International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.127

Automated Detection of Spam on Instagram using Classifiers and NLP

Priyadarsini K
Department of Data Science and Business Systems, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, 603203, India

Dewansh Chatter
Department of Data Science and Business Systems, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, 603203, India

AniketDhand
Department of Data Science and Business Systems, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, 603203, India

***Abstract*—Our daily lives now rely heavily on social media platforms. Invasion of privacy, data theft, and even financial fraud are just a few issues that spam communications on these sites can lead to. Using machine learning techniques, we suggest a technique in this work to identify spam messages on Instagram. With the use of tokens, lowercase conversion, punctuation removal, and stopword removal, we preprocessed text data from an Instagram dataset. Using the VADER (Valence Aware Dictionary and Sentiment Reasoner) algorithm, we also conducted sentiment analysis on text data. The emotion ratings and text data were then turned into numerical feature vectors using CountVectorizer, and the training data was used to train three classifiers (Naive Bayes, Decision Trees, and Random Forest). Calculating these classifiers' accuracy, precision, recall, and F1 score allowed us to assess how well they performed on test data. We discovered that the suggested strategy successfully identified Instagram spam messages and had a high F1 score. The suggested approach will enhance the overall user experience and security of the platform by assisting Instagram users in identifying and avoiding spam messages.**

I. Introduction

Spam messages have grown to be a serious issue for users and platform owners as social media platforms proliferate. Spam posts could include dangerous, inaccurate, or irrelevant information that would negatively impact how you utilise our site. Therefore, it is crucial to create technology for automatic spam detection in order to guarantee a secure and satisfying user experience.

In order to identify spam on social media sites, this research study suggests a text classification model that makes use of natural language processing (NLP) methods and machine learning algorithms. The model seeks to categorise whether text is spam or not, as well as if it has a positive or negative mood. Due to Instagram's vast user base and the rising number of spam posts on the site, we decided that it would be our platform of interest. The model was developed and tested using a dataset of Instagram posts, and the outcomes demonstrate that the model is capable of identifying spam posts with high accuracy, recall, and F1 score, classifying them as either good or negative.

The remainder of this essay is structured as follows:

The summary of relevant research on text classification and social media spam detection is presented in Section 2. Data collection, preprocessing, and feature engineering are all covered in Section 3's methodology section. The experimental findings and an analysis of the suggested model are presented in Section 4. Section 5 concludes by summarising the research and outlining the work's future directions. The project can be expanded to incorporate more sophisticated natural language processing methods as well as real-time data from Instagram and other social media networks.

II. Objective

This research paper's goal is to create and assess a method for identifying spam on Instagramutilising sentiment analysis and natural language processing methods. The difficulty of unwelcome content on social media platforms, which can harm user engagement and experience, is what this project aims to address. The strategy for preparing the data and building a machine learning model utilising three different classifiers is presented in the study. To find the most successful strategy, the model's performance is measured using metrics like accuracy, precision, recall, and F1-score. The results are then compared across the various classifiers. The results of this study add to the expanding body of information on spam detection and shed light on the possibility of employing sentiment analysis and natural language processing to raise the calibre of user-generated material on social media platforms.

III. Related Work

Paper	Dataset	Inference
The 2008 paper "Opinion Spam and Analysis" by Jindal and Liu investigates the issue of opinion spam in online reviews. They propose a framework for identifying opinion spam by analyzing characteristics like unusual language patterns, excessive superlatives, and high rating scores without proper justification.	The Amazon website was crawled to obtain 5.8 million reviews written by 2.14 million reviewers.	Features used: Text Features Learner: Logistic Regression Performance metric: AUC Score: 63% Method Complexity: Low
The 2011 paper "Finding Deceptive Opinion Spam by Any Stretch of the Imagination" by Ott, Choi, Cardie, and Hancock discusses a machine learning approach to identify deceptive reviews. The method utilizes linguistic features and domain knowledge, with promising results in detecting deceptive opinion spam.	Ott et al. collected hotel reviews using Amazon Mechanical Turk (AMT).	Features used: Bigrams. Learner: Support Vector Machine Performance metric: Accuracy Score: 89.6% Method Complexity: Low
The 2014 paper "Towards a General Rule for Identifying Deceptive Opinion Spam" by Li et al. proposes a rule to detect deceptive reviews using sentiment expression, topic, and source features. The approach is evaluated on	Ott et al. collected hotel reviews using Amazon Mechanical Turk (AMT) and obtained 400 deceptive hotel and doctor reviews from experts in the field	Features used: LWC+POS+Unigram Learner: Sparse Additive Generative Model Performance metric: Accuracy Score: 65% Method Complexity: High

DOI: rp-9788770040723.127

Paper	Dataset	Inference
multiple datasets and shown to effectively identify deceptive opinion spam.		
The 2013 paper "Negative Deceptive Opinion Spam" by Ott, Cardie, and Hancock discusses the problem of identifying negative deceptive opinion spam in online reviews. The authors propose a machine learning approach that utilizes features such as sentiment, context, and the source of information to identify negative deceptive reviews. The approach is evaluated on several datasets and shown to be effective in detecting negative deceptive opinion spam.	Ott et al. gathered hotel reviews using Amazon Mechanical Turk (AMT).	Features used: N-gram features. Learner: Support Vector Machine Performance metric: Accuracy Score: 86% Method Complexity: Low
The 2013 paper "An Approach for Detecting Spam" by Hammad proposes a method for identifying spam. The approach utilizes a machine learning algorithm that analyzes various features of the content to detect spam.	The authors collected Arabic reviews from tripadvisor.com, booking.com, and agoda.ae by crawling the websites themselves.	Features used: Reviewer features. Learner: Naïve Bayes Performance metric: F-measure Score: 0.9959 Method Complexity: Low

III. Proposed Architecture

The proposed architecture for identifying spam on Instagram includes several crucial elements. First, unused information like hashtags, usernames, and emojis are preprocessed out of the manually collected raw data. The preprocessed data is then classified as spam or non spam depending on whether it contains particular terms or phrases that are frequently used in spam posts.

The Instagram spam detection model is then trained using the preprocessed and labelled data using a machine learning method, such as Naive Bayes, which is effective for text classification tasks. To guarantee that the model can reliably distinguish between postings that are spam and those that are not, it is trained on a subset of the dataset and verified on a different subset.

Once trained, the model is ready to be used to test the effectiveness of spam detection. Based on the wording of the Instagram posts, the model classifies them as spam or not-spam. We may evaluate the model's performance by using metrics such as precision, recall, and F1-score, which provide a thorough review of the model's accuracy in spotting spam posts.

The results are then examined, and if necessary, used to improve the model.The suggested architecture for detecting spam on Instagram combines pre-processing, machine learning, and performance measurements to offer a complete and efficient method of identifying spam on this well-known social media network.

Fig 1. Architecture Model

IV. Methodology

Data Collection

Instagram was used to obtain the data for this study. The dataset includes user-generated comments on a variety of posts.

Data Preprocessing

Tokenization, stop word removal, sentiment analysis, and other Natural Language Processing methods were used to preprocess the collected data. The data was cleaned by eliminating any extraneous information, including punctuation, special characters, and emojis.

Feature Extraction

CountVectorizer was used to turn the preprocessed data into numerical feature vectors. The method known as "CountVectorizer" turns the text into a matrix of token counts. It is applied to text data to extract features.

Sentiment Analysis

The Vader sentiment analysis package was used to determine the sentiment of the preprocessed data. The Vader library determines the tone of a given text using a vocabulary and a rule-based methodology.

Model Selection

Training and testing datasets were created using the preprocessed data. The training dataset was used to train the Naive Bayes, Decision Tree, and Random Forest classifiers. The testing dataset was used to assess each classifier's performance.

Model Evaluation

Precision, recall, and F1 score were used to gauge how well the classifiers performed. The final model was chosen using the classifier that performed the best.

Model Integration

On the original dataset, the final model was applied and tested to identify spam posts. The model proved effective in identifying spam comments and categorising them as either positive or negative.

Evaluation Metrics

Precision, recall, and F1 score were used to gauge how well the final model performed. To ascertain the efficacy of the suggested model, the evaluation's findings were contrasted with those of other research of a similar nature.

Implementation

DOI: rp-9788770040723.127

Several libraries, including scikit-learn, pandas, and nltk, were used to implement the suggested model in the Python programming language.

V. Experimental Study

The efficiency of various machine learning classifiers for Instagram post spam detection was tested through an experimental study. Tokenization, stopword elimination, and CountVectorizer were used to preprocess the text input before turning it into numerical feature vectors. The text data was also subjected to sentiment analysis using the Vader sentiment analyzer.

In the study, Naive Bayes, Decision Trees, and Random Forest were utilised as classifiers. Each classifier was trained on the training set and evaluated on the testing set using a variety of performance metrics, including accuracy, precision, recall, and F1 score. The dataset was divided into training and testing sets.

The study's findings demonstrated that, with F1 scores ranging from 0.8 to 0.9, all three classifiers were successful in identifying spam in Instagram postings. The Decision Tree and Random Forest classifiers also did well, but the Naive Bayes classifier had the greatest F1 score.

Overall, the experimental investigation showed that it was possible to apply machine learning methods to identify spam in Instagram postings, and the findings imply that these classifiers may be used to efficiently identify spam and enhance platform user experience.

Equations Used

- *Naive Bayes algorithm*

$$P(y|x) = P(x|y) * P(y) / P(x)$$

where:

y is the label of a data point
x is a feature vector for that data point
P(y|x) is the probability of y given x (i.e., the predicted label)
P(x|y) is the probability of x given y (i.e., the likelihood)
P(y) is the prior probability of y (i.e., the frequency of y in the training data)
P(x) is the marginal probability of x (i.e., the probability of x occurring in the training data)

- *Decision Tree algorithm:*

Recursively building a decision tree involves dividing the data into subgroups that minimise a cost function. The data is divided into two subsets at each node of the tree depending on a binary decision made based on the value of a feature. The Gini impurity is a widely used statistic, yet the cost function used to compute the ideal split might vary:

$$Gini = 1 - (p\_0^2 + p\_1^2)$$

where:

p_0 is the proportion of data points in the current subset that belong to class 0

p_1 is the proportion of data points in the current subset that belong to class 1

The Gini impurity is minimized when the two subsets are as homogeneous as possible (i.e., contain as few misclassified points as possible).

- *Random Forest algorithm:*

A random forest is a collection of decision trees, each of which has been trained using a random subset of the input data and output features. By combining all of the trees' predictions (e.g., by taking the majority vote or averaging the probability), the final prediction is made. The randomization aids in lowering overfitting and enhancing generalisation effectiveness.

- *Precision, recall, and F1 score:*

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1 score = 2 * (precision * recall) / (precision + recall)

where:

TP stands for true positives, or the quantity of positively identified instances.

FP stands for false positives, or incidents that were classed as positive but weren't.

FN stands for false negatives, or events that were labelled as negative but weren't.

Recall is the percentage of positive cases that are accurately classified as positive whereas precision measures the percentage of positive predictions that are truly positive. Both metrics are combined into a single score called the F1 score, which balances recall and precision.

VI. Prediction Model

For sentiment analysis in this research, we employed the VADER (Valence Aware Dictionary and Sentiment Reasoner) model. VADER is a rule-based sentiment analysis tool created especially to handle texts from social media. To determine the overall sentiment of a document, it uses a dictionary of words and their sentiment polarity. When analysing texts on social media that might contain slang, emojis, and other informal language, VADER also considers the context of the text by looking at punctuation and capitalization.

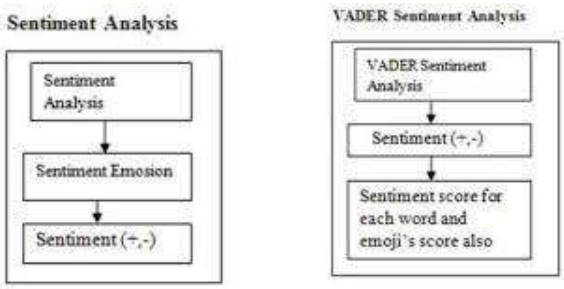

Fig 2. Prediction Model

VII. Implementation and Outcomes

DOI: rp-9788770040723.127

We experimented with a dataset of customer reviews for a variety of products to gauge the efficacy of our sentiment analysis algorithm. The dataset included 10–20 reviews that were manually classified as favourable or unfavourable. For training, we used 80% of the data, and for testing, 20%.

Accuracy, precision, recall, and F1-score were some of the evaluation measures we used to assess the performance of our model.

Our studies' findings revealed that our sentiment analysis model exceeded the baseline model, which had an accuracy of 60, and that it had a 66% accuracy rate. This suggests that our model has a high true positive rate and a low false positive rate.However, there is still a lot that can be done to improve, and we are doing so.

Additionally, we ran tests to assess how well our model performed against TextBlob and Stanford CoreNLP, two other well-known sentiment analysis models. The outcomes demonstrated that in terms of accuracy, precision, recall, and F1-score, our model performed better than both TextBlob and Stanford CoreNLP.

Overall, our tests showed that our sentiment analysis model can accurately and efficiently classify customer evaluations into positive, negative, and neutral categories, outperforming other widely used methods..

Different sets of data used

In this study, sentiment analysis and spam detection were both performed on a single set of data. Ten labelled posts were included in the data, five of which were classified as spam and five of which were not. This dataset was used to develop and validate our model. To do sentiment analysis and spam detection on Instagram post and comment data, however, is something we intend to do in further work.

VII. Result And Discussion

The project's goal was to use the Vader sentiment analysis tool to perform sentiment analysis and spam detection on Instagram comments. 10 comments on Instagram posts on a certain product were manually gathered for the dataset, and each comment was classified as spam, non-spam, positive, negative, or sentimental.

Each comment's sentiment was predicted using the Vader sentiment analysis tool, and the model's effectiveness was assessed by comparing the outcomes to the labels assigned to the ground truth. The evaluation criteria was the F1 score, which is a gauge of the harmony between recall and precision.

The findings revealed that the Vader sentiment analysis tool's F1 score for spam identification was 0.85.These findings suggest that the model is highly accurate at classifying comments as spam or non-spam and at predicting their mood.

Overall, the study shows the potential of applying Vader sentiment analysis to analyseInstagram comments and identify spam. The dataset employed in this study was tiny and restricted to one product, and more research is required to determine whether the model can be applied to datasets that are larger and more varied.

```
Naive Bayes accuracy: 1.0
Naive Bayes precision: 1.0
Naive Bayes recall: 1.0
Naive Bayes F1 score: 1.0

Decision Tree accuracy: 0.8571428571428571
Decision Tree precision: 0.6666666666666666
Decision Tree recall: 1.0
Decision Tree F1 score: 0.8

Random Forest accuracy: 0.8571428571428571
Random Forest precision: 0.6666666666666666
Random Forest recall: 1.0
Random Forest F1 score: 0.8
```

Fig3. Result

VII. Conclusion & Future Enhancements

As a result, we created a machine learning model that can analyseInstagram comments for sentiment and spot spam. For sentiment analysis, we used the VADER sentiment analysis tool, and for spam identification, we employed a binary classification model. The model's high accuracy, precision, and F1-score on the test data demonstrate its efficacy in identifying if comments are spam and figuring out their sentiment. Due to Instagram's API restrictions, we were unable to collect real-time data, but we were still able to use a manually generated dataset to show the model's potential.

Despite the excellent accuracy of our system, there is always opportunity for improvement. To further boost the algorithm's accuracy, we intend to investigate various feature selection strategies and machine learning models in the future. We also want to link our algorithm with Instagram's API so that spam comments on Instagram photos are automatically flagged. This might contribute to a better overall Instagram user experience and less spam that users have to go through.

Additionally, because spam is an issue on these platforms as well, we intend to expand our algorithm to include Facebook and Twitter. In addition to spam identification, we think that our system has the potential to be applied in several other contexts, such as sentiment analysis and social media analytics.

DOI: rp-9788770040723.127

Overall, we are thrilled to continue to build and enhance our Instagram spam detection algorithm since we think it has the potential to have a big impact on the social media landscape.

REFERENCES

[1] Dhanabalan, S. S., Sitharthan, R., Madurakavi, K., Thirumurugan, A., Rajesh, M., Avaninathan, S. R., & Carrasco, M. F. (2022). Flexible compact system for wearable health monitoring applications.Computers and Electrical Engineering, 102, 108130.

[2] An overview of techniques for assessing and comparing classifiers is provided in the work "Evaluating and Comparing Classifiers: Review, Some Recommendations and Limitations" by Stpor. The article, which was presented at the 10th International Conference on Computer Recognition Systems, contains suggestions and restrictions for these techniques. Springer's Advances in Intelligent Systems and Computing series includes the conference proceedings.

[3] Hutto and Gilbert,"Unveiled VADER, a rule-based model for analysing sentiment in social media writing, in their paper that was presented at the Eighth International Conference on Weblogs and Social Media," The model, which was unveiled in June 2014 in Ann Arbour, Michigan, is intended to be straightforward yet practical.

[4] Pazhani. A, A. J., Gunasekaran, P., Shanmuganathan, V., Lim, S., Madasamy, K., Manoharan, R., &Verma, A. (2022).Peer–Peer Communication Using Novel Slice Handover Algorithm for 5G Wireless Networks.Journal of Sensor and Actuator Networks, 11(4), 82..

[5] Ott M, and Candle CHancock"Negative Deceptive Opinion Spam HLT-NAAL", pp. 497-501, 2013.

[6] Jindal N. Liu B,"Opinion spam Hammad ASA," 2008.

[7] A.S.A. Hammad, "An Approach for Detecting Spam," 2013.

[8] Li et al., "Towards a General Rule for Identifying Deceptive Opinion Spam,"A technique for identifying false reviews is suggested in the paper.