Type of the Paper (Article, Review, Communication, etc.)

HATE SPEECH DETECTION ON INSTAGRAM COMMENTS USIND ML & DEEP LEARNING

Durga Prasad Thota 1, Sai Chitra Nagarageri 2 and Varun Kumar Polisetty 3, \*

|  |
| --- |
| **Citation:** To be added by editorial staff during production.  Academic Editor: Firstname Lastname  Received: date  Revised: date  Accepted: date  Published: date    **Copyright:** © 2024 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). |

1 Affiliation 1; [dthota@saintpeters.edu](mailto:dthota@saintpeters.edu), [snagarageri@saintpeters.edu](mailto:snagarageri@saintpeters.edu), [vpolisetty@saintpeters.edu](mailto:vpolisetty@saintpeters.edu)

**\*** Correspondence: [dthota@saintpeters.edu](mailto:dthota@saintpeters.edu); Tel.: +1 2014486265

**Featured Application: Authors are encouraged to provide a concise description of the specific application or a potential application of the work. This section is not mandatory.**

**Abstract**: Social media platforms such as Instagram, Facebook, and YouTube have become an integral part of our daily lives. However, alongside the positive aspects of social media, the misuse of these platforms has led to a significant rise in hate speech and abusive comments. This behavior not only disrupts the online experience but can also cause psychological distress to individuals.

Instagram serves as a platform where users share personal moments and ideas through posts. The comment section fosters communication by allowing individuals to express their opinions, thoughts, and views. Unfortunately, some users exploit this space to spread hateful content, including vulgarity, demotivating remarks, and abusive language. Recognizing this as a growing problem, we chose to focus on analyzing the text data from Instagram’s comment sections to develop a hate speech detection system. This system will leverage Machine Learning (ML) and Deep Learning (DL) techniques to identify and mitigate such harmful content.

**Keywords:** Hate speech; Instagram; comment section; Abusive Comments; Machine learning; Deep learning; Text Data Analysis.

0. How to Use This Template

The template details the sections that can be used in a manuscript. Note that each section has a corresponding style, which can be found in the “Styles” menu of Word. Sections that are not mandatory are listed as such. The section titles given are for articles. Review papers and other article types have a more flexible structure.

Remove this paragraph and start section numbering with 1. For any questions, please contact the editorial office of the journal or support@mdpi.com.

1. Introduction

The rise of social media has transformed how individuals communicate and interact, but it has also introduced challenges, such as managing harmful and inappropriate content. This project presents a machine learning-based framework that leverages state-of-the-art natural language processing (NLP) and deep learning models to detect and moderate inappropriate comments. By utilizing scalable cloud-based tools and robust machine learning techniques, the solution ensures efficient and real-time content moderation.

1. Materials and Methods

Data Source and Collection:  
The primary data source for this project includes Instagram comments collected using both manual CSV files and scraping live comments using scraping tools(https://exportcomments.com/export-instagram-comments). The collected data was anonymized to ensure user privacy.

**Data Storage and Preprocessing:**  
The data was stored in Azure Data Lake Storage Gen2 (ADLS Gen2) and preprocessed using Azure Data Factory and Azure Databricks. Preprocessing steps included:

* Cleaning the text data to remove special characters, stopwords, and irrelevant content.
* Tokenization and normalization of textual data.
* Splitting the dataset into training and testing subsets in an 80:20 ratio.

**Exploratory Data Analysis:**

The EDA process has been performed on the preprocessed data.

**Model Selection and Training:**  
The model was developed and trained using Azure Machine Learning Studio. Several machine learning and deep learning models were evaluated to select the most effective architecture for hate speech detection. The chosen model was DistilBERT, fine-tuned for text classification tasks.

**Model Evaluation:**  
The model's performance was evaluated on the test dataset using metrics such as accuracy, precision, recall, and F1-score. These results were used to refine the model and optimize its hyperparameters.

**Model Deployment:**  
The finalized model was deployed using Streamlit or an alternative UI framework. Real-time data was fed to the deployed model, enabling live hate speech detection for Instagram comments. Detected hate speech triggered specific actions:

* Flagging the comment.
* Reporting the user account.
* Sending warning messages.
* Deleting the comment.

**Ethical Compliance:**  
All data was processed in compliance with GDPR and other relevant privacy standards. No personal data was stored or shared beyond the scope of this study.

**Data and Code Availability:**  
The dataset used in this study and the corresponding scripts are available on GitHub at [link to repository]. If restricted, access will be granted upon request.

3. Results

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

3.1. Subsection

3.1.1. Subsubsection

Bulleted lists look like this:

* First bullet;
* Second bullet;
* Third bullet.

Numbered lists can be added as follows:

1. First item;
2. Second item;
3. Third item.

The text continues here.

3.2. Figures, Tables and Schemes

All figures and tables should be cited in the main text as Figure 1, Table 1, etc.

****

**Figure 1.** This is a figure. Schemes follow the same formatting.

**Table 1.** This is a table. Tables should be placed in the main text near to the first time they are cited.

|  |  |  |
| --- | --- | --- |
| **Title 1** | **Title 2** | **Title 3** |
| entry 1 | data | data |
| entry 2 | data | data 1 |

1 Tables may have a footer.

The text continues here (Figure 2 and Table 2).

|  |  |
| --- | --- |
| C:\Users\martin\Downloads\testFigure.tif | C:\Users\martin\Downloads\testFigure.tif |
| (**a**) | (**b**) |

**Figure 2.** This is a figure. Schemes follow another format. If there are multiple panels, they should be listed as: (**a**) Description of what is contained in the first panel; (**b**) Description of what is contained in the second panel. Figures should be placed in the main text near to the first time they are cited.

**Table 2.** This is a table. Tables should be placed in the main text near to the first time they are cited.

|  |  |  |  |
| --- | --- | --- | --- |
| **Title 1** | **Title 2** | **Title 3** | **Title 4** |
| DistilBERT \* | data | data | data |
| data | data | data |
| data | data | data |
| LSTM | data | data | data |
| data | data | data |
| CNN | data | data | data |
| data | data | data |
| data | data | data |
| data | data | data |
| SVM | data | data | data |
| data | data | data |

\* Tables may have a footer.

3.3. Formatting of Mathematical Components

This is example 1 of an equation:

|  |  |
| --- | --- |
| a = 1, | (1) |

the text following an equation need not be a new paragraph. Please punctuate equations as regular text.

This is example 2 of an equation:

|  |  |
| --- | --- |
| a = b + c + d + e + f + g + h + i + j + k + l + m + n + o + p + q + r + s + t + u + v + w + x + y + z | (2) |

the text following an equation need not be a new paragraph. Please punctuate equations as regular text.

Theorem-type environments (including propositions, lemmas, corollaries etc.) can be formatted as follows:

**Theorem 1.** Example text of a theorem. Theorems, propositions, lemmas, etc. should be numbered sequentially (i.e., Proposition 2 follows Theorem 1). Examples or Remarks use the same formatting, but should be numbered separately, so a document may contain Theorem 1, Remark 1 and Example 1.

The text continues here. Proofs must be formatted as follows:

**Proof of Theorem 1.** Text of the proof. Note that the phrase “of Theorem 1” is optional if it is clear which theorem is being referred to. Always finish a proof with the following symbol. □

The text continues here.

4. Discussion

Authors should discuss the results and how they can be interpreted from the perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

5. Conclusions

This section is not mandatory but can be added to the manuscript if the discussion is unusually long or complex.

6. Patents

This section is not mandatory but may be added if there are patents resulting from the work reported in this manuscript.

**Supplementary Materials:** The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1: title; Table S1: title; Video S1: title.

**Author Contributions:** For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used “Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.” Please turn to the [CRediT taxonomy](https://img.mdpi.org/data/contributor-role-instruction.pdf) for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

**Funding:** Please add: “This research received no external funding” or “This research was funded by NAME OF FUNDER, grant number XXX” and “The APC was funded by XXX”. Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding. Any errors may affect your future funding.

**Institutional Review Board Statement:** In this section, you should add the Institutional Review Board Statement and approval number, if relevant to your study. You might choose to exclude this statement if the study did not require ethical approval. Please note that the Editorial Office might ask you for further information. Please add “The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval).” for studies involving humans. OR “The animal study protocol was approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval).” for studies involving animals. OR “Ethical review and approval were waived for this study due to REASON (please provide a detailed justification).” OR “Not applicable” for studies not involving humans or animals.

**Informed Consent Statement:** Any research article describing a study involving humans should contain this statement. Please add “Informed consent was obtained from all subjects involved in the study.” OR “Patient consent was waived due to REASON (please provide a detailed justification).” OR “Not applicable.” for studies not involving humans. You might also choose to exclude this statement if the study did not involve humans.

Written informed consent for publication must be obtained from participating patients who can be identified (including by the patients themselves). Please state “Written informed consent has been obtained from the patient(s) to publish this paper” if applicable.

**Data Availability Statement:** We encourage all authors of articles published in MDPI journals to share their research data. In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Where no new data were created, or where data is unavailable due to privacy or ethical restrictions, a statement is still required. Suggested Data Availability Statements are available in section “MDPI Research Data Policies” at https://www.mdpi.com/ethics.

**Acknowledgments:** In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

**Conflicts of Interest:** Declare conflicts of interest or state “The authors declare no conflicts of interest.” Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results must be declared in this section. If there is no role, please state “The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results”.

**Appendix A**

The appendix is an optional section that can contain details and data supplemental to the main text—for example, explanations of experimental details that would disrupt the flow of the main text but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data is shown in the main text can be added here if brief, or as Supplementary data. Mathematical proofs of results not central to the paper can be added as an appendix.

**Appendix B**

All appendix sections must be cited in the main text. In the appendices, Figures, Tables, etc. should be labeled starting with “A”—e.g., Figure A1, Figure A2, etc.

References

References must be numbered in order of appearance in the text (including citations in tables and legends) and listed individually at the end of the manuscript. We recommend preparing the references with a bibliography software package, such as EndNote, ReferenceManager or Zotero to avoid typing mistakes and duplicated references. Include the digital object identifier (DOI) for all references where available.

Citations and references in the Supplementary Materials are permitted provided that they also appear in the reference list here.

In the text, reference numbers should be placed in square brackets [ ] and placed before the punctuation; for example [1], [1–3] or [1,3]. For embedded citations in the text with pagination, use both parentheses and brackets to indicate the reference number and page numbers; for example [5] (p. 10), or [6] (pp. 101–105).

1. Mullah, N.S.; Zainon, W.M.N.W. Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review. IEEE Access 2021, 9, 88364–88376.
2. Toktarova, A.; Syrlybay, D.; Myrzakhmetova, B.; Anuarbekova, G.; Rakhimbayeva, G.; Zhylanbaeva, B.; Suieuova, N.; Kerimbekov, M. Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods. Int. J. Adv. Comput. Sci. Appl. 2023, 14(5), 1–9. DOI: [10.14569/IJACSA.2023.0140542](http://dx.doi.org/10.14569/IJACSA.2023.0140542).
3. Subramanian, M.; Sathiskumar, V.E.; Deepalakshmi, G.; Cho, J.; Manikandan, G. A Survey on Hate Speech Detection and Sentiment Analysis Using Machine Learning and Deep Learning Models. J. Comput. Sci. 2023, X, 1–10
4. Akhter, M.P.; Jiangbin, Z.; Naqvi, I.R.; et al. Abusive Language Detection from Social Media Comments Using Conventional Machine Learning and Deep Learning Approaches. *Multimedia Syst.* 2022, *28*, 1925–1940. DOI: [10.1007/s00530-021-00784-8](https://doi.org/10.1007/s00530-021-00784-8).
5. Munyer, T.; Tanvir, A.A.; Das, A.; Zhong, X. DeepTextMark: A Deep Learning-Driven Text Watermarking Approach for Identifying Large Language Model Generated Text. IEEE Access 2024, 12, 40508–40520. DOI: 10.1109/ACCESS.2024.3376693.
6. Mekala, S.; Bhuvana, M.; Gupta, D.B.; Bhatt, V.; Kunekar, P.; Manoharan, G. Natural Language Processing and Deep Learning Techniques to Improve Sentiment Analysis in Social Media Texts. *Proceedings of the 2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, Gautam Buddha Nagar, India, 2023, pp. 1751–1755. DOI: 10.1109/IC3I59117.2023.10397735.
7. Akhter, M.P.; Jiangbin, Z.; Naqvi, I.R.; Zia, T. Abusive Language Detection from Social Media Comments Using Conventional Machine Learning and Deep Learning Approaches. *Multimedia Syst.* 2022, *28*(6), 1925–1940. DOI: [10.1007/s00530-021-00784-8](https://doi.org/10.1007/s00530-021-00784-8). Published: 01 December 2022.
8. Subramanian, M.; Sathiskumar, V.E.; Deepalakshmi, G.; Cho, J.; Manikandan, G. A Survey on Hate Speech Detection and Sentiment Analysis Using Machine Learning and Deep Learning Models. Alexandria Engineering Journal, 80, 110–121, 2023. Available online: <https://doi.org/10.1016/j.aej.2023.08.038> (accessed on 14 January 2025)

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.