# Instagram Comment Spam Detection using ALBERT

CS4442


Ameena Malik
*Faculty of Engineering Sciences*
Western University
London, Ontario
amali62@uwo.ca

Talia Wilk
*Faculty of Science*
Western University
London, Ontario
twilk2@uwo.ca

Brendan Bain
*Faculty of Science*
Western University
London, Ontario
bbain7@uwo.ca



***Abstract*—Social media platforms like Instagram have become integral parts of modern communication. However, the rise of spam accounts and automated bots has posed a significant challenge to the user experience, with spam comments cluttering the comment sections of posts and potentially influencing user behavior. Traditional rule-based methods often fail to differentiate between spam and non-spam comments, necessitating the use of new and sophisticated algorithms that take into account context and natural language patterns. This project aims to develop and explore the results of a spam detection model capable of distinguishing between non-spam and spam Instagram comments using the pre-trained NLP model ALBERT. This research hopes to contribute to the development of effective spam mitigation strategies for social media platforms by investigating the effectiveness of these techniques and evaluating their performance against non-spam-world Instagram data.**



***Keywords—Spam Detection, Natural Language Processing (NLP), Instagram, ALBERT Model, Machine Learning, Comment Analysis, Text Classification, Algorithmic Filtering, Contextual Analysis, Deep Learning, Data Preprocessing, Model Evaluation***


## I. INTRODUCTION

Instagram's spam filters have historically relied upon rule-based methods in social media, which are progressively rendered ineffective by the advent of sophisticated spamming techniques. These methods need to be revised to combat the sophistication of contemporary spamming strategies. Despite its success, BERT's extensive computational and memory requirements have prompted research into more efficient models.

The ALBERT model was developed in response to these challenges, introducing critical design modifications to improve computational and memory efficiency without significantly compromising performance.

## II. RELATED WORK

The proliferation of social media has been marred by the rise of spam comments, challenging existing detection systems. Conventional rule-based spam detectors are inadequate for the intricate language patterns of spam, underscoring the need for advanced natural language processing (NLP) techniques [5]. The ALBERT model has emerged as an effective tool in this regard, displaying a significant improvement over traditional methods by utilizing innovative features such as cross-layer parameter sharing and the Sentence Order Prediction (SOP) task [2], [3].

Despite ALBERT's effectiveness, its application to the specific context of Instagram—a platform with rapidly evolving spam tactics and user interactions—remains underexplored [2], [3]. This study seeks to address this gap by fine-tuning NLP methods to Instagram's unique environment to enhance spam detection. ALBERT's efficiency is rooted in its architecture, optimized to reduce computational demand while maintaining performance. It starts with an embedding layer that condenses tokenized inputs, preparing them for the encoder's multiple transformer layers that share parameters to streamline complexity and improve learning. The architecture's multi-head attention mechanism captures diverse contexts, enriching its understanding of the input data. As detailed in Figure 1, ALBERT's encoder is followed by a Bi-LSTM layer that adds depth to temporal context comprehension—vital for differentiating spam from genuine content. ALBERT's pre-training, featuring MLM and SOP, further refines its predictive accuracy and textual coherence.

The softmax layer then classifies inputs as spam or non-spam, with the model's scalable design enabling more

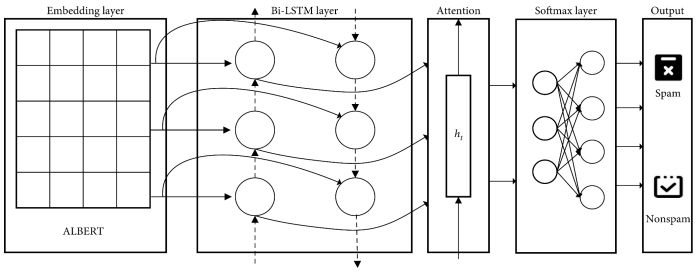


Figure 1: Architectural Overview of ALBERT Model for Spam Detection

profound neural network architectures without excessive computational overheads. Hence, ALBERT stands as a scalable solution for spam detection, advancing beyond BERT's capabilities for NLP challenges in social media contexts.

## III. RESEARCH OBJECTIVES

Each research objective has an Id. Each of these objectives is designed to collectively enhance the understanding of spam detection on Instagram, aiming for a solution that is more sophisticated and effective than the platform's existing measures:

### A. *Objective ID: RO1*

Significance: Develop a nuanced comprehension of the spam dynamics prevalent on Instagram by acquiring and preprocessing a varied collection of Instagram comments. This task will involve the translation of specific terminologies from related domains (e.g., YouTube comments) to their Instagram counterparts, such as converting “subscribe to” to “follow” and “channel” to “account”. This adaptation ensures the dataset's relevance and authenticity to Instagram's context.

### B. *Objective ID: RO2*

Significance: Employ the pre-trained ALBERT NLP model to develop and train a spam detection model, integrating keyword-based classification methods. This objective is crucial for leveraging advanced AI techniques to improve spam detection efficiency beyond traditional rule-based systems.

### C. *Objective ID: RO3*

Significance: This evaluation will comprehensively evaluate the spam detection model's performance using accuracy, precision, recall, and F1-score metrics. It will test the model's effectiveness and capability to adapt to and identify evolving spam tactics.

### D. *Objective ID: RO4*

Significance: Conduct an in-depth analysis of the developed model's strengths and weaknesses, with a particular focus on its versatility in addressing various spam manifestations. This will include conducting a comparative analysis of word frequencies within spam and non-spam comments to uncover insights for potential model refinements.

## IV. RESEARCH METHODOLOGY

The data collection process for this project involves utilizing a dataset from Kaggle, originally curated as a public set of YouTube comments collected for spam research [6]. The dataset comprises 1,956 non-spam messages gathered from five highly viewed YouTube videos. The messages are classified as spam (indicated as 1) or non-spam (indicated as 0), with 51.4% of the messages being spam and 48.6% being non-spam. We opted for this dataset and later adapted it to suit Instagram comments as Instagram’s policies prohibit automated data extraction, while YouTube’s policies do not, thus making datasets from YouTube more accessible [7].

## V. DATA MODIFICATION

The next step in collecting the dataset involves modifying the dataset to align it with real-life Instagram comments. This entails analyzing the dataset and identifying words or phrases that need to be adjusted to better fit the context of Instagram. For instance, terms such as “subscribe to” may be replaced with “follow”, and “channel” may be substituted with “account”, ensuring the dataset accurately represents Instagram interactions. The data set is then read over in order to ensure that all sentences are grammatically correct and understood in the context of an Instagram comment after being modified.

## VI. EXPLORATORY ANALYSIS

Before proceeding with Data Preprocessing, we conduct exploratory analysis on the data to identify significant elements crucial for model training. This analysis aims to discern key attributes within the test data that warrant attention. Among these factors are specific keywords that could help the model in distinguishing between real and spam comments based on the frequency of their appearance across the spam or non-spam sets. Our exploration revealed the top 10 most frequent words in the spam dataset are "Check," "this," "check," "follow," "post," "account," "Instagram," "please," "like," and "just" (refer to Figure 2). It's noteworthy that we treat "Check" and "check" as distinct words, indicating the former typically appears at the beginning of sentences. To assess the utility of these words in classifying comments, we analyzed their distribution in spam versus non-spam comments (see Figure 3). Notably, "please" and “Check” were found only in spam comments, suggesting that the use of the word “please” or “Check” specifically at the beginning of a sentence could be used as a potential indicator of a spam comment. Moreover, "follow," "check," "account," and "Instagram" exhibit an 85% or higher frequency in spam comments compared to non-spam ones. The words “this”, “post”, and “like” appear in relatively equal percentages across both datasets.

Thus, while certain words may show up often in spam comments, their importance lies in how they compare between the two datasets, helping to spot potential spam indicators.

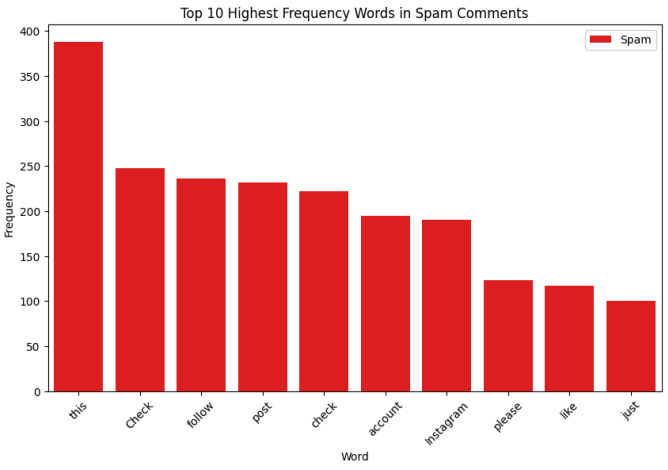


Figure 2: Top 10 Frequency Words in Spam Comments

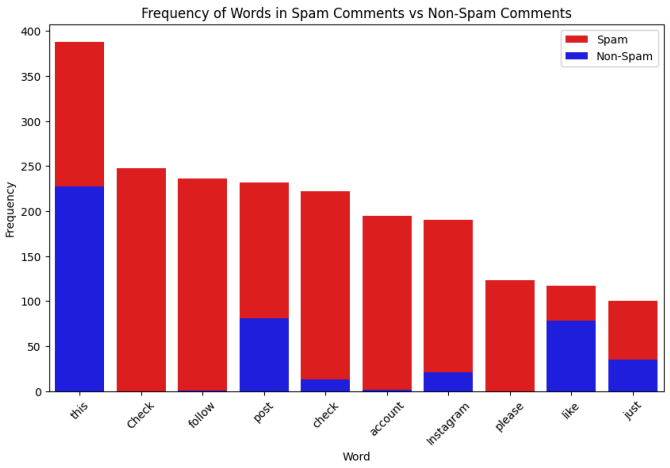


Figure 3: Frequency of Words in Spam Comments vs Non-Spam Comments

## VII. DATA PREPROCESSING

The data collected has already been cleaned up, it will next undergo pre-processing of tokenization and removal of special characters to prepare it for model training. Pre-processing also included changing the encoding from 'cp1252' to 'utf-8-sig' and combining all datasets into one csv. This is a crucial step done to ensure that the data is usable to train and be used by the ALBERT model.

## VIII. MODEL TRAINING

The pre-trained ALBERT model will be fine-tuned on the collected Instagram comment dataset using supervised learning techniques. During training, the model will learn to distinguish between genuine and spam comments by adjusting its parameters based on the provided labeled data. ALBERT was chosen due to its efficiency and effectiveness in handling NLP tasks. Its lightweight architecture allows for faster training and inference compared to other transformer models like BERT, making it suitable for real-time applications such as spam detection on social media platforms. It is a relatively new model and is thus conducting research on it's effectiveness within this application can be useful for improving spam detection models in the future. Supervised learning was chosen as the primary approach for model training due to the availability of labeled data. By providing the model with labeled examples of genuine and spam comments, it can learn to differentiate between the two categories and make predictions on unseen data.

## IX. EVALUATION METRICS

The performance of the spam detection model will be evaluated using various metrics such as accuracy, precision, recall, and F1-score. Additionally, the model will be tested on unseen Instagram data to assess its generalization capability. Precision, recall, and F1-score were selected as evaluation metrics to provide a comprehensive assessment of the model's performance. Precision measures the accuracy of the model in identifying spam comments, while recall measures its ability to detect all spam comments in the dataset. The F1-score provides a balance between precision and recall, taking into account both false positives and false negatives.

Once trained, the model is tested using a dataset of 100 real Instagram comments gathered ourselves from popular instagram posts. 47 of the collected comments were spam comments and 53 of them were non-spam. The precision, recall and F-1 score can then be compared to that of the trained model's fold scores, to see how capable the model was at evaluating real comments.

## X. RESULTS

The development and testing of the spam detection model using the pre-trained NLP model ALBERT underwent multiple iterations to refine its performance and address inherent challenges. Here, we delineate the key findings and provide a comprehensive analysis:

*First Iteration: Initial Model Setup and Resource Assessment*

The primary objective of the first iteration was to ascertain the feasibility of implementing ALBERT for our spam detection task. We conducted an initial test run without incorporating K-fold cross-validation or a validation set. The dataset, comprising 1975 samples, underwent an 80-20 split for training and testing, respectively, with shuffling to ensure randomness.

Given the resource-intensive nature of ALBERT, unexpected challenges arose due to its heavy computational demands. The model was executed on a system equipped with a 16GB RAM and an Intel Core i7 processor. However, during runtime, CPU and memory usage occasionally surpassed the 90% threshold. Notably, the absence of GPU support hindered efficient model training, prompting us to cancel the first iteration after a runtime of 2 hours with minimal progress.

*Second Iteration: Implementation of GPU Support and Progress Visualization*

Building upon the insights gained from the initial iteration, our focus shifted towards enhancing computational efficiency by incorporating GPU support. We installed and enabled CUDA toolkit 11.8 for Windows and updated PyTorch to support CUDA, thereby leveraging GPU acceleration for model training. Additionally, we introduced progress visualization using tqdm to monitor training progress in a more structured manner.

Table I: Results of First Iteration

| Epoch | Batch/Second | Loss |
|---|---|---|
| 1/3 | 4.79 | 0.347 |
| 2/3 | 4.59 | 0.152 |
| 3/3 | 4.60 | 0.229 |

*Third Iteration: Integration of K-fold Cross-Validation and Performance Metrics*

In the third iteration, our efforts were directed towards enhancing model robustness and evaluation methodologies. We implemented K-fold cross-validation to mitigate overfitting and variability issues, thereby enhancing the model's generalization capabilities. Furthermore, we explored the possibility of reducing the number of training epochs and introduced additional performance metrics to provide more comprehensive insights into model performance. However, concerns were raised regarding the adequacy of the dataset size, prompting us to further evaluate dataset balance and adequacy.

Table II: Results of Third Iteration

| Fold Number | Batch/Second | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| Fold 1 | 2.06 | 0.8686 | 0.8138 | 0.9757 | 0.8874 |
| Fold 2 | 2.21 | 0.9433 | 0.9836 | 0.9045 | 0.9424 |
| Fold 3 | 2.09 | 0.9536 | 0.9592 | 0.9495 | 0.9543 |
| Fold 4 | 2.09 | 0.8737 | 0.9811 | 0.7723 | 0.8643 |
| Fold 5 | 2.08 | 0.7321 | 0.9889 | 0.4611 | 0.6290 |
| Average Across 5 folds | | 0.8743 | 0.9453 | 0.8126 | 0.8555 |

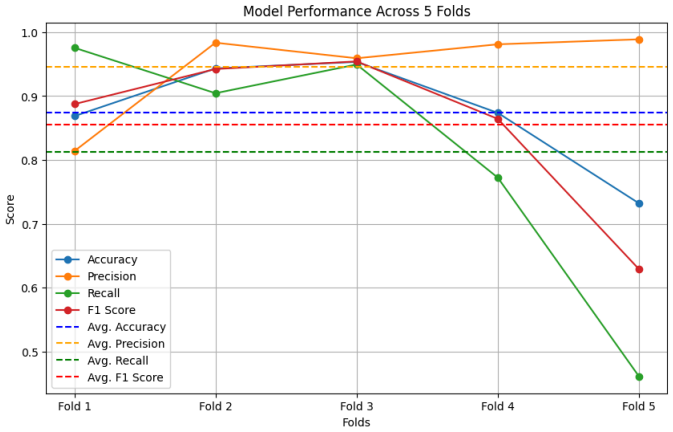


Figure 4: Model Performance Across 5 Folds

Table III: Results of the Real Instagram Dataset using Third Iteration Model

| Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|
| 0.65 | 0.6579 | 0.5319 | 0.5882 |

## XI. ANALYSIS AND INTERPRETATION OF RESULTS

The average results of iteration 3 seen in Table II and Figure 4 indicate that the model was able to accurately categorize whether a comment was spam or not 87% of the time. The fifth fold was the least balanced between recall and precision, with this imbalance reflected in the F1-score. This fold also has the lowest accuracy of the five, however, looking at the metrics as an average across all folds is more meaningful in determining the model's performance.

In the case of this project, we are aiming for a high accuracy and precision score because our goal is to maximize the detection of spam comments on Instagram in order to maximize their removal. We care greatly about precision, which indicates how many spam comments categorized as spam are actually spam, because false positives, meaning mistakenly flagging a comment as spam is detrimental to the functioning of instagram. Because of this, high precision is favourable to high accuracy, which means that the high average precision score of 94.5% across the training set is a very good result for this use case.

Once the model was trained, we tested the dataset using a set of 100 real Instagram comments. 47 of the comments were spam and 53 of them were non-spam. Table III shows the results of using this test set on the model.

When comparing the training set's average metrics (see Table II) to the test set, it is found that the test set was 22.43% less accurate, was 28.74% less precise and had 28.07% less recall than the training sets. The F1 score of the set: 0.5882 is lower than the training set's average F1 score by 0.2673. All of this indicates that the trained model was not very successful in categorizing real life Instagram comments into spam or non-spam.

Reasons for the vast difference in metrics between the training sets' and test sets' classification is likely due to the data gathering of the test set coming directly from Instagram posts, as opposed to Youtube comments that were then modified. There appears to be a fundamental difference in spam between YouTube and Instagram, with the latter platform's unique functionalities needing to be adequately represented in the training data (e.g. Instagram has stories and direct messaging, while Youtube does not). The rapid evolution of social media vernacular and the potential outdatedness of the YouTube dataset could contribute to the model's reduced effectiveness in current contexts. The Youtube dataset did not account for the varied forms of spam on

Instagram, including engagement-boosting bot comments, taking the form of a singular emoji and random word (e.g. 😸fighters, 💐outputs, within☝), diverging from the primarily promotional spam observed on YouTube.

Modern spam bots' strategies, including emojis and alternative fonts to evade detection, were not sufficiently represented in the training process either, indicating a need for more sophisticated data preprocessing and model training approaches (e.g. 1o.ooo **FOḶḶOWERS** 7**O**$, ⟦🅒🅗🅔🅒🅚.🅜🅨.🅢🅣🅞🅡🅨❤️). Using the ALBERT model to filter spam comments is not a novel concept. The novelty of this project comes in the form of exploring the assumption that one social media's spam comments can be tailored to fit the form of another social media's with comparable accuracy.

## XII. CONCLUSION

The project successfully demonstrated the application of the ALBERT model for detecting spam on Instagram, marking a significant step in addressing the shortcomings of rule-based systems with advanced NLP techniques. Despite challenges, the model showed promising results in understanding the subtleties of human language, signaling its potential for sophisticated, context-aware spam detection.

The results substantiate the hypothesis that sophisticated NLP models like ALBERT can outperform traditional spam detection methods. However, they also highlight the importance of relevant, platform-specific training data and the dynamic nature of spam tactics, emphasizing the need for continual model refinement.

To improve upon the training process of the ALBERT model, attributes other than the content of a comment could be considered. Often, the account name and verification status of a commenter are used as proof in distinguishing real people, as spam bots typically have strange or basic sounding names and cannot verify themselves.

## REFERENCES


[1] Solomon, "Instagram Spam Detector," GitHub repository. Available: https://github.com/Solomon04/InstagramSpamDetector. Accessed on: Apr. 8, 2024.

[2] Y. Zhang and L. Sun, "Instagram Spam Detection," Semantic Scholar. [Online]. Available: https://www.semanticscholar.org/paper/Instagram-Spam-Detection-Zhang-Sun. Accessed on: Apr. 8, 2024.

[3] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," arXiv preprint arXiv:1909.11942, 2019.

[4] Hindawi, "Spam detection with ALBERT," Hindawi Journals, 2021.

[5] NotShrirang, "Spam Filter using ALBERT," GitHub repository. Available: https://github.com/NotShrirang/SpamFilterUsingALBERT. Accessed on: Apr. 8, 2024.

[6] L. N, “YouTube Spam Collection Data Set,” Kaggle. Available: https://www.kaggle.com/datasets/lakshmi25npathi/images. Accessed on: Mar. 6, 2024.

[6] “Master Instagram scraping: A guide for prospecting and Competitive Research,” Magical Text Expander & Autofill. Available: https://www.getmagical.com/blog/how-to-scrape-instagram. Accessed on: Mar. 6, 2024.