International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

DATA ANALYTICS SYSTEM FOR OFFENSIVE MEMES TEXT CLASSIFICATION IN SOCIAL NETWORKS

Akshaya.M, Geethapriya.G, Janani.R, Dr.G.Pushpa Ph.D
Department of Computer Science and Engineering
E.G.S Pillay Engineering College,Nagapattinam,TamilNadu,India

---

ABSTRACT:

The dramatic evolution of memes on social media has raised the challenge of new content moderation challenges, most notably the identification of hate speech and objectionable content made up of a combination of visual and text data. This paper suggests a Data Analytics System for Offensive memes Text Classification using Optical Character Recognition (OCR), Natural Language Processing (NLP), and sentiment analysis to identify and moderate toxic meme content on the internet.

The system proposed performs text extraction from memes via OCR, processes it with text mining methods (tokenization, stop word elimination, stemming), and classifies sentiment via the VADER (Valence Aware Dictionary and sEntiment Reasoner) algorithm. The tiered moderation model applies escalating interventions—warnings for initial transgressions, image blocking for second offenses, and account suspension for habitual perpetrators—while informing users via SMS. Operates in Python and MySQL, the system is a real-world solution to battling hate speech online through memes. Transfer learning and multimodal analysis are potential future extensions to enhance detection performance for other cultural backgrounds. The work contributes to automatic content moderation through solving the specific issues of hate speech in memes through end-to-end OCR-NLP.

KEYWORDS:

Optical Character Recognition (OCR), Natural Language Processing (NLP), sentiment analysis, meme classification, hate speech detection, social media moderation, VADER algorithm, text mining, content filtering, deep learning.

INTRODUCTION

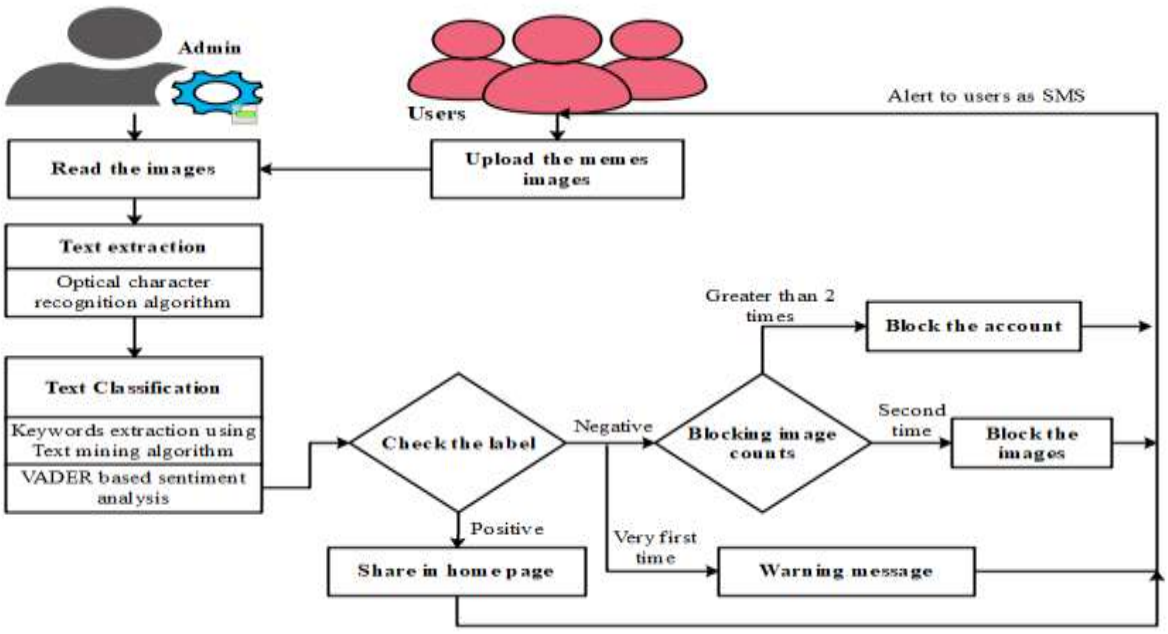
The rapid and sudden explosion of social media has made memes a very powerful medium of communication, combining words and images to convey satire, humor, and social commentary. However, this popularity has also come with the propagation of hate speech, cyberbullying, and abusive content. Current content moderation frameworks are pushed to their limits by memes since they are multimodal and the frameworks cannot generally detect problematic content that is predicated upon the interaction of visual and textual information. Text-based or image-based old-fashioned methods will not suffice, and manual moderation is not efficient or scalable. This gap suggests the need for an automated system to effectively analyze and classify offending content in memes to create safer online communities. To

address this issue, we propose an analytics system for offending meme classification that integrates Optical Character Recognition (OCR), Natural Language Processing (NLP), and sentiment analysis.

We scrape text out of memes using OCR, then normalize it applying NLP methods like tokenization and stop-word filtering, and estimate sentiment by the VADER algorithm. Our system uses multi-level moderation and gives a warning upon first offense, escalating to suspension of an account upon repeated offense. Combining these technologies, our system attempts to achieve greater detection accuracy while minimizing human moderation reliance. This research contributes to the emerging body of multimodal content analysis by offering a scalable solution to counter hate speech on memes.

LITERATURE REVIEW

Existing research on identifying hate content has also been unimodal, such that the text or images were being processed in either the domain or the other. Sentiment analysis and keyword search-based textual approaches using natural language processing (NLP) have worked effectively in identifying hate speech in plain text (Schmidt & Wiegand, 2017). Similarly, image-based techniques based on convolutional neural networks (CNNs) can only identify overt visual content but not textual content of memes (Kiela et al., 2019). However, unimodal techniques like these cannot recognize the nuanced coexistence between visual and textual content typical of meme communication and hence are plagued with humongous false positives and negatives. This constraint has driven recent work on multimodal systems that integrate OCR technology-based text extraction and end-to-end machine learning models (Gomez et al., 2020). Hybrid models that combine OCR-extracted visual features and visual inspection have been demonstrated in recent work to be effective.

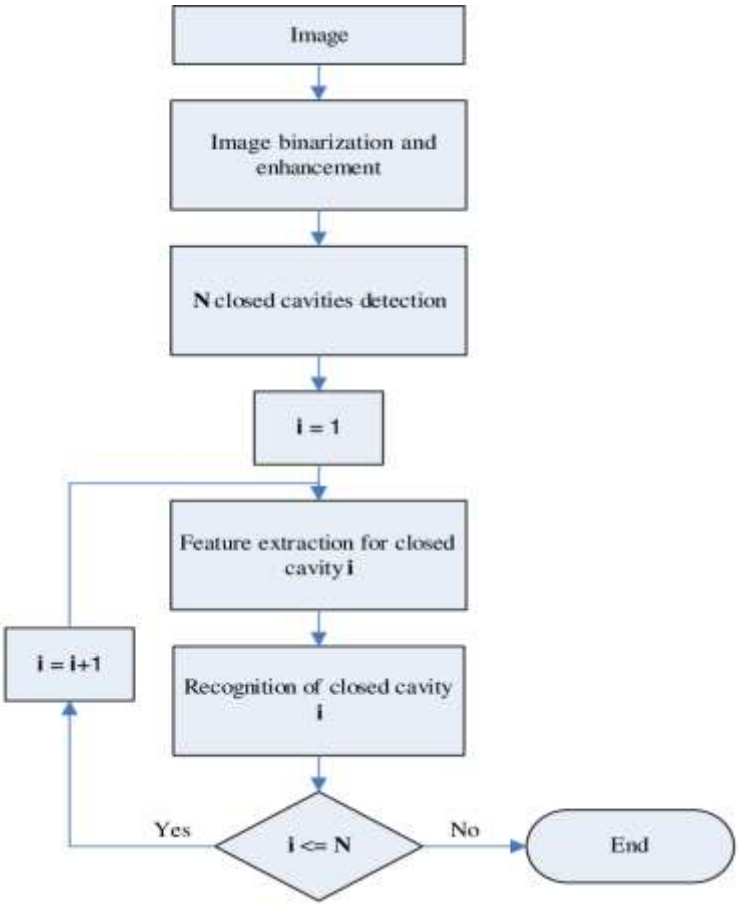


The Hateful Memes Challenge dataset (Kiela et al., 2020) has emerged as a benchmark for

capability and limitation testing of such multimodal systems. While more accurate transformer-based models like BERT and Vision Transformers are now present, they are still computationally costly and culturally biased (Kiela et al., 2021). In this work, the deficiencies are treated with the inclusion of a more sophisticated pipeline leveraging state-of-the-art OCR ability and VADER sentiment analysis to provide a balanced strategy that is as effective at detection and yet scalable enough to be used to real social media platforms. This paper is an extension of prior work and addresses concerns of computational efficiency and cultural sensitivity required in meme classification.

PROPOSED DESIGN

The offensiveness memes text classification system design proposed here combines Optical Character Recognition (OCR), Natural Language Processing (NLP), and sentiment analysis to automatically identify and moderate offensive memes in social network memes. The system starts with a simple upload interface for memes, which get parsed through an OCR engine to achieve text. The text pulled out is preprocessed (tokenization, stemming, stop word removal) prior to sentiment analysis via VADER algorithm to label content as positive, negative, or neutral.



Offending content (negativesentiment) has a graduated moderation approach where first offense is warnings, subsequent offense is image blocking, and further violation is suspension of an account, with all process escalated to users through SMS notifications. The Python backend processors, MySQL data storage, and commodity hardware (Intel processor, 4GB RAM) are envisioned. Transfer learning to enhance classification and domain adaptation to enhance cultural context understanding are future enhancements. The overall approach rectifies existing deficits in meme moderation through the synergy of text extraction, machine learning analysis, proportional enforcement action, and scalability in social media sites.

REQUIREMENTS

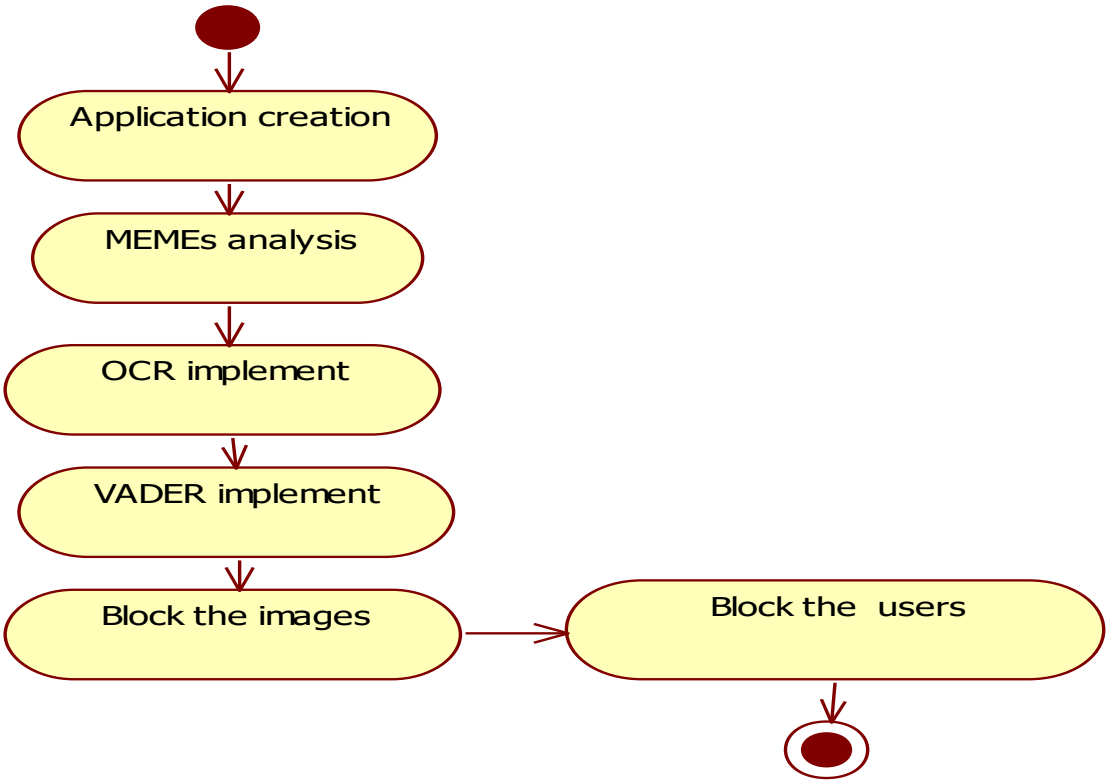
HARDWARE REQUIREMENTS

Processor-Intelprocessor2.6.0GHZ
Ram - 4 GB
Hard Disk - 160 GB
CompactDisk-650Mb
Keyboard - Standard keyboard
Monitor-15inchcolormonitor

SOFTWARE REQUIREMENT

OperatingSystem–WindowsOS
Frontend: Python
Backend:MYSQL
IDE – PYCHARM

ACTIVITY DIAGRAM



ADDITIONAL DEPENDENCIES AND CONSTRAINTS

Dependencies

The hate speech detection system based on memes relies significantly on some significant dependencies and is constrained by some design and deployment limitations. First, it relies significantly on large-scale annotated multimodal and multilingual source meme databases. The datasets must have diversified meme categories, languages, and cultural sensitivities to facilitate the successful training and testing. Design also relies on third-party libraries such as Tesseract or Google Vision for Optical Character Recognition (OCR) and Natural Language Processing libraries such as VADER, NLTK, or Spacy to detect sentiment analysis and text classification. More sophisticated machine learning models such as BERT or RoBERTa can also be employed to achieve more contextual perception with ambiguous or multilingual content. In addition, there has to be backend infrastructure in the shape of cloud storage, image processing infrastructure, and SMS APIs to enable real-time alerting and scale-out operation.

Even so, there are certain performance and reliability constraints for the system. One of them is a tradeoff between classification accuracy and real-time processing, which in fact is extremely critical in social network usage. The second one is possible dataset bias, leading to biased or erroneous classification and thus the necessity of ethical analysis and fairness checking.
Multilingual text, code-mixed text, and hierarchical nature of meme communication (sarcasm and satire) introduce contextual ambiguity.

Further Constraint

Besides the inherent limitations, there are also several other limitations of the proposed system that need to be taken into consideration in order to apply it properly and practically make use of it. One such constraint is model generalizability—in the event a system has already been trained on a specific kind or source of meme content, it will fail if it's presented with memes from another platform, sub-community, or culture. The other issue the model creates is how quickly meme shape and lexicon develop because memes draw heavily from current subjects, new slang, or neologisms and constantly retraining or updating the model to keep it up-to-date.

The system can also fail on multi-layered or visually encoded hate speech, where hateful information is symbolically encoded through symbols, fonts, emojis, or implicit image suggestions, which cannot be detected by simple OCR and NLP pipelines. There is also a regulatory and legal norm, with hate speech definition differing across geographies and jurisdictions. This compels one to be unable to build a world-model for fear of under-enforcing or over-enforcing in a foreign legal environment.

Hardware, system downtime and network reliance would render real-time scale moderation impossible in countries with undeveloped internet infrastructures. Storage constraints could arise from the amount of image data and analysis logs used to store moderation history and possible legal audit. Furthermore, resistance and backlash from users against content moderation technology are a social constraint—users may view the system as biased, intrusive, or censorious, modifying user engagement and platform trust.

Finally, there is more problematic adversarial activity. Malicious agents manually manipulate memes by distorted text, obfuscation, or creative designs in order to bypass detection algorithms. It is a cat-and-mouse that requires adversarial defense mechanisms and constant system learning for countering new threats

METHODOLOGY

1. Image Upload & Preprocessing

The process begins when users upload memes through a social network interface. The system preprocesses these images to enhance text clarity by applying noise reduction, contrast adjustment, and resolution optimization. This step ensures that the OCR algorithm can accurately detect and extract text from memes, even those with complex backgrounds or low-quality images.

2. Text Extraction Using OCR

Optical Character Recognition (OCR) converts the text within memes into machine-readable format. The system employs an OCR algorithm that identifies character boundaries, constructs a Bag of Visual Words (BoVW) framework for pattern recognition, and uses a pre-trained model to consolidate predictions. This step is crucial for accurately capturing slang, abbreviations, and culturally specific phrases commonly found in memes.

3. Text Mining & Feature Extraction

The extracted text undergoes preprocessing to refine its structure for analysis. Tokenization breaks the text into unigrams, bigrams, and n-grams, while stop word removal and stemming eliminate irrelevant terms. Special characters are filtered out, and key phrases are extracted to identify offensive language. This step ensures that the sentiment analysis module receives clean, meaningful input for accurate classification.

4. Sentiment Analysis with VADER

The VADER algorithm evaluates the emotional tone of the extracted text by calculating positive, negative, neutral, and compound sentiment scores. A compound score below 0 indicates offensive content, triggering moderation actions. VADER's context-aware approach allows it to detect sarcasm and implicit hate speech, which are prevalent in memes but often missed by traditional classifiers.

5. Moderation & User Feedback System

Based on sentiment analysis, the system enforces a tiered moderation policy. First-time offenders receive warnings, repeat offenders face image blocks, and persistent violators have their accounts suspended. SMS alerts notify users of moderation actions, ensuring transparency. The system logs all actions in a MySQL database, enabling administrators to review trends and refine detection algorithms for better accuracy over time.

CONCLUSION

The proposed system of offensive meme categorization provides a valid contribution to the newly arising issue of determining offensive content in memes on social media. By applying Optical Character Recognition (OCR) to detect text and combining it with Natural Language Processing (NLP) and sentiment analysis based on the VADER algorithm, the system effectively categorizes a meme as being offensive or non-offensive. Utilizing a progressive moderation system—warning for first-time offenders, blocking images for repeat offenders, and account suspension for habitual offenders—ensures equitable and automated content moderation. Further, integrating the system with an intuitive social network interface as well as SMS-based notification system ensures usability and transparency. Notwithstanding that the present developments include successful text extraction, classification, and moderation operations, future incorporation of transfer learning and domain adaptation will further improve accuracy on a range of meme patterns and cultural contexts. This research illustrates the power of AI-based solutions against online hate speech with scalability and effectiveness in content moderation.

REFERNCES:

[1] R. Richard and G. Giorgi, "What is a meme, technically speaking?" Inf., Commun. Soc. pp. 1–19, Feb. 2023.

[2] Z. Mansur, N. Omar, and S. Tiun, "Twitter hate speech detection: A systematic review of methods, taxonomy analysis, challenges, and opportunities," IEEE Access, vol. 11, pp. 16226–16249, 2023, doi: 10.1109/ACCESS.2023.3239375.

[3] E. K. Boahen, B. E. Bouya-Moko, F. Qamar, and C. Wang, "A deep learning approach to online social network account compromisation," IEEE Trans. Computat. Social Syst., vol. 10, no. 6, pp. 3204–3216, Dec. 2023, doi: 10.1109/TCSS.2022.3199080.

[4] K. Abbas, M. K. Hasan, A. Abbasi, U. A. Mokhtar, A. Khan, S. N. H. S. Abdullah, S. Dong, S. Islam, D. Alboaneen, and F. R. A. Ahmed, "Predicting the future popularity of academic publications

using deep learning by considering it as temporal citation networks,'' IEEE Access, vol. 11, pp. 83052–83068, 2023.

[5] H. Hosseinmardi, S. Arredondo Mattson, R. Ibn Rafiq, R. Han, Q. Lv, and S. Mishra, ''Detection of cyberbullying incidents on the Instagram social network,'' 2015, arXiv:1503.03909.

[6] H. Zhong, H. Li, A. C. Squicciarini, S. M. Rajtmajer, C. Griffin, D. J. Miller, and C. Caragea, ''Content-driven detection of cyberbullying on the Instagram social network,'' in Proc. IJCAI, vol. 16, 2016, pp. 3952–3958.
[7] Y. Chen and F. Pan, ''Multimodal detection of hateful memes by applying a vision-language pre-training model,'' Plos One, vol. 17, no. 9, 2022, Art. no. e0274300.

[8] Y. Zhou, Z. Chen, and H. Yang, ''Multimodal learning for hateful memes detection,'' in Proc. IEEE Int. Conf. Multimedia Expo. Workshops (ICMEW), Jul. 2021, pp. 1–6, doi: 10.1109/ICMEW53276.2021.9455994.

[9] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, ''The hateful memes challenge: Detecting hate speech in multimodal memes,'' 2020, arXiv:2005.04790.

[10] T. Deshpande and N. Mani, ''An interpretable approach to hateful meme detection,'' in Proc. Int. Conf. Multimodal Interact., New York, NY, USA, Oct. 2021, pp. 723–727, doi: 10.1145/3462244.3479949.

[11] A. A. Ahmed, M. K. Hasan, M. M. Jaber, S. M. Al-Ghuribi, D. H. Abd, W. Khan, A. T. Sadiq, and A. Hussain, ''Arabic text detection using rough set theory: Designing a novel approach,'' IEEE Access, vol. 11, pp. 68428–68438, 2023.

[12] A. K. Thakur, F. Ilievski, H. Sandlin, Z. Sourati, L. Luceri, R. Tommasini, and A. Mermoud, ''Multimodal and explainable internet meme classification,'' Dec. 2022, arXiv:2212.05612.

[13] A. Radford, J. Wook Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, ''Learning transferable visual models from natural language supervision,'' 2021, arXiv:2103.00020.

[14] Y. Qu, X. He, S. Pierson, M. Backes, Y. Zhang, and S. Zannettou, ''On the evolution of (Hateful) memes by means of multimodal contrastive learning,'' in Proc. IEEE Symp. Secur. Privacy (SP), San Francisco, CA, USA, May 2023, pp. 293–310, doi: 10.1109/sp46215.2023.10179315.

[15] K. Abbas, M. K. Hasan, A. Abbasi, S. Dong, T. M. Ghazal, S. N. H. S. Abdullah, A. Khan, D. Alboaneen, F. R. A. Ahmed, T. E. Ahmed,andS.Islam,''Co-evolvingpopularitypredictionintemporal bipartite networks: A heuristics based model,'' IEEE Access, vol. 11, pp. 37546–37559, 2023.

[16] A. Bhandari, ''Bias in AI: A comprehensive examination of factors and improvement strategies,'' Int. J. Comput. Sci. Eng., vol. 10, no. 6, pp. 9–14, Jun. 2023, doi: 10.14445/23488387/ijcse-v10i6p102.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, ''Learning transferable visual models from natural language supervision,'' in Proc. Int. Conf. Mach. Learn., 2021, pp. 8748–8763.

[18] J. Badour and J. A. Brown, ‘‘Hateful memes classification using machine learning,’’ in Proc. IEEE Symp. Ser. Comput. Intell. (SSCI), Orlando, FL, USA, Jun. 2021, pp. 1–8, doi: 10.1109/SSCI50451.2021.9659896.

[19] S. Prabhakaran. Cosine Similarity—Understanding the Math and How it Works (With Python Codes). Accessed: Oct. 20, 2023. [Online]. Available: https://www.machinelearningplus.com/nlp/cosine-similarity/

[20] K. E. Koech. Softmax Activation Function—How it Actually Works. Towardsdatascience.com. Accessed: Oct. 5, 2023. [Online]. Avail able: https://towardsdatascience.com/softmax-activation-function-how-it actually-works-d292d335bd78

[21] A. van den Oord, Y. Li, and O. Vinyals, ‘‘Representation learning with contrastive predictive coding,’’ 2018, arXiv:1807.03748.

[22] M. S. Hee, R. K.-W. Lee, and W.-H. Chong, ‘‘On explaining multimodal hateful meme detection models,’’ in Proc. ACM Web Conf., New York, NY, USA, Apr. 2022, pp. 3651–3655, doi: 10.1145/3485447.3512260.

[23] P.Lippe,N.Holla,S.Chandra,S.Rajamanickam,G.Antoniou,E.Shutova, and H. Yannakoudakis, ‘‘A multimodal framework for the detection of hateful memes,’’ 2020, arXiv:2012.12871.