + (1 - y) \log(1 - p)] + \lambda||w||^2$$

The equation uses y as the actual class label, while p represents the predicted probability, and λ functions as the regularization factor, which prevents overfitting through weight penalties.

Decision Rule: The system classifies messages as tasks when their probability P(Task | x) reaches 0.5, which organizations can modify to achieve their desired precision and recall needs.

Algorithm :

INPUT: labeled text documents

OUTPUT: predicted class label

STEP 1: CLEAN TEXT

lowercase → remove punctuation

remove stopwords → lemmatize

STEP 2: TF-IDF VECTORIZE

TF(t,d) = term count / doc length

IDF(t) = log(N / df(t))

TF-IDF(t,d) = TF × IDF

normalize each doc vector → L2

STEP 3 : TRAIN LOGISTIC REGRESSION

initialize weights W and bias b = 0

Repeat until convergence:

z = X·W + b

ŷ = sigmoid(z) ← or softmax if multiclass

loss = CrossEntropy(y, ŷ) + λ||W||²

W -= α · ∂loss/∂W

b -= α · ∂loss/∂b

STEP 4: PREDICT

clean → tfidf_transform → z = x·W + b

return argmax(softmax(z))

F.Algorithm 2: XLM-RoBERTa

XLM-RoBERTa is a multilingual transformer based model that generates dense contextual embeddings through multi-head self-attention mechanisms.

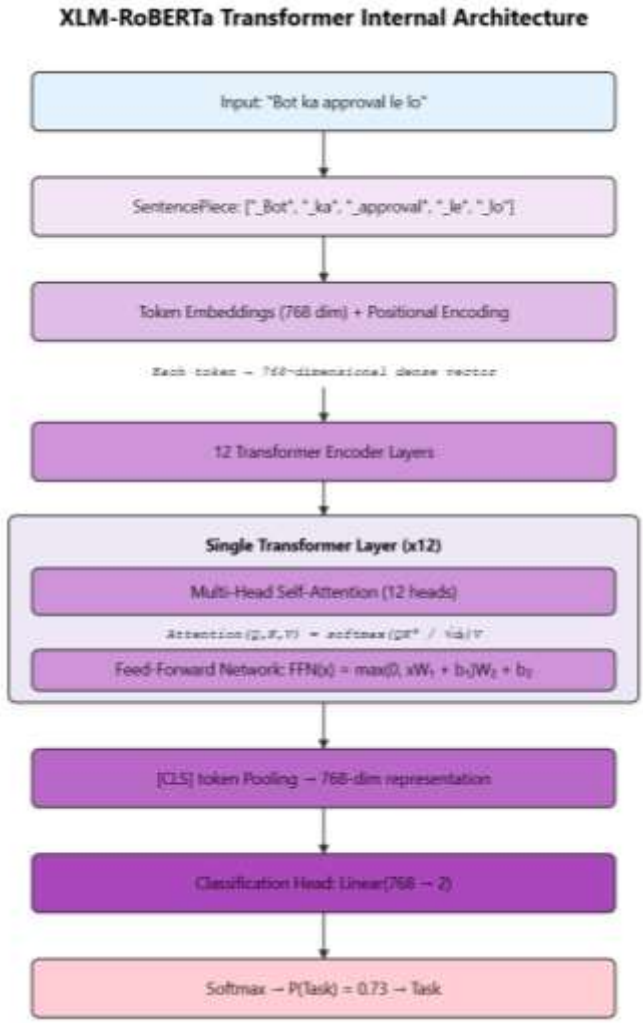


Figure 4: Internal architectures of the XLM-RoBERTa transformer with multi-head self-attention and dense contextual representations.

SentencePiece Tokenization

Instead of tokenizing text word by word, XLMRoBERTa does it subword by subword so unknown words, multilingual phrases, and Hinglish are all dealt with effortlessly.

`"Bot ka approval le lo" → ["_Bot", "_ka","_approval", "_le", "_lo"]`

The small is just an indicator of where a new word begins.

Token and Positional Embeddings

Each subword token has a detailed 768-dimensional personality—it's not just a number, but a vector full of meaning. And position is embedded too, so the model never forgets what came first.

$$E = \text{TokenEmbedding(token)} + \text{PositionalEncoding(position)}$$

Unlike TF-IDF's sparse vectors, each and every one of those 768 dimensions is alive and buzzing with meaning.

Transformer Encoder Layers

The architecture of XLM-RoBERTa consists of 12 transformer encoder layers. Each of the 12 layers

performs two tasks:

Self-Attention: where every word looks at everyother word and asks, "How much do you matter to what I mean?" Q asks the question, K has the answer,and V has the payload:

$$\text{Attention} = \text{softmax}(Q \cdot K^T / \sqrt{dk}) \cdot V$$

12 heads do this in parallel: one might focus on grammar rules, another might pick up on time related phrases.

Feed-Forward Network: where each token is then individually optimized:

$$FFN(x) = \max(0, xW_1+b_1)W_2+b_2$$

Classification + Softmax

After 12 rounds, the [CLS] token has quietly taken in the meaning of the entire sentence. It goes straight to the classifier:

$$\text{logits} = W\_class \cdot h[CLS] + b\_class$$

$$P(\text{Task} \mid x) = \exp(\text{logit\_task}) / \Sigma \exp(\text{all logits})$$

All parameters, including the embedding, the attention weights, and the classifier, are learned jointly until the model gets it right.

Algorithm

INPUT: raw text

OUTPUT: predicted class label

STEP 1: TOKENIZE

text → [CLS] + subwords + [SEP]

convert to token IDs + attention mask

STEP 2: EMBEDDING

x = TokenEmbedding(ids) +

PositionEmbedding()

x = LayerNorm(x)

STEP 3: TRANSFORMER ENCODER (repeat× 12)

Q, K, V = x · Wq, x · Wk, x · Wv

attn = softmax(Q·K^T / √dk) · V

x = LayerNorm(x + attn)

x = LayerNorm(x + FFN(x))

STEP 4: CLASSIFICATION

cls = x[:, 0, :]

logits = Dropout(cls) · W + b

ŷ = softmax(logits)

STEP 5: FINE-TUNE

loss = CrossEntropy(ŷ, y) + λ||W||²

W -= α · ∂loss/∂W

STEP 6: PREDICT

return argmax(softmax(logits))

G.Evaluation Framework

There exists a significant class imbalance problem in work-related communication. The accuracy metric does not function as an effective performance assessment method. Our evaluation process, which we use to measure performance, includes the following assessment metrics:

Precision: The proportion of predicted tasks that are actual tasks:

$$\text{Precision} = TP / (TP + FP)$$

The system achieves high-precision results because it generates fewer false alerts, which helps decrease notification fatigue.

Recall: The proportion of actual tasks that are correctly identified:

$$\text{Recall} = TP / (TP + FN)$$

The system achieves high recall results because it detects all tasks, which prevents critical tasks from being missed.

F1-Score: The harmonic mean of precision and recall:

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

The F1 score provides a balanced measure that shows equal importance to both major and minor task categories.

## 4. Results

This section will display the updated experimental results for the comparison of the performance of TF-IDF + Logistic Regression, XLM-RoBERTa-base, and the optimized XLM-RoBERTa-base model on the task classification dataset. The performance assessment of all models uses standard classification metrics which emphasize the evaluation of minority task class data that exists when is_task equals 1

A.Confusion Matrix Analysis

The analysis of confusion matrices for all three models enabled a more precise evaluation of their classification performance. The TF-IDF + Logistic Regression model was able to correctly classify 164 non-task and 39 task messages, with 24 and 11 misclassifications, respectively, indicating a moderate level of accuracy with a slight bias toward the majority class. The baseline XLM-RoBERTa model was able to correctly classify more non-task messages (174 correct classifications) but was less accurate on task messages, misclassifying 16 task messages as non-task. The model needs domain-specific training to accurately classify informal task-oriented messages. The XLM-RoBERTa-base (Optimized) model was able to correctly classify 178 non-task and 190 task messages with negligible errors (10 and 11, respectively), indicating balanced accuracy on both classes.

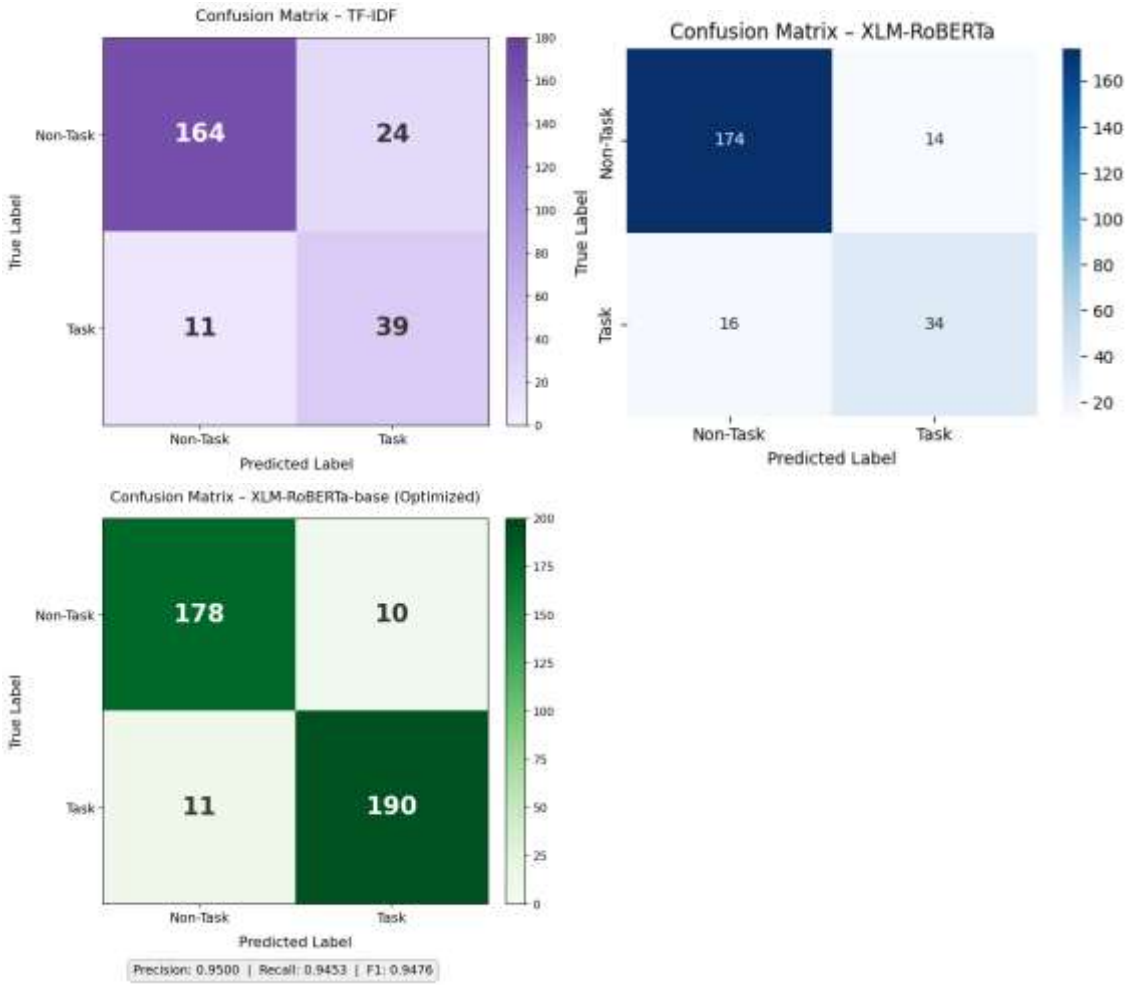


Fig. 5: the confusion matrices

The analysis of the confusion matrices above confirms the progressive improvement from one model to the next and validates the observations highlighted in the previous comparison section.

B.Overall Performance Comparison

| Model | Precision | Recall | F1-Score |
|---|---|---|---|
| TF-IDF + Logistic Regression | 0.77 | 0.82 | 0.79 |
| XLM-RoBERTa-base | 0.70 | 0.68 | 0.69 |
| XLM-RoBERTa-base (Optimized) | 0.95 | 0.94 | 0.94 |

Table 4: The performance of all models on the task class (is_task=1) is summarized.

The results demonstrate that all methods show distinct performance results. The TF-IDF and logistic regression algorithms create a strong system that achieves 82 percent recall and 77 percent precision. The lexical baseline proves its efficacy because it identifies most of the actionable messages while maintaining an acceptable number of false positives. The F1-score value of 0.79 demonstrates that traditional feature-driven approaches maintain their ability to identify minimal context tasks efficiently.

XLM-RoBERTa-base (Standard Configuration) delivers its typical performance through its accuracy metrics, which show 0.70 precision, 0.68 recall, and 0.69 F1 score. The current base transformer model configuration provides performance that is lower than the lexical baseline because the system requires task-level optimization to derive maximum benefit from the dataset using contextual understanding.

XLM-RoBERTa-base (optimized)

The optimized XLM-RoBERTa-base model provides a dramatic boost in performance because of its precision score of 0.95 and recall score of 0.94 which results in an F1 score of 0.94. The system configuration reduces both false positives and false negatives because transformer models provide better task detection performance after effective tuning and calibration techniques.

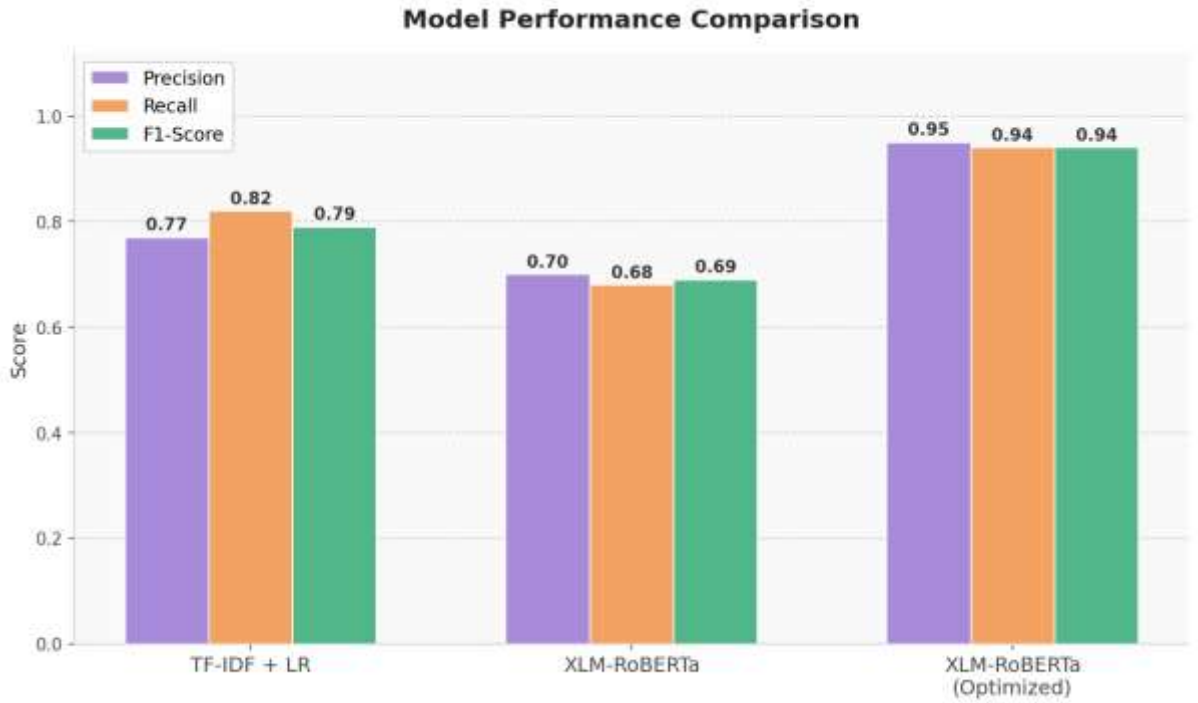


Figure 6: Model Performance Comparison for Task Class (is_task=1)

C.Key Findings

The optimized XLM-RoBERTa-base model performs better than both TF-IDF plus LR and the standard transformer model because it achieves higher scores in all three evaluation metrics of precision, recall, and F1-score.

The optimized transformer model outperforms all other models except for TF-IDF plus LR because it achieved an F1-score of 0.94. However, TF-IDF + LR still provides tough competition to the optimized model because it obtains an F1-score of 0.79. This outcome proves that the lexical models still provide effective results as reliable approaches to detect minimal context tasks.

The large difference between the standard model (F1 = 0.69) and the optimized model (F1 = 0.94) for XLM-RoBERTa proves that hyperparameter optimization and training and threshold tuning act as critical components of model development.

The transformer models can provide their best results when these models are provided appropriate optimization because these models can leverage contextual semantics to analyze the content of short texts that previous models could not analyze.

The optimized model now provides high precision and high recall because the trade-off between precision and recall has been eliminated to a great extent.

## 5. Discussion

The experimental results demonstrate significant insights into the behavior of the models in minimal context conditions. Instead of supporting the widely held assumption that increased model complexity is equivalent to increased performance, the experimental results demonstrate that optimization is the key factor.

A.Why TF-IDF + LR Works

The TF-IDF + Logistic Regression model proved to be a surprisingly strong baseline, with an F1-score of 0.79. This is because the messages in workplace tasks follow certain linguistic conventions on the surface level, such as the use of imperative verbs, temporal words, and urgency words. These words are magnified by the TF-IDF scores, and the stability of the logistic regression model helps it perform well, especially with an unbalanced dataset. The simplicity of the model also gives it certain advantages, such as faster running times and less memory usage.

B.Why the Standard Transformer Underperforms

The standard XLM-RoBERTa-base model is architecturally advanced, but it still produced the worst results in terms of performance, with an F1-score equal to 0.69. This is a well-known problem that transformer models, which have been pre-trained on large general-purpose corpora, cannot be used in narrow-domain tasks without proper fine-tuning [9]. In a low-context scenario, messages are brief and have limited discourse context in which to operate, and therefore the advantage of self-attention is diminished in favor of a more lexical-based approach.

C.What Makes the Optimised Transformer Work

The lack of syntactic structures and discourse features prevents accurate classification of short texts, which will remain an unresolved issue. The optimized transformer model achieved a high F1 score of 0.94 because it succeeded in processing minimal context, which led to peak performance results. Successful performance in this particular scenario requires the following three elements to be present. The dataset needs to maintain its properties while the domain must show consistent characteristics. The organization requires its members to follow specific communication patterns. The system needs to establish optimal methods for fine-tuning processes. The system requires the development of consistent patterns that will determine how different classes will be distributed.

D.Model Selection in Practice

The results suggest that model selection should be guided by the operational requirements of the deployment environment. The TF-IDF-logistic regression approach is appropriate for situations where speed of deployment is important or where computational resources are scarce. The optimized transformer model is appropriate in situations where classification accuracy is important and computational resources are not a concern. In multilingual or code-mixed communication situations, XLM-RoBERTa benefits from an additional advantage that is not available in lexical models

## 6. Conclusion

This paper investigated automated task identification from short, informal workplace messages under minimal-context conditions. The comparative analysis has shown the effectiveness of classical lexical-based approaches as viable and efficient solutions, while the application of transformer-based approaches, with appropriate optimization, has shown promising task detection capability. One of the important findings of the current research is that the complexity of the model does not guarantee the effectiveness of the task detection capability. Another important finding is the confirmation of the single utterance classification as a viable approach for task detection in the context of the notification-driven environment. The findings are important for developing task detection capability with appropriate application for informal and multilingual communication in the workplace environment.

Future research could focus on creating hybrid models that combine lexical feature representation with transformer-based models to generate contextual embeddings and improve task detection through conversational context utilization.

## 7. Limitations

This study has some limitations. The dataset of 1,190 messages, which were all collected within a single organizational environment, thus limits the generalizability of the study to environments with different communication norms and cultural conventions. Another limitation is that classification is limited to binary classification, where messages can only be classified as belonging to a particular type of task or not, thus limiting the practical use of such a system, as real-world task management is not limited to such classification.

## References

[1] Sachdeva K, Maynez, J. and Siohan, 0 2021, December: Action item detection in meetings using pretrained transformers. In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 861-868). IEEE

[2] Cohen, A. Kantor; A. Hilleli, S. and Kolman, E 2021, August. Automatic rephrasing of transcripts-based action items. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 2862-2873)

[3] Conneau; Khandelwal, K Goyal, N. Chaudhary, V: Wenzek; G,, Guzman, F, Grave, E Ott, M, Zettlemoyer; L. and Stoyanov; V 2020. July: Unsupervised cross-lingual representation learning at scale In Proceedings of the S8th annual meeting of the association for computational linguistics (pp. 8440-8451).

[4] Golia, L. and Kalita, 2023 December Action-item-driven summarization of long meeting transcripts. In Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval (pp. 91-98).

[5] Asthana, S,, Hilleli, S. He, P and Halfaker; A 2025. Summaries, Highlights, and Action items: Design; implementation and evaluation of an LLM-powered meeting recap system. Proceedings of the ACM on Human-Computer Interaction, 92), pp.1-29

[6] Larson, S,, Mahendran, A , Peper; JJJ,, Clarke, C Lee, A Hill, P, Kummerfeld, JK Leach K Laurenzano, MA. Tang; LJ and Mars, 2019. An evaluation dataset for intent classification and out-of-scope prediction: arXiv preprint arXiv:1909.02027.

[7] Zhang; H: and Shafiq; M.O. 2024. Survey of transformers and towards ensemble learning using transformers for natural language processing: Journal of big Data, 11(1), p.25 .

[8] Shu; K Mukherjee, S. Zheng; G. Awadallah A.H,, Shokouhi, M. and Dumais, S_ 2020, July: Learning with weak supervision for email intent detection: In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1051-1060)_

[9] Minaee, S. Kalchbrenner; N. Cambria, E Nikzad, N. Chenaghlu; M. and Gao, J,, 2021. Deep learning--based text classification: comprehensive revicw: ACM computing surveys (CSUR); 54(3) , pp.1-40

[10] Cunha, W. Viegas, F, Franca, C , Rosa, T Rocha, L. and Goncalves, MA: 2023 comparative survey of instance selection methods applied to non-neural and transformer-based text classification: ACM Computing Surveys, 55(13s) , pp.1-52.

[11] Said, A. and Ismail, A., 2025. Trends in natural language processing for text classification: A comprehensive survey. International Journal of Science and Research Archive, 14, pp.1540-1547.

[12] Taha, K Yoo, PD Yeun, C. Homouz, D. and Taha; A , 2024. comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. Computer Scicnce Review, 54, p.100664.

[13] Zhu; Y, Wang; Y, Mu; J,, Li, Y, Qiang; J,, Yuan, Y and Wu, X, 2024. Short text classification with soft knowledgeable prompt-tuning Expert Systems with Applications, 246, p.123248

[14] Hong, M,, Jiang; D. Song; Y and Zhang; CJ, 2024. Neural-Bayesian Program Learning for Few-shot Dialogue Intent Parsing: arXiv preprint arXiv:2410.06190

[15] Li, J,, Fang; J and Du, C 2025, March Research on Short Text Classification Based on BERT with Fusion of Key Features and Temporal Features In Proceedings of the 2025 4th International Conference on Cyber Security, Artificial Intelligence and the Digital Economy (pp. 546-551)

[16] Li, C. Xie, Z. and Wang; H,, 2025. Short Text Classification Based on Enhanced Word Embedding and Hybrid Neural Networks. Applied Sciences, 15(9), p.5102.

[17] Yang; S,, Xing; J,, Liu, Z_ and Sun; Y 2025_ Short-Text Sentiment Classification Model Based on BERT and Dual-Stream Transformer Gated Attention Mechanism_ Electronics, 14019),p.3904._

[18] Wu; H,, Zhang; Y, Han, Z Hou; Y Wang; L, Liu; S,, Gong, Q. and Ge Y 2024. Quartet logic: four-step reasoning (qlfr) framework for advancing short text classification: arXiv preprint arXiv:2401.03158.

[19] Mustafa, S. and Hama Saced, M 2025. Empowering text classification with NLP and explainable AI for enhanced interpretability: Journal of Electrical Systems and Information Technology; 12(1), p.81.

[20] Rostam, ZR and Kertesz, G 2025. Advances in Pre-trained Language Models for Domain-Specific Text Classification: Systematic Review: ACM Transactions on Intelligent Systems and Technology; 16(6), pp.1-41.

[21] Chen; J,, Hu; Y Liu, J , Xiao, Y. and Jiang; H,, 2019, July. Deep short text classification with knowledge powered attention. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 6252-6259).

[22] Han; X Sun Y Huang; W, Zheng, H. and Du; 2025. Towards Robust Few-Shot Text Classification Using Transformer Architectures and Dual Loss Strategies. arXiv preprint arXiv:2505.06145.