DESIGN AND DEVELOPMENT OF FRAMEWORK FOR DETECTION OF HATE CONTENT

A Thesis submitted
in the Partial Fulfilment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in

INFORMATION TECHNOLOGY

by

Anusha Chhabra
(2K20/PHDIT/502)

Under the Supervision of
Prof. Dinesh Kumar Vishwakarma
Department of Information Technology

To the
Department of Information Technology

Delhi Technological University
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

August, 2024

ACKNOWLEDGEMENT

I am profoundly grateful to my esteemed PhD supervisor, **Prof. Dinesh Kumar Vishwakarma**, whose exceptional guidance and unwavering support have been instrumental in completing this thesis. His exemplary discipline, unyielding focus, and relentless work ethic have inspired me and set a high standard for academic excellence. I am fortunate to have benefited from his expertise and commendable commitment throughout this challenging yet rewarding journey. I am also grateful to **Prof. Girish Kumar, Mechanical Department** for his invaluable suggestions to my growth as a researcher.

No one is complete without family, who silently toil behind the scenes to help fight for the dreams. I thank my parents, **Mr. Ashok Chhabra** and **Mrs. Meera Chhabra,** for being the best parents one could hope for. I also thank my in-laws **Mr. Dharamveer Singh Grewal** and **Mrs. Indumati Grewal** for their tremendous support in taking care of my two kids **Siya Grewal** and **Aman Grewal**. I specially thank to my spouse **Mr. Paras Grewal** for being the best life partner and supporter throughout my journey. I am also grateful for unwavering support of my brother **Mr. Rahul Chhabra** and sisters-in-law **Ms. Kirti Ahlawat, Ms. Neetu Verma**, and **Ms. Ritika Chhabra**.

Finishing PhD is a highly challenging journey, and the seniors who helped navigate this path need a special mention. To this end, I am extremely grateful for the support of my PhD senior, **Dr. Chhavi Dhiman, Dr. Deepika Varhsney,** and **Dr. Ankit Yadav.**

As I went through the most challenging phase of my PhD, my lab mates ensured that I never gave up and always came back stronger. I am grateful to **Dr. Deepak Dagar, Dr. Ananya Pandey, Dr. Ashish Bajaj, Abhishek Verma** and **Bhavana Verma** for all the light-hearted conversations.

I extend my heartfelt appreciation to the state-of-the-art research lab established by my supervisor. Equipped with cutting-edge NVIDIA GPUs, it was pivotal in facilitating the success of the computationally expensive deep learning-based research experiments throughout my PhD.

Last but not least, I thank God for giving me the persistence and strength to show up at my lab each day and work through the ups and downs of this PhD journey.

Anusha Chhabra
2K20/PHDIT/502

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I certify that the dissertation titled "***Design and Development of Framework for Detection of Hate Content***" that I am submitting for the Doctor of Philosophy degree is solely my own and has not been previously submitted for any other degree or certification from any other academic institution. The research conducted in this thesis is original and has been independently carried out by me duringthe period from 13/01/2021 to 27/08/2024 under the guidance of Prof. Dinesh Kumar Vishwakarma.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Anusha Chhabra
2K20/PHDIT/502

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR

Certified that Anusha Chhabra (2K20/PHDIT/502) has carried out their research work presented in this thesis entitled “Design and Development of Framework for Detection of Hate Content” for the award of Doctor of Philosophy from Department of Information Technology, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Signature:

Name of the Supervisor: Prof. Dinesh Kumar Vishwakarma
Designation: Professor**;** Department of Information Technology
Address: Rohini, Delhi

Date: 09/12/2024

ABSTRACT

Due to the widespread usage of social media platforms, an alarming problem has emerged in an era characterised by the rapid spread of hateful contents in any of the multimedia formats. The combination of the simplicity and complexity of these innovations presents a substantial risk to the clean and reliable conversations. This thesis emphasises the necessity to create novel systems for detecting hate content by using the potential of machine learning and deep learning techniques. The susceptibility of multimodal content towards hatefulness has significantly increased, reaching unprecedented levels. This results in implementing modern technologies that facilitate the production of counterfeits with a high degree of authenticity. The objective of this study is to leverage the capabilities of machine and deep learning to identify and mitigate hateful content effectively. Given that social media platforms are the main channels for sharing information, the suggested detection systems utilising machine learning and deep learning aim to ensure the mitigation of hate content detection. As a result, this will enhance the establishment of a digital ecosystem characterised by increased reliability and credibility. This thesis tackles this detection challenge by proposing four novel architectures.

The first two techniques are dedicated to the problem of tacking the textual hate content in an efficient manner**.** In the first approach, it is seen that reducing features using Truncated SVD along with hyper parameter tuning helped in increasing balanced accuracy and F1 score for algorithms like Logistic Regression, SVM and XGBoost when compared to the baseline results. Still, the proposed approach is lacking in handling uncertain or imprecise data. The second approach focuses on handling the uncertainty and vagueness in the data by implementing the fuzzy classifiers. An empirical evaluation of seven classifiers is presented for hate speech detection on two commonly used benchmarks of different data characteristics providing essential insights into their detection in terms of accuracy for their deployment in real-world applications. Fuzzy classifiers outperformed the other two classifiers out of the three.

Next two models are dedicated to the multimodal hate content detection. The first framework presents a dual-branch network which is composed of knowledge distillation attention for extracting the essential information from the caption modality and multi-kernel attention for collecting pertinent information from the images. Extensive testing on three publicly accessible datasets showed that the suggested architecture outperformed baseline models, claiming better results in terms of accuracy and AUC scores. Numerous ablation trials on the available multimodal datasets are conducted to conclude that the proposed architecture is contributing to the performance of hate content identification in memes. The second proposed model “MHS-STMA” explored the problem of learning complementary information between multimodal data. The architecture utilizes transformers for capturing the dependencies and relationships between the elements in a sequence. The proposed architecture also utilizes attention mechanisms at multiple levels and focuses on crucial regions in the images based on the attended textual features. Self-attention mechanism is implied at the end to remove any redundancy from the multimodal data. The experimental results conducted on three popular datasets show that our method performs efficiently.

Lastly, A novel robust approach MHM-HGraph is proposed to effectively capture the contextual dependencies within two modalities (Visual and Text). To better capture the underlying patterns within the data, this model makes use of hypergraph convolution layers to investigate the application of non-local information, identifying high-order correlations on hypergraph, and exploit the “enhancement connection” to perform non-linear mapping on the features.

In conclusion, this thesis presents substantial discoveries and identifies potential areas for future research on the subject of identifying hate content detection.

LIST OF PUBLICATIONS

Publications Arising from Research Work in this Thesis

1) **Anusha Chhabra**, Dinesh Kumar Vishwakarma. "A Literature Survey on Multimodal and Multilingual Automatic Hate Speech Identification." Published in **Multimedia Systems**, 2023, (Pub: Springer): **DOI: https://doi.org/10.1007/s00530-023-01051-8**.

2) **Anusha Chhabra**, Dinesh Kumar Vishwakarma. "Multimodal Hate Speech Detection via Multi-Scale Visual Kernels and Knowledge Distillation Architecture." Published in **Engineering Applications of Artificial Intelligence** (126), 2023, (Pub: Elsevier), **DOI: https://doi.org/10.1016/j.engappai.2023.106991**.

3) **Anusha Chhabra**, Dinesh Kumar Vishwakarma. "Hate Content Detection via Novel Pre- Processing Sequencing and Ensemble Methods." **https://arxiv.org/submit/5841063/view**

4) **Anusha Chhabra**, Dinesh Kumar Vishwakarma. "MHS-STMA: Multimodal Hate Speech Detection via Scalable Transformer- Based Multilevel Attention Framework" **https://arxiv.org/submit/5841081/view**

5) **Anusha Chhabra**, Dinesh Kumar Vishwakarma. “Fuzzy and Machine Learning Classifiers for Hate Content Detection: A Comparative Analysis.” **IEEE Conference: 4th International Conference on Artificial Intelligence and Speech Technology (AIST)**, *IGDTUW, Delhi. (2022) [**Presented**],* **DOI: 10.1109/AIST55798.2022.10064822**

6) **Anusha Chhabra**, Dinesh Kumar Vishwakarma. “A Truncated SVD Framework for Online Hate Speech Detection on the ETHOS Dataset.” **IEEE Conference: 4th International Conference on Innovative Trends in Information Technology (ICIIIT)**, *IIIT Kottayam, Kerala. (2023) [**Presented**],* **DOI: 10.1109/ICITIIT57246.2023.10068574**

7) **Anusha Chhabra**, Dinesh Kumar Vishwakarma. “Multimodal Hate Speech Identification in Memes Based on Hypergraph.” **IEEE Conference: International Conferences on Signal Processing and Advance Research in Computing (SPARC-2024), Amity School of Engineering and Technology, Amity University, Lucknow, UP, India. https://arxiv.org/submit/5841006/view**

Publications Arising from Research Work Outside this Thesis

SCIE Journal Papers

Sajal Aggarwal, **Anusha Chhabra**, Dinesh Kumar Vishwakarma. "Multimodal Hateful Meme Detection using Knowledge Distillation and Hierarchical Vision Transformer Framework." *Communicated* in **Engineering Applications of Artificial Intelligence (1st Revision- Under Review)** (Pub: Elsevier).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

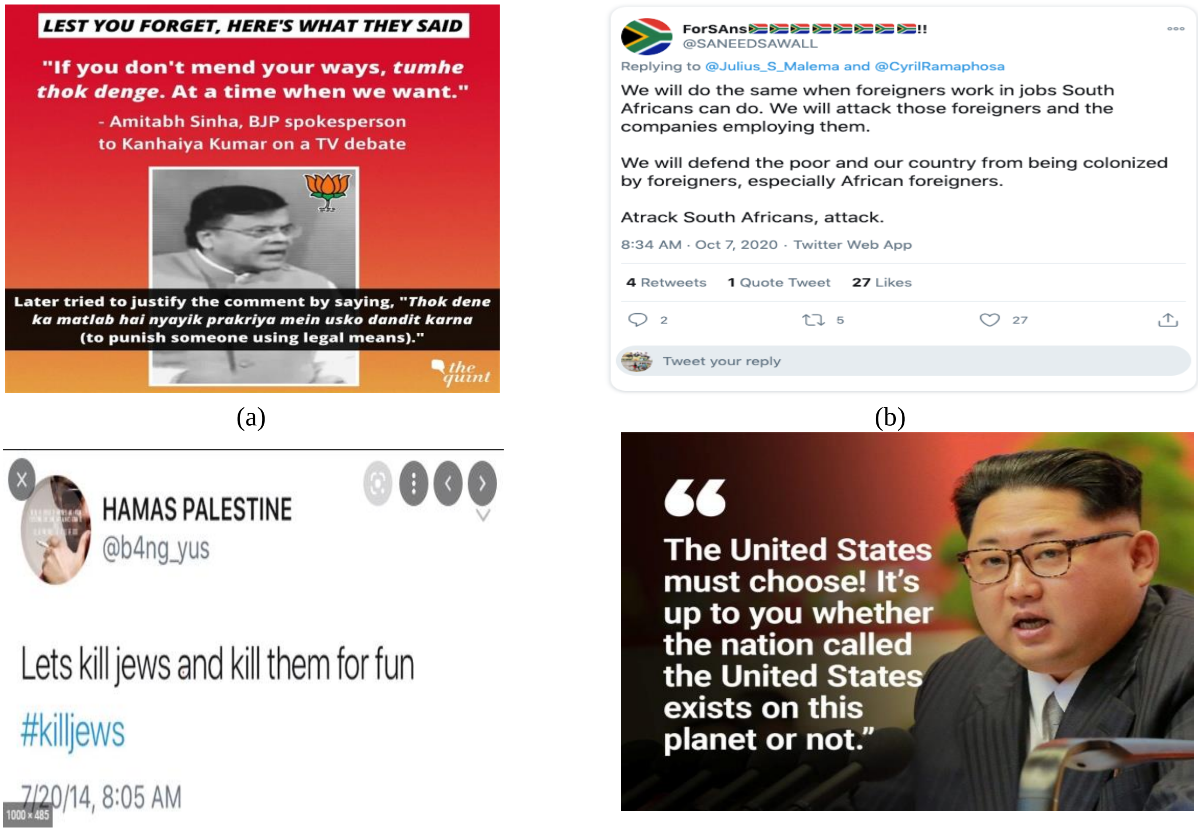
Social media is a more prevalent, common and powerful platform for communication to share views about any topic or article, which consequently leads to unstructured toxic, and hateful conversations. With the growing internet and technology, a large amount of information content is present on online community networks as multimodal data (Text, Pictures, and videos). The last decade has witnessed a tremendous rise in social media platforms. An extensive online presence has become a normal part of daily human lives. The number of active users on social media has grown tremendously, from just over 2 million active users at the beginning of 2015 to almost 4 million active users by the end of 2020 [1]. Also, the average person had about 8.6 social media accounts in 2020 [1]. It is an understatement to say that social media has become integral to everyday life. The importance of social media is discussed as follows:

- Social media connects people together.
- Social media provides a platform for sharing information, exchanging ideas, expressing opinions, etc.
- Social media also attracts a large number of passive information consumers. Users create and share multimedia data and view and explore data shared by other community users, group, organization, etc.
- Social media has an enormous impact on individuals' mental and emotional states.

1.1Hate Speech Examples

If we look at the statistics, we can visualize that a large number of people using social media is escalating at a very great speed, and people can easily present their views to each other via various social media platforms. The type of content on the online social media stages contributes to the propagation of hate speech and misleading people. Right now, controlling this kind of media information is very important. Therefore, hate speeches harm individuals and impact society by raising hostility, terrorist attacks, child pornography, etc. **Fig. 1.1** shows a portion of hate speech and offensive expressions posted on social media or the web. **Fig. 1.1(a)** shows a clear example of encouraging violence during huge fights against CAA, NRC, and NPR across India in Jan 2020 [2]. **Fig. 1.1(b)** shows the tweet released under #putsouthafricansfirst, a person openly tweeting to attack the foreigners working in South

Africa. **Fig. 1.1(c)** shows a tweet posted in 2014 advocating killing Jewish people for fun after the synagogue shooting in Pittsburg [3]. **Fig. 1.1(d)** shows a post posted in Jan 2018 that a supreme leader is giving a genuine threat statement to the US for war [4].



LEST YOU FORGET, HERE'S WHAT THEY SAID

"If you don't mend your ways, *tumhe thok denge*. At a time when we want."

- Amitabh Sinha, BJP spokesperson to Kanhaiya Kumar on a TV debate

Later tried to justify the comment by saying, "*Thok dene ka matlab hai nyayik prakriya mein usko dandit karna* (to punish someone using legal means)."

(a)

ForSAns🇿🇦🇿🇦🇿🇦🇿🇦🇿🇦🇿🇦🇿🇦🇿🇦!!
@SANEEDSAWALL

Replying to @Julius_S_Malema and @CyrilRamaphosa

We will do the same when foreigners work in jobs South Africans can do. We will attack those foreigners and the companies employing them.

We will defend the poor and our country from being colonized by foreigners, especially African foreigners.

Atrack South Africans, attack.

8:34 AM · Oct 7, 2020 · Twitter Web App

4 Retweets 1 Quote Tweet 27 Likes

(b)

HAMAS PALESTINE
@b4ng_yus

Lets kill jews and kill them for fun
#killjews

7/20/14, 8:05 AM

"The United States must choose! It's up to you whether the nation called the United States exists on this planet or not."

(c) (d)

Fig. 1.1 Examples of hate speech and offensive expressions present over social media

The recent instances of high-profile politicians making speeches were an apparent attempt at inciting violence, which led to large-scale violence. These instances are yet to be dealt with by law enforcement agencies. Hence, the integrity of identifying hate instances is one of the most significant challenges in social media stages, and research-based analysis of this type of content is necessary. The following section describes the definition analysis of hate speech from various sources.

1.2 Hate Speech: Definition Perspective and Analysis

There is a general agreement among researchers to define hate speech, and researchers have described it as a language that attacks an individual or a society dependent on characteristics like race, shading, nationality, sex, or religion [5]. This section provides some state-of-the-art definitions of hate speech (**Table 1.1**). Although many authors and social media platforms have given their purposes for hate speech, researchers are following them to understand the forms and classifications of hate speech. The definitions from the various sources are as follows:

- Some of the scientific definitions include the community's perspective.
- Major social networking sites like Facebook, YouTube, and Twitter are the most used platforms where hate speech occurs regularly.

Table 1.1 Some of the prominent definitions by some state-of-the-art

References	Hate Speech Definitions
[6]	An antagonistic, malevolent speech focused on an individual or a social event of people taking into account a part of their genuine or intrinsic qualities. It communicates unfair, scary, objecting, hostile, or potentially biased perspectives toward those attributes, including sex, race, religion, identity, shading, public beginning, incapacity, or sexual direction.
[7]	Hate Speech is a conscious and hardheaded public assertion expected to slander a gathering of individuals.
[8]	Hate Speech is a quick attack on individuals subject to race, identity, sex, character, and veritable sickness or impediment. We portray assail as horrible or dehumanizing talk, clarifications of deficiency, or calls for dismissal or seclusion.
[9]	Hate speech alludes to content that advances viciousness or scorn against the public dependent on specific ascribes, like ethnic or race beginning, religion, inability, sex, age, veteran reputation, and sexual direction/sex personality.
[10]	Content that attacks people based on actual or perceptual race, ethnicity, country of origin, religion, gender, sexual orientation, disability, or illness is not permitted. It is considered a potential threat or attack for the content that many people find offensive (jokes, Stand-up comedy, lyrics of popular songs, etc.
[11]	Hate speech attacks an individual or get-together depending on characteristics like religion, race, ethnicity, insufficiency, sexual heading, or sex character.
[12]	Hate Speech attack others dependent on racism, ethnicity, public start, sexual bearing, sex, character, age, handicap, or genuine illness.
[13]	The language used to convey hate speech towards a selected bunch.
[14]	Hate Speech is a purposeful attack on a specific social occasion of people motivated by the pieces of the group's character.
[15]	Hate Speech assails or prompts malignance against gatherings in light of explicit qualities like actual looks, religion, ethnicity, sexism, and many more. Moreover, individuals with diverse phonetic styles in unobtrusive construction can happen.

The definition analysis **(Table 1.2)** mainly relies on various sources like multiple definitions from scientific papers and powerful social media platforms. The dimensions used for analysis are “ violence,” “attack,” “ specific targets,” and “ status.”

Table 1.2 Definition Analysis

Ref.	**Dimensions**			
	Specific Targets	**Status**	**Violence**	**Attack**
[6]	Yes	No	Yes	No
[7]	Yes	No	Yes	No
[8]	Yes	Yes	No	Yes
[9]	Yes	No	Yes	No
[10]	Yes	Yes	Yes	No
[11]	Yes	No	No	Yes
[12]	Yes	No	No	Yes
[13]	Yes	No	Yes	No

Ref.	Dimensions			
	Specific Targets	Status	Violence	Attack
[14]	Yes	Yes	Yes	Yes
[15]	Yes	Yes	Yes	No

After a thorough definition analysis, we have also portrayed the definition of Hate Speech as follows:

"Hate Speech is a toxic speech attack on a person's individuality and likely to result in violence when targeted against groups based on specific grounds like religion, race, place of birth, language, residence, caste, community, etc."

1.3Hate Speech: Forms and Related Words

Fig. 1.2 shows significant hate forms of speech like Cyberbullying, Toxicity, Flaming, Abusive Language, Profanity, Discrimination, etc., and **Table 1.3** presents the definitions of the above forms of hate speech found in the literature with their distinction from hate speech.

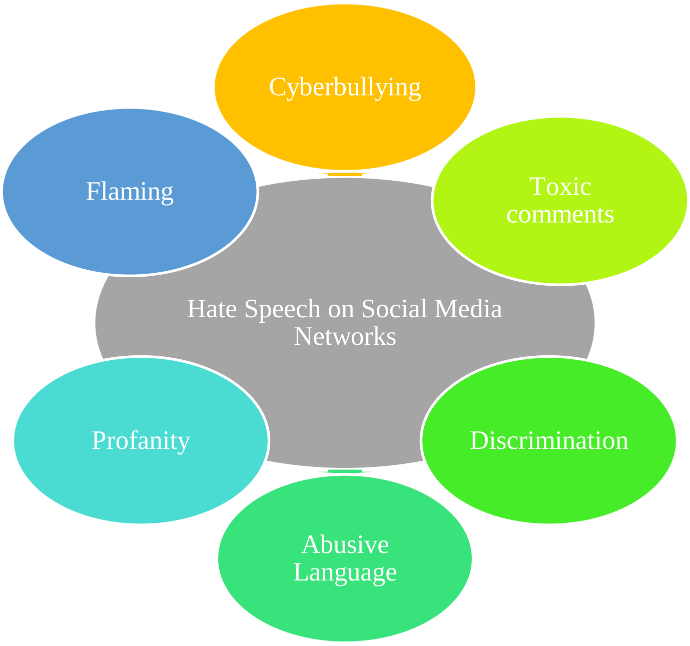


Fig. 1.2 Forms of Hate Speech

Table 1.3 Comparison between Hate Speech and its various forms

Forms	Definitions of forms	Distinction from hate speech
Cyberbullying	Characterized as a deliberate demonstration completed by a social occasion or individual using electronic stages [15].	Hate speech is abusive speech explicitly directed toward a unique, non-controllable attribute of a group of people.
Discrimination	Interaction via a distinction and afterward utilized as the premise of unreasonable treatment [16].	Hate speech is a virulent form of discrimination.
Flaming	Flaming describes antagonistic, profane, and threatening remarks that can upset and offend	Unlike flaming, hate speech can occur in any context.

Forms	Definitions of forms	Distinction from hate speech
	other members of the forums, generally called trolls [17].	
Abusive Language	The term abusive language seeks to diminish or humiliate some person or group [18].	Hate Speech is a type of abusive language.
Profanity	Hostile or indecent words or expressions.	Hate speech can use profane words but not always.
Toxic language	Conveying content that is disrespectful, abusive, unpleasant, and harmful [19].	Not all toxic comments contain hate speech.

Hence, analyzing hate speech on the web is one of the critical areas to study due to the following reasons:

- Reduce conflicts and disputes created among human beings due to toxic language and offensive expressions.
- The broad availability and notoriety of online web-based media, like Facebook, Twitter, Instagram, web journals, microblogs, assessment sharing sites, and YouTube, boost communication and allow people to freely share information in the form of their thoughts, emotions, and feelings among strangers.
- Moreover, Click baiting takes massive attention and encourages visitors to click on the link, harming readers' emotions.
- Hate speeches can incite violence and cause irreparable loss of life and money.
- The latest incident was triggered by online hate speech in the Philippines, citing the example of the Christchurch mosque shooting in 2019 [20].
- To forestall bigot and xenophobic viciousness and separation spread among Asians and individuals of Asian drop uniquely in this pandemic. As per the report distributed by US Today in May 2021, more than 6600 hate and offensive incidences against Asian-Americans and Asians have been accounted for [7].
- To save our society from being gravely damaged.

From the points mentioned above, it has been observed that detecting and restraining hate speech at an initial stage is very crucial and, indeed, a challenging task. Major online media stages like Facebook, Twitter, and YouTube are trying to eliminate hate speeches and other harmful content at an initial step as part of their ongoing projects, using advanced AI techniques. However, keeping an eye on an individual is vital to have hate off platforms. Social media platforms and an individual can adopt the following suggestions:

- The most significant source of hate speech on the internet is trolls. A person should block, mute, or report these trolls instead of giving recognition.
- A person should do a proper data analysis and facts before forwarding the posts.
- Social media firms should follow strong policy rules against abusive behavior.

1.4 General Framework of Hate Speech Detection

Fig. 1.3 provides a framework for the process of hate speech identification. The foremost step is to search the powerful source platform where most hate speech/ offensive languages occur. Most state-of-the-art adopted significant social media firms like Facebook and Twitter. The second step is to collect data either in the form of posts or tweets. Gathering a great measure of information from web-based media stages nowadays is one of the significant research challenges for researchers and academia. The platforms provide a simple and quick approach to gathering and storing information through inbuilt APIs [21].

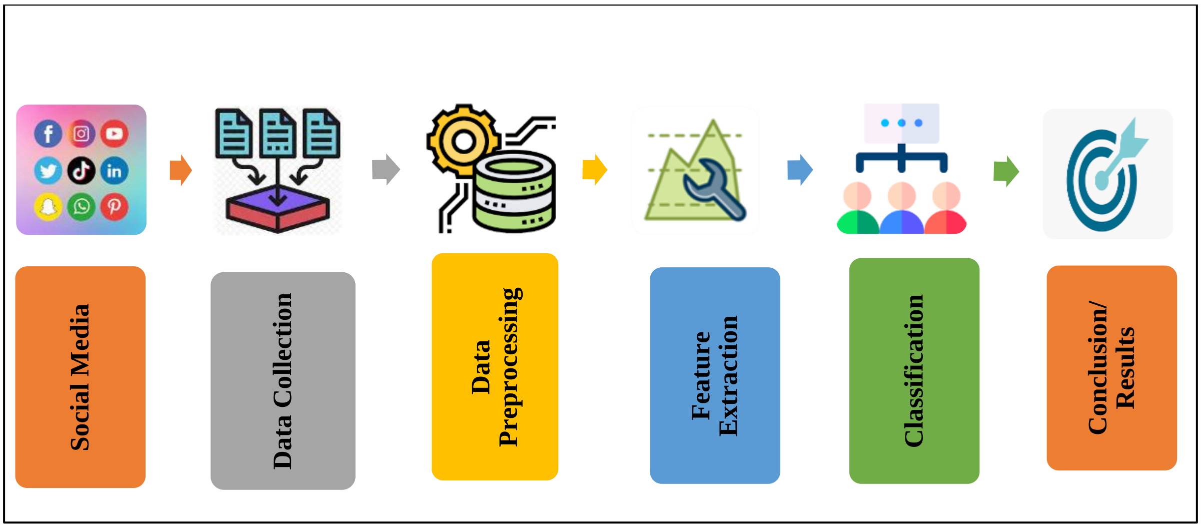


Fig. 1.3 General Framework of Hate Speech Detection

A large amount of hate speech data collection is from two powerful social media platforms: Twitter, Instagram, Facebook, and these two platforms are actively working on combating hate speeches. The next phase includes data normalization and feature extraction for training a model, and the last step performs classification to classify the problem.

1.5Motivations for Detection of Hate Content

Recently, more users are actively participating on social media in the form of WhatsApp posts, Facebook updates, YouTube videos, reviews, and comments, among other forms, on various themes. People are sharing their opinions, which has resulted in a vast volume of data online. The information needs to be examined for future study. The data from the last five years has been considered for visualization and motivation behind conducting this research. **Fig. 1.4** illustrates the count of hate content publications from Web of Science (WoS) and Scopus databases.

From **Fig. 1.4**, it is observed that there is a tremendous amount of hate content generation on social media, giving the motivation to conduct research in this field. Moreover, the publications on multimodal hate speech detection are very few. If we take the data from WoS and Scopus databases, the number of publications via keyword search as "multimodal hate speech" is 1% and 6.6%, respectively. Nowadays, people communicate via memes more often than texts, emphasizing authors' focus on multimodal data.

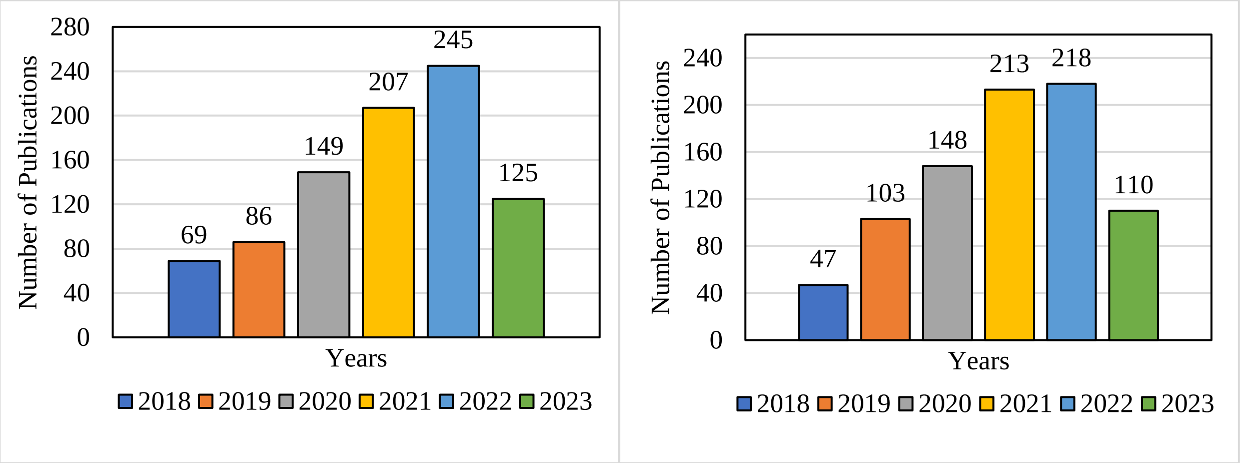


Fig. 1.4 depicts the number of publications in hate content identification over the last five years taken from (a) Web of Science and (b) Scopus database

1.6Challenges

Different categories of images, texts, or combining two could create various memes. It can be a benign text confounder, benign image confounder, contrapositive, or counterfactual meme (refer to **Fig. 1.5**). **Fig. 1.5 (a)** stands for misleading memes. The left image depicts a hateful meme, whereas the other two illustrate its confounders, changing its label and resulting in a not-hateful meme. The confounding meme can be generated by changing the original meme's image or label. **Fig. 1.5 (b)** depicts contrapositive, and **Fig. 1.5 (c)** represents counterfactual meme. The challenge lies in the diversities of these types of memes which can contain objects that the existing classifiers cannot identify the result. The solution is essential via incorporating additional information from different sources to improve classification accuracy. An optical character recognition (OCR) technique can be applied to extract the linguistic part of the meme, and object detection methods can be utilized to encode the correlated image part of the meme. Aspects of several modalities have been combined at three levels in prior research to solve this issue: early fusion, late fusion, and hybrid fusion. Late fusion takes place at the decision or score level; early fusion happens at the feature level, and hybrid fusion mixes the two.

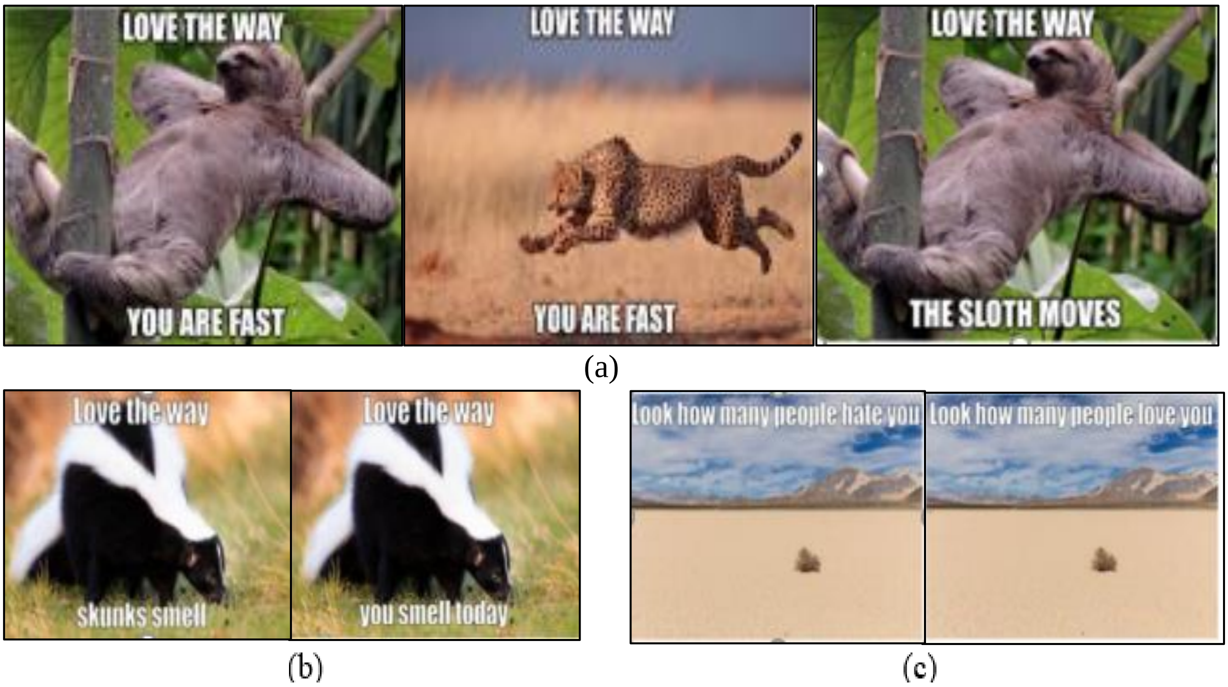


Fig. 1.5 (a) Samples of Text and Image confounders. Left Image of (a) shows a hateful meme. In contrast, the middle and right images of (a) illustrate its confounders via changing text or image to a non-hateful meme, (b) and (c) represents a contrapositive meme

One of the most important research issues is combining text and visual modalities. [22] asserts that examining a post's multimodal content yields better outcomes than processing it separately. In feature-level fusion, the features gathered independently for each modality are fused and input into the classification model [23]. A late fusion technique, on the other hand, is when the characteristics obtained for each data mode are provided to several classification models, after which the overall scores of numerous models are combined [[24], [25]]. Early fusion may not accurately capture the tightly coupled correlation among the pertinent modalities, and decision-level fusion may not accurately portray the interplay between several modalities. To address this problem, attention-based intermediate fusion models were created [[26], [27]]. The proposed approach has significant difficulty sustaining the validity of meaning while bridging the relationship between modalities. Although early and late fusion is one of the primary methods for connecting various modalities meaningfully, it does not appear straightforward. A study [28] found that late fusion outperforms early fusion in integrating text, audio, and facial expression features.

*The following research works form the basis of this chapter***:**

- ❖ **Anusha Chhabra**, Dinesh Kumar Vishwakarma. "A Literature Survey on Multimodal and Multilingual Automatic Hate Speech Identification." Published in **Multimedia Systems**, 2023, (Pub: Springer): **DOI:** https://doi.org/10.1007/s00530-023-01051-8.

1.7 Thesis Overview

The remaining section of the dissertation is structured in the following manner.

- **Chapter 2** The literature review examines the existing state-of-the-art techniques for multimodal hate content detection.
- **Chapter 3** Describes the most potent methodologies for detecting textual hate content.
- **Chapter 4** Describes the detection of hate content in multimedia data.
- **Chapter 5** dives into the robust analysis of multimodal hate content.
- **Chapter 6** summarizes the conclusions inferred from this research work and highlights the potential future work in this area.

CHAPTER 2
LITERATURE REVIEW

This chapter reviews the literature review of work done in detecting hate speech considering textual, visual, and multimodal aspects.

2.1Feature Extraction Techniques in Automatic Hate Speech Detection

A feature is the closed characteristics of an entity or a phenomenon. [29] focus on natural language processing (NLP) to explore the automation of understanding human emotions from texts. Word references and lexicons are the most straightforward and basic approaches for feature extraction in text analysis. Identifying the appropriate features for classification is more tedious when using machine learning. The fundamental step in traditional and deep learning models is tokenization, in which the primary and straightforward approach is dictionaries/ lexicons. Dictionary is a method that generates a set of words to be looked at and included in the text. Frequencies of terms are used directly as features. Features play an essential role in machine learning models. Machine learning approaches cannot work on raw data, so feature extraction techniques are needed to convert text into vectors of features. Many basic features like BOW, Term Frequency- inverted Term Frequency, Word references, etc., are used.

2.1.1 Bag-of-Words (BOW)

BOW is an approach like word references extensively used for document classification [[14], [30], [31]] The frequency of each word is used as a characteristic for training a classifier after gathering all the words. The burden of this technique is that the sequencing of words is disregarded, whether it is syntactic or semantic information. Both pieces of information are crucial in detecting hate content. [32] used BOW to represent Arabic hate features as text pre-processing before applying various machine learning classifiers. [33] derived a method for detecting Arabic religious hate speech using different features with the machine and deep learning models. Consequently, it can prompt misclassification of whether the terms are utilized in multiple contexts. N-grams were executed to overcome the issue.

2.1.2 N-grams

The N-grams approach is the most utilized procedure in identifying hate speech and offensive language [[30], [18], [34], [35], [36], [37], [38]]. The most widely recognized N-grams approach combines the words in sequence into size N records. The objective is to enumerate all size N expressions and check their events. It further increases the performance of all classifiers since it incorporates each word context [39]. Rather than utilizing words, it is additionally conceivable to use the N-grams approach along with characters. [40] proved character N-gram features are more predictive in detecting hate speech than token N-gram features, whereas it is not valid in the case of identifying offensive language. Although N-grams also have limitations, like all the related words have maximum distance in a sentence [30], an answer for this issue lies in incrementing the N value. However, it lowers the processing speed [41]. [35] proved that greater N values perform better than lower N-values (unigrams and trigrams). The authors in [[5], [34]] observed that character N-gram features perform better when combined with extra-linguistic features. The authors generated one hot N-gram and N-gram embedding feature to train the model and analyzed better performance by N-gram embedding [42].

2.1.3 Lexicon-based and Sentiment based

Lexical features use unigrams, and bigrams of the target word, whereas syntactic features include POS tags and various components from a parse tree. The parser used in NLP, proposed by the Stanford NLP Group [43], was used to catch the linguistic conditions inside a sentence [41]. Lexicon-based methods are crucial in identifying the sentiments of speech. For example, nigga is an offensive word and must be prohibited in ordinary language [44]. Hateful speech on a social stage cannot be a positive polarity because awful grammar provides a negative inclination by the speaker to the listeners and readers. Authors in [[35], [45], [46], [37], [47], [48], [49]] consider sentiments as a characteristic for identifying hate speech. Some authors [35] used the sentiment features in combination with others, which proved in result enhancement. [50] presents metaheuristic approach for sentiment analysis and proved that the optimization methods can be alternatively used against machine leaning models with promising results.

2.1.4 Topic Modeling

This method is also famous for topic classification, which focuses on extracting topics that occur in a corpus. Topic modeling is also used for detecting hateful comments from central social media platforms like Youtube [[51], [52]] used the Latent Dirichlet Allocation model [53] to discover abstract topics and use them in classifying multimodal data. [54] derived text clusters from LDA for multilingual hate speech detection and proved that topic modeling is not giving any major incite for classification.

2.1.5 TF-IDF

TF-IDF is a scoring measure broadly used in information retrieval and is planned to reflect how important a term is in a given record. TF-IDF is the most common feature extraction technique used by traditional classification methods for hate speech identification [[55], [56]]. TF-IDF differs from a bag of words technique or N-gram technique because the word recurrence offsets the frequency of each term in the corpus, which clarifies that a few words show up more often than expected (for example, stop words). [57] used N-grams and TF-IDF values to perform a comparative analysis of the machine learning models to detect hate speech and offensive language and claimed that the L2 normalization of TF-IDF outperforms the baseline results.

2.1.6 Part-of-speech

POS tagging is a well-known task in NLP. This approach refers to the technique of classifying words into their parts of speech. Moreover, it improves the value of the context and identifies the word's role in the context of a sentence [58]. Some authors [36] used this approach to classify racist text. PoS tagging with TF-IDF gives a better result in Indonesian Hate Speech Detection [59].

2.1.7 Word Embedding

The most widely recognized technique in text analysis of hate content is the utilization of word references. This methodology comprises all words (the word reference) that are looked at and included in the message. The frequencies are utilized straightforwardly as features and for calculating scores. In NLP, Word embedding is used for representing of words while performing text analysis. [60] uses word2vec embedding for extracting hate content features for grouping the semantically related words. [61] applies attention based neural networks and word embedding feature extraction methods for classification. Hate speech detection in Spanish language [62] uses word embedding methods like Word2Vec, Glove, FasText for feature extraction. Another procedure used in text analysis of hate content is the distance metric, which can be used to supplement word reference-based methodologies. A few investigations have called attention when the negative words are obscured with a purposeful incorrect spelling [63]. Instances of these terms are @ss, sh1t [18], nagger, or homophones, for example, joo [63].

2.1.8 Rule-Based Approach

Text analysis uses a rule-based feature selection technique for finding the regularities in data, for example, IF-THEN clauses. [64] Proves that rule-based methods do not include learning but depends on word reference of subjectivity pieces of information. This particular approach is used to extract subjective sentences to generate hate content classifiers for unlabeled corpus [46]. [65] works on the combination of dictionary-based classifiers along with rule-based classifiers to generate the semantic features for hate speech classification.

2.2 Textual Aspect

The natural language processing (NLP) branch is working to close the comprehension and understanding gap between human languages and computers. Specifically, the efficiency and effectiveness of deep learning architectures have made extensive advancement in the areas like sentiment analysis [66], atomistic simulation [67], self-monitoring systems, object detection, life prediction analysis, online education like music teaching [68], classification, filtering, language translation, etc., covering diverse languages. With the ease of social media and the vast use English language, hate speech is also flowing on the web in various regional languages like Urdu, Portuguese, Bengali, Hindi, etc. Hindi-English code-switched language models for efficient text representation are encountered by [69]. The study [70] concentrates on brief sentences since learning-sufficient qualities in news information are lacking. More research is needed to find the augmented features using a web search for longer material. The lengthy text may be divided into shorter sentences to acquire probabilistic pre-decision. By combining the outcomes of the pre-decision, the final decision regarding the class can be deduced. [71] illustrates that traditional features collected from news articles outperform previous models built using text embedding approaches. Most online publications in this field are available in a single modality, i.e., Text only, and most of the work has been done in English only [72]. Traditional machine learning algorithms rely on feature engineering, making the process complex and time-consuming [73]. The performance analysis is measured in [74] using various traditional learning, deep learning, and fuzzy logic classifiers for hate speech detection, showing the outstanding performance of fuzzy classifiers. In hate content identification, the Support Vector Machine (SVM) classifier includes manually extracted text features to determine whether the provided content is abusive [63]. N-gram characteristics were employed by [75] to categorize whether or not the speech was abusive. [75] have classified whether the speech is abusive using n-gram features. There are many text-based datasets available for hostility recognizable proof [76], the recognition of hate speech [37], and the identification of offensive language [77] The relevancy of cross-dataset and cross-lingual generalization in this area is remarked by [[78], [79]]. Authors in [76] work on unigrams and examples of the text for recognizing hate content. After being meticulously crafted by hand, these patterns are given to machine learning models for further classification. [77] addresses some of the difficulties in identifying offensive content and the classification of hateful tweets in German. This research emphasizes features specific to a single modality, such as text and manual feature extraction. Nowadays, the advancement in NLP has introduced pre-trained linguistic models like BERT [80], and its variants like RoBERTa [81], DistilBERT [82] and XLNet [83] are widely used in hate content classification. BERT uses large amounts of unlabeled data to build models with adjustable parameters as needed for smaller amounts of supervised data to enhance performance [84]. The variants of BERT are assessed for hate content classification incorporating characters with subword embedding [85]. Authors in [86] designed a framework for detecting hate content by combining DNNs with static BERT embedding to extract contextual information better.

2.3 Visual Aspect

Due to the tremendous data increase in various modalities, deep learning models are used. We work with memes spanning text and images, and deep learning algorithms automatically extract features from these memes. The research till now majorly relies on object detection where only images are considered [87]. CNN is frequently employed in machine vision tasks as it provides the benefit of processing pixels in images [88]. It has also been acknowledged that error gradients in deep networks or recurrent neural networks can build and result in very large gradients during the training phase. Furthermore, the benefits of successfully avoiding gradient explosion are investigated using a bidirectional long- and short-term memory (Bi-LSTM) neural network fed with the CNN characteristics [89]. In image-based hate content detection, Convolution Neural Networks (CNNs) models are the most prominent in identifying offensive content in the form of nudity [[90], [91], [92]], appropriate or in- appropriate images for children [93], offensive/ non- compliant logos [93], pornographic web pages [94].

2.4 Multimodal Aspect

The related work till now focuses on unimodal aspects, i.e., either text or image individually. Now a day, Information in the form of internet memes is the most common on social media platforms, and they form the text- accompanied images. A shared task on the analysis of memotion was already available in SemEval2020 [77]. Very few datasets are available in this aspect [72]. The reliability and robustness of the algorithms should be taken into account prior to working on multimodal data [95]. [96] designed a popular Facebook hateful memes challenge dataset in May 2020 and obtained a human accuracy score of 84.70% after the annotation process. [96] also applied a VilBERT [97] on conceptual caption [98], providing a winning solution by claiming test accuracy of 69.47% and an AUC score of 75.44%. Authors in [99] implemented R-CNN for the visual branch and BERT for a textual branch on Facebook hateful memes dataset, giving an accuracy score of 75.80% and AUC of 82.80%. An impressive accuracy score of 71.08% on [96] is also claimed by [100] using BERT and Xception models. A dataset of about 150K images by the name "MMHS150K" [101] is the largest dataset available publicly on hate speech. They tried to build a model which can differentiate between publications that use offensive language, and that target particular communities. They, therefore, experimented with utilizing hate scores instead of binary labels for each tweet, but the data showed no significant differences between the various labels of hatred in the experimental part. Authors in [101] used the Feature concatenation model, textual kernel model, and spatial concatenation model to boost the performance with an accuracy of 68.50%. A dataset with only 743 offensive memes is also publicly available and implemented stacked LSTM and VGG16 for evaluating precision, recall, and F-score [87]. [99] applied R-CNN and BERT on the MultiOff dataset, calculating precision, recall, and F1-score as 64.50%, 65.10%, and 64.60% respectively. Currently, sarcasm is also a developing research area in relation to sentiment analysis, emoji detection, and hate/ offensive content. Authors in [102] also focus primarily on sarcasm detection using a multimodal attention mechanism followed by emoji and sentiment analysis [66].

2.5 Research Gaps

On the basis of the literature presented in the above section, several research gaps have been identified.

- It has been observed that there is a scarcity in multimodal hate content detection.
- From the last few years, authors are focusing on multilingual hate speech identification by creating own datasets, but the authors do not publish those, which really makes difficult to compare the results.
- Choosing informative, independent, and discriminating features are crucial in classification problems. The above-mentioned studies are lacking in this aspect.

2.6 Research Objectives

The research objectives for this thesis are:

- To propose a novel architecture by combining state-of-the-art works in terms of already available textual datasets and models.

- To conduct a comparative performance analysis of several state-of-the-art method by proposing more robust and accurate hate speech detection on available textual datasets.
- Identification of model for hate speech detection under multimodal dataset, achieving high robustness against the most competent and prevailing approaches.
- To perform a comparative analysis of several state-of-the-art methods by proposing more efficient hate speech detection algorithm on available image datasets.

2.7 Research Contributions

The main objective of the thesis is to design and develop novel machine and deep learning architectures capable of identifying hate content in multimedia data. Hence, the following architectures and frameworks are proposed to accomplish this:

- Proposed a novel architecture by combining state-of-the-art works in terms of already available textual datasets and models. A novel architecture utilizes the concept of truncated singular value decomposition (SVD) for detecting hate content on the textual dataset. Compared with the baseline results, our framework has performed better compared to various machine learning algorithms.
- Proposed a comparative analysis using fuzzy pattern classifiers, including both the top-down and bottom-up algorithms for identifying the hate contents on multiple datasets, compared to the baseline results obtained from diverse machine learning and deep learning classifiers. The result shows that fuzzy logic classifiers give decent results on available textual datasets.
- Proposed a novel multi-modal architecture for identifying hateful memetic information in response to the above observation. The proposed architecture contains a novel "multi-scale kernel attentive visual" (MSKAV) module that uses an efficient multi-branch structure to extract discriminative visual features. Additionally, MSKAV utilizes an adaptive receptive field using multi-scale kernels. MSKAV also incorporates a multi-directional visual attention module to highlight spatial regions of importance. The proposed model also contains a novel "knowledge distillation-based attentional caption" (KDAC) module. It uses a transformer-based self-attentive block to extract discriminative features from meme captions. The model claims high performance in identifying hate content from memes, beating SOTA multi-modal hate speech identification models.

- Proposed a novel architecture “MHS-STMA: Multimodal Hate Speech Detection via Scalable Transformer-Based Multilevel Attention” is proposed which consists of three main parts: a vision attention-based encoder for visual part, and a caption attention-based encoder for textual part, and a combined attention-based learning. To identify hate content, each component uses various attention processes and handles multimodal data in a unique way. Experimental results confirm the potency of the proposed architecture as it comfortably outperforms other state-of-the-art hate content detection approaches.

CHAPTER 3
TEXTUAL HATE CONTENT DETECTION

3.1 Scope of this Chapter

This chapter is dedicated to the problem of textual hate content detection. In the first approach, it is seen that reducing features using Truncated SVD along with hyper parameter tuning helped in increasing balanced accuracy and F1 score for algorithms like Logistic Regression, SVM and XGBoost when compared to the baseline results. Still, the proposed approach is lacking in handling uncertain or imprecise data. The second approach focuses on handling the uncertainty and vagueness in the data by implementing the fuzzy classifiers. An empirical evaluation of seven classifiers is presented for hate speech detection on two commonly used benchmarks of different data characteristics providing essential insights into their detection in terms of accuracy for their deployment in real-world applications. Fuzzy classifiers outperformed the other two classifiers out of the three.

3.2 A Truncated SVD Framework for Online Hate Speech Detection on the ETHOS Dataset

3.2.1 Abstract

Hate content on social media is currently one of the most significant risks, where the victim is either a single individual or a group of people. In the current scenario, online web platforms are one of the most prominent ways to contribute to an individual's opinions and thoughts. Free sharing of ideas on an event or situation also bulks on the web. Information sharing is sometimes a bane for society if primarily used platforms are utilized with some lousy intention to spread hatred for intentionally creating chaos/ confusion among the public. Users take this as an opportunity to spread hate to get some monetary benefits, the detection of which is of paramount importance. This article utilizes the concept of truncated singular value decomposition (SVD) for detecting hate content on the ETHOS (Binary-Label) dataset. Compared with the baseline results, our framework has performed better in various machine learning algorithms like SVM, Logistic Regression, XGBoost, and Random Forest.

3.2.2 Proposed Methodology

Hate speech is now a threat to society, affecting the dignity of an individual, unity, and nation. Many hate words are used alternatively. **Fig. 3. 1** shows the word cloud for hate speech explicitly generated from [37].

Fig. 3. 1 Word Cloud

Therefore, Eliminating and classifying hate content over social platforms is crucial and requires an hour. Data preprocessing is a component of data preparation. Several techniques are used to normalize the data before it is fed into any machine learning or AI development pipeline.

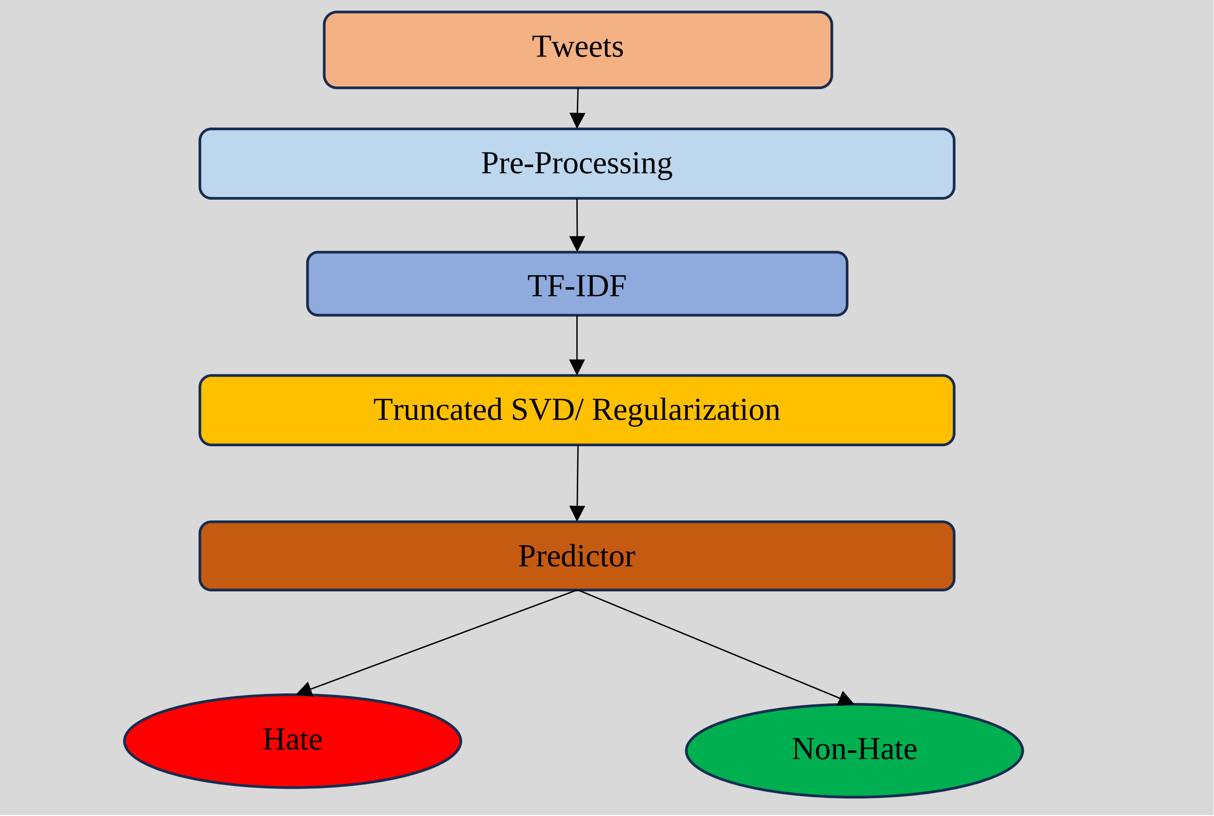


Fig. 3. 2 Proposed Flowchart

Fig. 3. 2 represents the flowchart adopted for implementation. During the data preprocessing in the first stage, tokenization is the initial step in any NLP pipeline, used to break unstructured data into chunks. Then, stemming is used as a normalized technique in which tokenized words are converted into short words to remove redundancies. Finally, the cleaned data is used to create a dictionary for key: value pairs. Second stage implements TF-IDF, to quantify the words. Truncated SVD is used for dimensionality reduction for simplifying the calculations. Hyper parameter tuning such as L1 regularization is also done for logistic regression, XGBoost and SVM. Finally, the Prediction is done using various machine learning algorithms like Logistic Regression, Random Forest, Support Vector Machines, and XGBoost.

3.2.3 Experimental Setup

Although the dataset size is very small. To prove that a dataset of higher quality is more useful than the larger datasets, we have considered a dataset D1 [37] which is approximately 24 times greater than ETHOS [103]. In this experiment, we train various machine learning models with default parameters on the ETHOS dataset and compare the results with D1 dataset. The results are compared in terms of F1 score and balanced accuracy.

F1 score (**Eqn. 3.1**) is defined as the combination of precision and recall of a classifier into a single metric by considering their harmonic mean.

$$F1 = \frac{(2 \times Recall \times Precision)}{Recall + Precision} \tag{3.1}$$

Table 3. 1 shows the results in the form of overall F1 scores, F1 Score (Hate) and F1 score (No Hate) of four machine learning models implemented on ETHOS and D1 datasets.

Table 3. 1 F1 Scores of ETHOS and D1 from SVM, LR, RF, XGBoost Models

Models	**ETHOS**			**D1**		
	F1 Score	**F1 Score (Hate)**	**F1 Score (No Hate)**	**F1 Score**	**F1 Score (Hate)**	**F1 Score (No Hate)**
SVM	67.71	59.60	73.63	75.47	12.86	79.30
LR	69.13	60.84	75.27	78.76	14.89	82.67
RF	67.01	58.85	73.03	67.21	12.73	70.55
XGBoost	65.30	54.50	73.44	75.39	10.62	79.35

The results are obtained when the models are trained on ETHOS and cross validation is done on D1 dataset. Balanced accuracy (**Eqn. 3.2**) is defined as the arithmetic mean of sensitivity and specificity. It is also considered as the further development in standard accuracy metric.

$$Balanced\ Accuracy = \frac{Specificity+Sensitivity}{2} \tag{3.2}$$

Balanced Accuracies are shown in the **Table 3. 2** representing that our proposed approach using truncated SVD and hyper parameter tuning gives better results than baseline results.

Table 3. 2 Comparison Table_Balanced Accuracy

Balanced Accuracy		
Models	**ETHOS_Our Approach**	**ETHOS_Baseline**
SVM	66.70	66.43
LR	67.07	66.94
RF	68.17	65.04
XGBoost	64.42	64.33

The graphical representation of balanced accuracies are shown in **Fig. 3. 3**.

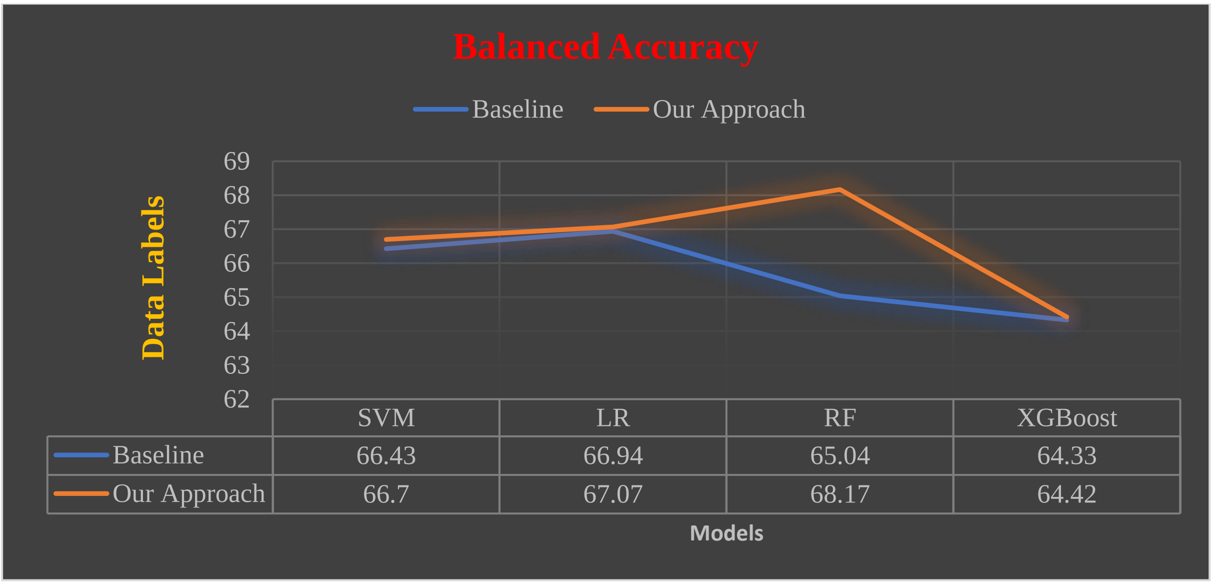


Fig. 3. 3 Performace Results: Balanced Accuracy

3.2.4 Conclusion

From the empirical evaluation done in the paper, it is seen that reducing features using Truncated SVD along with hyper parameter tuning helped in increasing balanced accuracy and F1 score for algorithms like Logistic Regression, SVM and XGBoost when compared to the baseline results. For Random Forest, only change in hyper parameter is giving good results.

3.3 Fuzzy and Machine Learning Classifiers for Hate Content Detection: A Comparative Analysis

3.3.1 Abstract

Hate content on social media is currently one of the most significant risks, where the victim is either a single individual or a group of people. In the current scenario, online web platforms are one of the most prominent ways to contribute to an individual's opinions and thoughts. Free sharing of ideas on an event or situation also bulks on the web. Information sharing is sometimes a bane for society if primarily used platforms are utilized with some lousy intention to spread hatred for intentionally creating chaos/ confusion among the public. Users take this as an opportunity to spread hate to get some monetary benefits, the detection of which is of paramount importance. This article includes various fuzzy pattern classifiers, including both the top-down and bottom-up algorithms for identifying the hate contents on multiple datasets, compared to the baseline results obtained from diverse machine learning or deep learning classifiers. Moreover, the result shows that fuzzy logic classifiers give decent results when classification is done on hate speech datasets.

3.3.2 Proposed Architecture

Hate speech is now a threat to society, affecting the dignity of an individual, unity, and nation. Many hate words are used alternatively. Therefore, Eliminating and classifying hate content over social platforms is crucial and requires an hour.

Data preprocessing is a component of data preparation. Several techniques are used to normalize the data before it is fed into any machine learning or AI development pipeline. Tokenization is the initial step in any NLP pipeline, used to break unstructured data into chunks of discrete values. Then, stemming is used as a normalized technique in which tokenized words are converted into short words to remove redundancies. Finally, the cleaned data is used to create a dictionary for key: value pairs. Word2Vec is used for mapping words to vectors of real numbers then a vector matrix is given as an input to classify as hate or non-hate. **Fig. 3. 4** explains the proposed methodology adopted. We have trained the data in the form of absolute values vectors. The training and testing ratio is 70:30, respectively. The training and testing labels are represented using one hot encoding. Then, we used three machine learning classifiers: Logistic Regression, Naïve Bayes, Support Vector Machines, and Random Forest, two deep learning classifiers: LSTM and Bi-LSTM, and two Fuzzy classifiers: Fuzzy Pattern classifiers and Fuzzy tree top down classifiers**.**

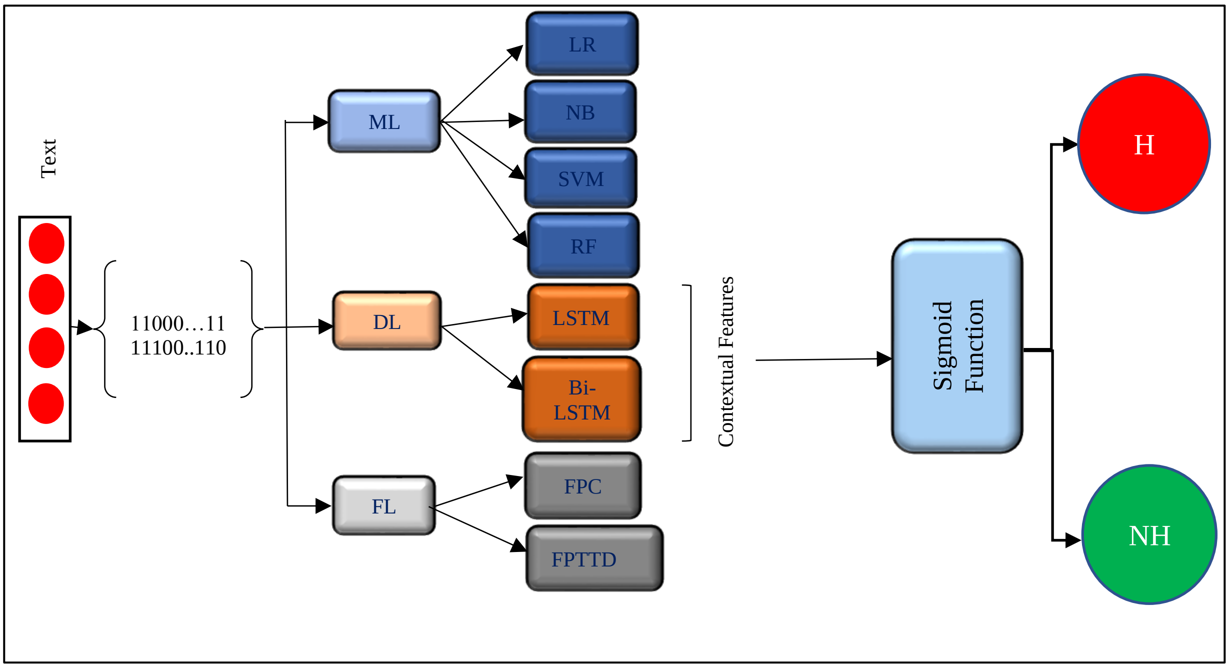


Fig. 3. 4 Proposed Architecture

Finally, the sigmoid function is used as a binary classification neural network. The fuzzy pattern tree used here follows bottom-to-top induction. Fuzzy pattern tree **(Fig. 3. 5**) takes the attributes or features in the form of a hierarchical structure in which the Number of features indicates leaf nodes, and inner nodes indicate the fuzzy arithmetic operators. The values are then combined to calculate the output.

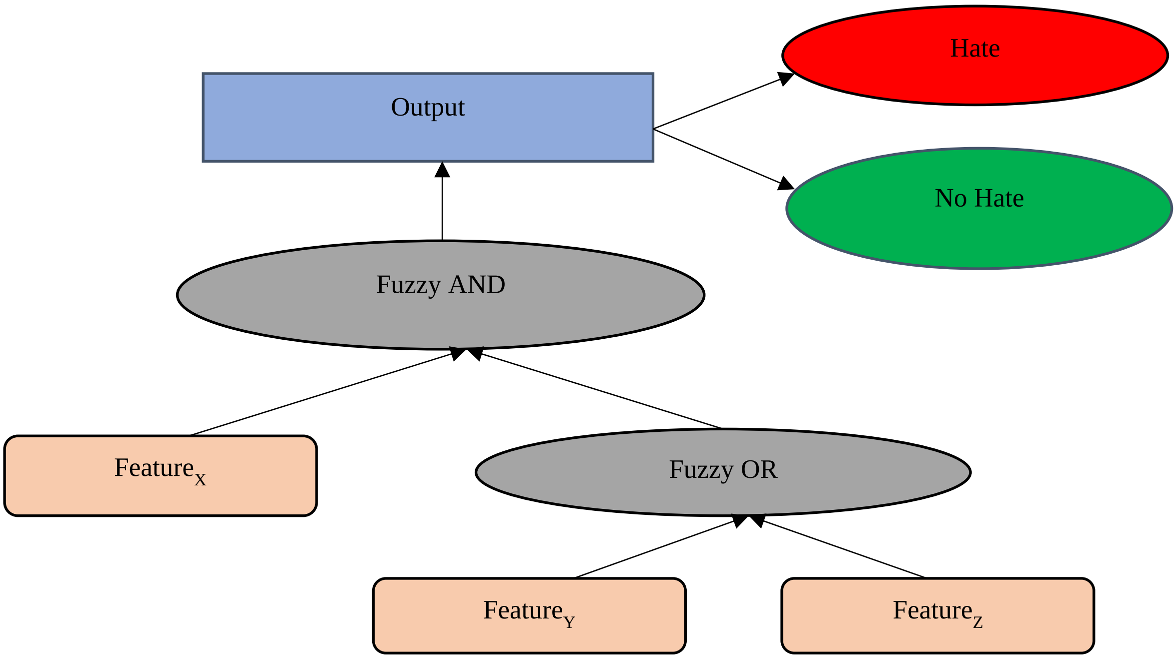


Fig. 3. 5 Fuzzy Pattern Tree

This iterative approach selects the best pattern tree with the least prediction error.

3.3.3 Experimental Setup

Two hate speech datasets of approximately similar size, named Davidson and El-Sherief, are adopted for experimentation. Datasets with their size and key statistics are shown in **Table 3. 3** and **Table 3. 4** respectively.

Table 3. 3 Dataset Size

Datasets	Source	Domain/Scope	Size
[37]	Twitter	HateBase Terms	24,802
[104]	Twitter	Hate Groups	22,584

Both datasets are multi-label and multiclass datasets. Both the datasets are from Twitter. Davidson dataset is annotated as Offensive, Hate, and Neither, making up 24,802 tweets. Similarly, El-Sherief dataset is categorized into grievances, inferiority, incitement, irony, stereotypes, threats, and a total of 22,584 tweets (**Table 3. 4)**. So, their labels are changed to 0 or 1 for binary classification. Three Machine Learning, Two Deep Learning, and Two Fuzzy Logic Classifiers are evaluated on these two datasets.

Table 3. 4 Dataset Description

Dataset	Class and Statistics
Davidson (Hate Speech and Offensive Language)	Offensive-19190
	Hate-1430
	Neither-4163
	Total~25K tweets
El-Sherief (Implicit Hate)	Grievance-24.2%
	Incitement: 20.0%
	Inferiority-13.6%
	Irony-12.6%
	Stereotypes- 17.9%
	Threats-10.5%
	Other-1.2%
	Total~22K tweets

The classification results in terms of accuracy are shown in **Table 3. 5.**

Table 3. 5 Comparison Table

Binary Classification			
Classifiers		**Accuracy**	
		[37]	**[104]**
Machine Learning	LR	87.59	70.15
	NB	68.71	42.80
	SVM	**88.78**	69.41
	RF	87.97	68.85
Deep Learning	LSTM	**89.41**	65.92
	Bi-LSTM	89.25	64.80
Fuzzy Logic	FPC	94.19	90.21
	FPTTD	**94.26**	89.19

It has been observed from **Table 3. 5** that SVM achieves better classification results from four machine learning classifiers; LSTM has outperformed Bi-LSTM in deep learning models whereas, in fuzzy classifiers, the Fuzzy tree top-down classifier has given acceptable performance.

3.3.4 Conclusion

An empirical evaluation of seven classifiers for hate speech detection is presented on two commonly used benchmarks "Davidson" and "El-Sherief" of different data characteristics providing essential insights into their detection in terms of accuracy for their deployment in real-world applications. Fuzzy classifiers outperformed the other two classifiers out of the three.

3.4 Significant Outcomes of the Chapter

- In the first approach, it is seen that reducing features using Truncated SVD along with hyper parameter tuning helped in increasing balanced accuracy and F1 score for algorithms like Logistic Regression, SVM and XGBoost when compared to the baseline results. Still, the proposed approach is lacking in handling uncertain or imprecise data.
- The second approach focuses on handling the uncertainty and vagueness in the data by implementing the fuzzy classifiers. An empirical evaluation of seven classifiers is presented for hate speech detection on two commonly used benchmarks of different data characteristics providing essential insights into their detection in terms of accuracy for their deployment in real-world applications. Fuzzy classifiers outperformed the other two classifiers out of the three.

The following research works form the basis of this chapter:

- ❖ **Anusha Chhabra**, Dinesh Kumar Vishwakarma. "A Truncated SVD Framework for Online Hate Speech Detection on the ETHOS Dataset." **IEEE Conference: 4th International Conference on Innovative Trends in Information Technology (ICIIIT)**, *IIIT Kottayam, Kerala. (2023).*
- ❖ **Anusha Chhabra**, Dinesh Kumar Vishwakarma. "Fuzzy and Machine Learning Classifiers for Hate Content Detection: A Comparative Analysis." **IEEE Conference: 4th International Conference on Artificial Intelligence and Speech Technology (AIST)**, *IGDTUW, Delhi. (2022)*

CHAPTER 4
MULTIMODAL HATE CONTENT DETECTION

4.1 Scope of this Chapter

This chapter is dedicated to the problem of detecting multimodal hate content detection by presenting the first framework as dual-branch network composed of knowledge distillation attention for extracting the essential information from the caption modality and multi-kernel attention for collecting pertinent information from the images. Extensive testing on three publicly accessible datasets showed that the suggested architecture outperformed baseline models, claiming better results in terms of accuracy and AUC scores. Numerous ablation trials are conducted on image datasets to conclude that the proposed architecture is contributing to the performance of hate content identification in memes. In the second approach, the proposed model “MHS-STMA” explored the problem of learning complementary information between multimodal data. The architecture utilizes transformers for capturing the dependencies and relationships between the elements in a sequence. The proposed architecture also utilizes attention mechanisms at multiple levels and focuses on crucial regions in the images based on the attended textual features. Self-attention mechanism is applied at the end to remove any redundancy from the multimodal data. The experimental results conducted on three popular datasets show that our method performs efficiently.

4.2 Multimodal Hate Speech Detection via Multi-Scale Visual Kernels and Knowledge Distillation Architecture

4.2.1 Abstract

People increasingly use social media platforms to express themselves by posting visuals and texts. As a result, hate content is on the rise, necessitating practical visual caption analysis. Thus, the relationship between image and caption modalities is crucial in visual caption analysis. Contrarily, most methods combine features from the image and caption modalities using deep learning architectures with millions of parameters already trained without integrating a specialized attention module, resulting in less desirable outcomes. This paper suggests a novel multi-modal architecture for identifying hateful memetic information in response to the above observation. The proposed architecture contains a novel "multi-scale kernel attentive visual" (MSKAV) module that uses an efficient multi-branch structure to extract discriminative visual features. Additionally, MSKAV utilizes an adaptive receptive

field using multi-scale kernels. MSKAV also incorporates a multi-directional visual attention module to highlight spatial regions of importance. The proposed model also contains a novel "knowledge distillation-based attentional caption" (KDAC) module. It uses a transformer-based self-attentive block to extract discriminative features from meme captions. Thorough experimentation on multi-modal hate speech benchmarks MultiOff, Hateful Memes, and MMHS150K datasets achieved accuracy scores of 0.6250, 0.8750, and 0.8078, respectively. It also reaches impressive AUC scores of 0.6557, 0.8363, and 0.7665 on the three datasets, respectively, beating SOTA multi-modal hate speech identification models.

4.2.2 Proposed Methodology

Most of the work is done on textual datasets [72] using majorly baseline machine learning methods to evaluate the results. However, it is also observed that there is an inclination towards deep learning models for text classification after the SemEval-2019 Task5 competition [105]. In addition, transformers-based attention mechanisms have also shown advancements in

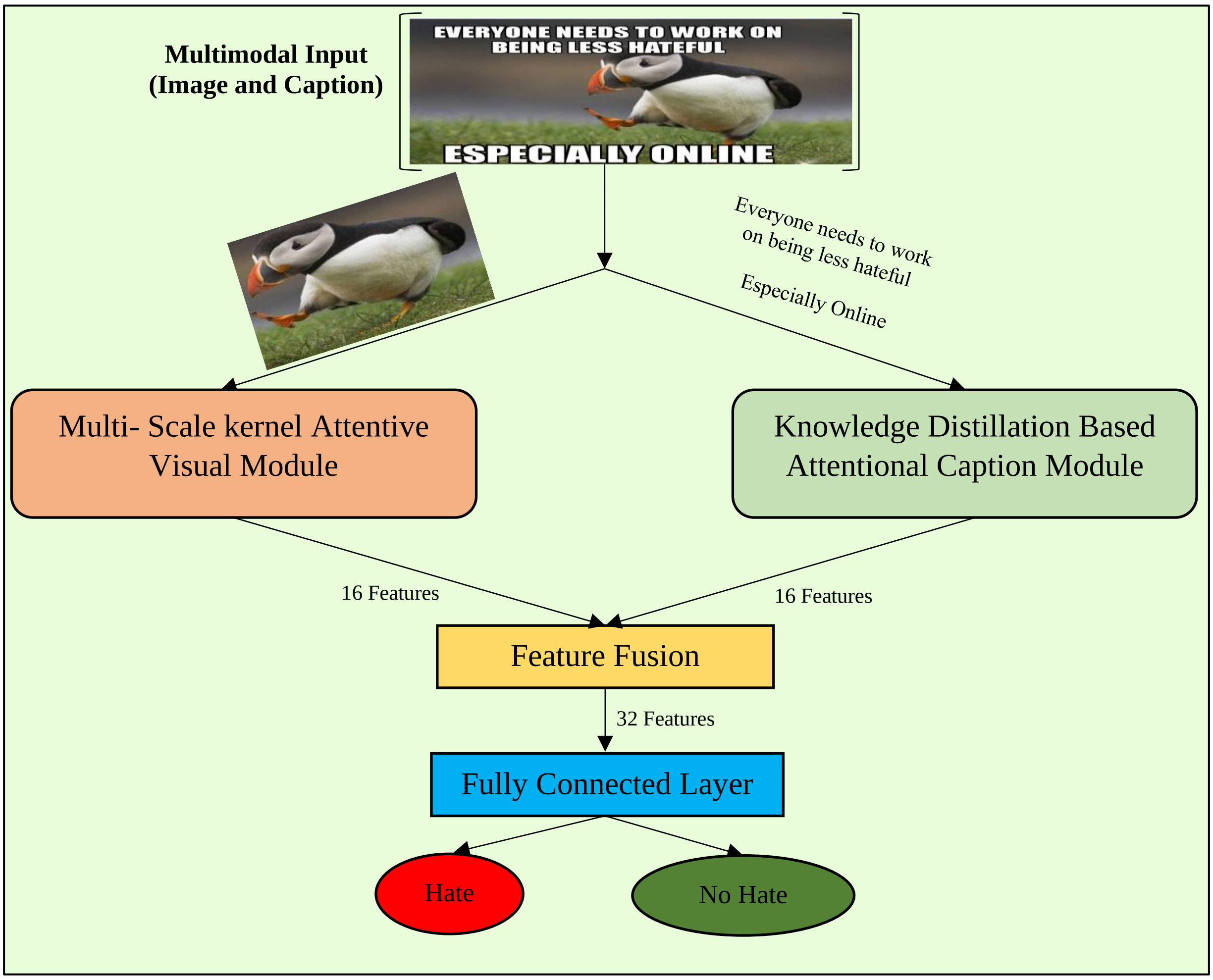

Multimodal Input
(Image and Caption)
EVERYONE NEEDS TO WORK ON BEING LESS HATEFUL
ESPECIALLY ONLINE
Everyone needs to work on being less hateful
Especially Online
Multi- Scale kernel Attentive Visual Module
Knowledge Distillation Based Attentional Caption Module
16 Features
16 Features
Feature Fusion
32 Features
Fully Connected Layer
Hate
No Hate

Fig. 4. 1 Proposed Architecture: Multi- Scale Kernel Attentive Visual Module cum Knowledge Distillation Attentional Caption Architecture learning the contexts in a more reliable manner [80]. The proposed architecture of the paper is composed of two modules "Multi-Scale Kernel Attentive Visual Branch" and "Knowledge Distillation based Attentional Caption Branch" each for images and captions respectively as shown in **Fig. 4. 1**. After feeding the respective inputs to their respective modules, the (16,16) features obtained from both branches are fused together (32 features) before passing through the fully connected layer resulting in the binary classification as hate or non-hate.

The following subsections describe the respective modules:

4.2.2.1 Multi-Scale Kernel Attentive Visual (MSKAV) Module

This module describes the vision branch that consists of a dynamic selection of multi-scale convolution kernels along with multi-branch architecture and visual Attention. The architecture is a simple, highly modularized network framework for image classification. It is made by repeating a building block with the same topology and stronger representations.

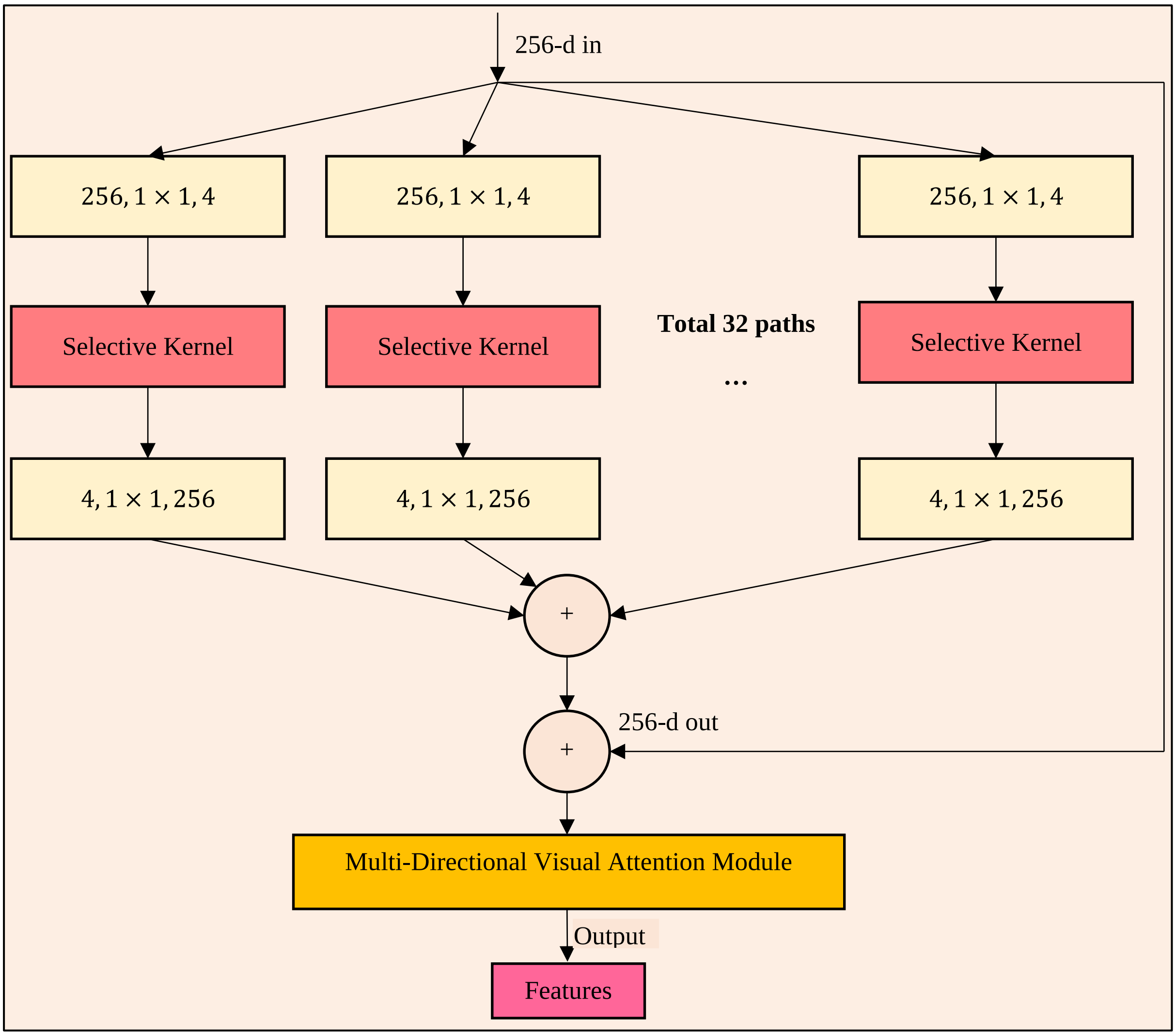


Fig. 4. 2 Visual Branch Architecture via aggregated residual transformation and attentional module

This concept exposes a new dimension, "cardinality" as an essential factor along with depth and width.

The visual branch architecture with cardinality 32 (total 32 paths) and 4d bottleneck width, which is further enhanced using selective kernel convolution followed by multi-directional visual attention as shown in **Fig. 4. 2**. For each path, Conv 1×1 –Selective Kernel- Conv 1×1 are done at each convolution path. Each path's internal dimension is designated as d (d=4). The cardinality C (C=32) represents the number of paths. It is also the dimension of 128, which is further raised to 256 if we add the dimension of each "selective kernel" block. Then all the paths are added together by the concatenation operator. The fundamental principle of choosing a "Selective Kernel" convolution is that it is helpful in optimization by enabling each neuron to adaptively change the size of its receptive fields (RF) based on various scales of input data. A detailed explanation of "Selective Kernel" and a multi-directional visual attention module is given in further sub-sections. The visual branch architecture which follows multi-branch architecture is used to learn characteristics from raw data via targeting a specific goal making it effective on detecting hateful memes. The idea of using the multi branch architecture is the presence of hundreds of layers making the training much easier.

4.2.2.1.1 Computationally Efficient Multi-Branch Feature Learning

In recent years, CNNs have become much deeper and deeper as multiple layers are better at generalizing. Inception architecture is a deep network architecture with the property of divide-transform-merge strategy proved to give good accuracy against previous deep networks with the limitation of high computational power. In comparison with the deeper or wider networks, it has been found that increasing the size of the set of transformations with the same architecture is more effective in terms of classification and complexity. The aggregated transformation is given as follows:

$$\sum_{n=1}^{D} w_n x_n \rightarrow \sum_{n=1}^{S} \mathrm{T}_n\,(x) \tag{4.1}$$

where, x is the D- channel input vector $(\alpha \times \beta \times \Upsilon)$,

$'S'$ is the size of the set of transformations T

LHS of the (**Eqn.** 4.1**)** shows the splitting of vector x as low-dimensional representation, then transformation is done with the scaling vector, w which is finally aggregated as $\sum_{n=1}^{D} w_n x_n$ whereas RHS of the equation is the aggregated transformation with an arbitrary function as $\mathrm{T}_n\,(x)$.

To further enhance the classification capability of the backbone architecture from [106], a “selective kernel” [107] has been used to exploit the dynamic selection of multi-size convolution kernels.

4.2.2.1.2 Selective Kernel Convolution for Adaptive Receptive Field

A computationally lightweight selective kernel network (**Fig. 4. 3**) was introduced with an adaptively adjustable property of receptive field size, making it more effective and efficient for producing the features in image classification. Additionally, it includes the triplet operations: divide, merge, and select.

- **Divide:** The divide procedure produces numerous pathways with varied kernel sizes that match different RF sizes of neurons. In this phase, two transformations Ƒ': X→Ũ ϵ $\mathfrak{R}^{D'\text{ and}}$ Ƒ'': X→Û ϵ $\mathfrak{R}^{D'}$ with variation in the kernel sizes as 3 × 3 and 5 × 5 respectively, are conducted on the feature map Ƒ ϵ $\mathfrak{R}^{D}$ with dimensions D as $\alpha \times \beta \times \Upsilon$. Both transformations are composed of depth-wise convolutions, Batch Normalization (BN), and ReLU function in series. The combination of all transformations splits the input feature map X into M (M=2).
- **Merge:** The Merge operation embeds the global information from the two branches (M) via an element-wise summation by the Global Average Pool (GAP), generating the vector 's' as channel-wise statistics of 1 × 1 dimension, then adds a fully linked layer to mix and aggregate the information from two paths Ƒ=Ƒ’+Ƒ’’ to create an extensive and global representation of selection weights.

  Further, in this operation, a compact feature descriptor $'z'$ (**Eqn. 4.2**) is designed to facilitate precise selection guidance through the use of a single and fully connected layer.

  $$z = ReLU\left(BN(Ws)\right) \tag{4.2}$$

  where, $W \in R^{d\times 1}$; $d\ is\ the\ combination\ of\ \alpha\ and\ \beta,$

  BN is Batch Normalization,

  's' is the channel wise statistics of 1 × 1 dimension.

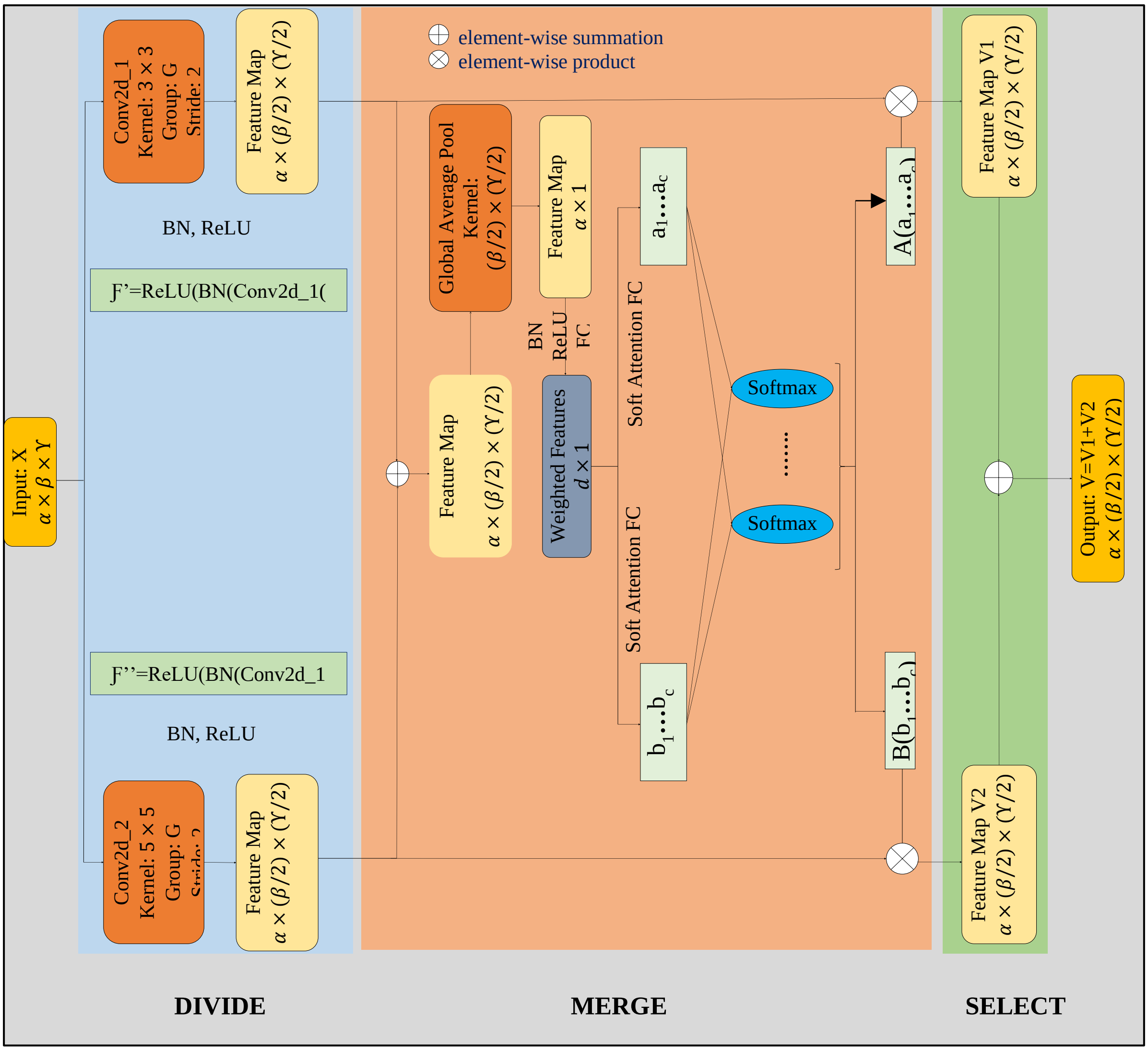

element-wise summation
element-wise product
Conv2d_1 Kernel: 3 × 3 Group: G Stride: 2
Feature Map α × (β/2) × (ϒ/2)
BN, ReLU
Ƒ'=ReLU(BN(Conv2d_1(
Input: X α × β × ϒ
Ƒ''=ReLU(BN(Conv2d_1
BN, ReLU
Conv2d_2 Kernel: 5 × 5 Group: G Stride: 2
Feature Map α × (β/2) × (ϒ/2)
Global Average Pool Kernel: (β/2) × (ϒ/2)
Feature Map α × 1
Feature Map α × (β/2) × (ϒ/2)
BN ReLU FC
Weighted Features d × 1
Soft Attention FC
Soft Attention FC
a1…ac
b1…bc
Softmax
……
Softmax
A(a1…ac)
B(b1…bc)
Feature Map V1 α × (β/2) × (ϒ/2)
Feature Map V2 α × (β/2) × (ϒ/2)
Output: V=V1+V2 α × (β/2) × (ϒ/2)
DIVIDE
MERGE
SELECT

Fig. 4. 3 Selective Kernel Convolution for Adaptive Receptive Field (Divide, Merge, and Select)

Select: The select procedure aggregates the feature maps from various kernel sizes, i.e., 3 × 3 and 5× 5, in accordance with the selection weights A and B to generate the feature maps V1 and V2, respectively. Then select $V = A \otimes Ƒ' + B \ \otimes$ Ƒ''.

The output feature map V is fed into the multi-directional visual attention module for capturing the cross-dimension between spatial and channel dimensions of the input. The description of the visual attention module is given in the further sub-section.

4.2.2.1.3 Multi-Directional Visual Attention

The attention mechanism in deep neural networks includes either channel or spatial or both attentions. Contrary to spatial attention, which weighs each pixel in a single feature map, channel attention essentially weighs each feature map or channel in the tensor. The proposed attention mechanism is a three-branch multi-directional visual attention (**Fig. 4. 4**) involves both attention (channel + spatial). When computing attention on a single pixel channel, there is no dependency between channel dimension and spatial dimension, which could result in a significant loss of spatial information, this module is used to collect cross-dimension dependencies via multi-branch to solve the information loss problem. The cross-dimensional dependencies are introduced by independently capturing the dependencies between the input tensor's (C, H), (C, W), and (H, W) dimensions. The detailed description is as follows:

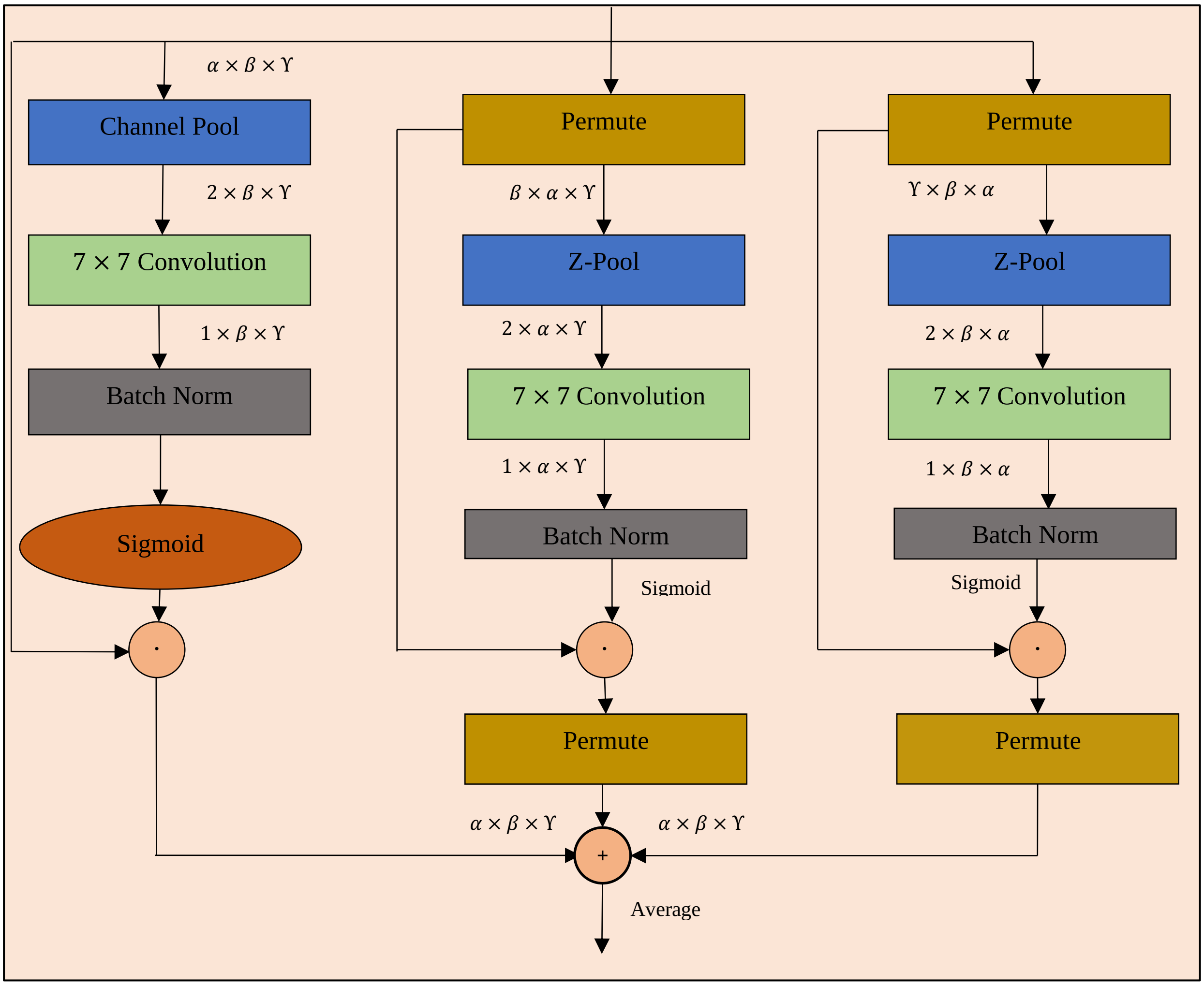


Fig. 4. 4 Multi-Directional Visual Attention Module [108]

Each of the three branches of the three-branch multi-directional visual architecture (**Fig. 4. 4**) is in charge of calculating and applying the attention weights across two of the three dimensions of the input tensor. The bottom branch computes fundamental spatial attention, whereas the top two branches calculate channel attention weights against each of the two spatial dimensions. The input tensor is rotated in the top two branches to change its dimensions, and then the Zeroth Pool (Z-Pool) operator is used to reduce the zeroth dimension to two by concatenating its average-pooled and maximum-pooled features across that dimension. After going through a single spatial convolution layer and sigmoid activation, the output of this tensor is then relayed, further reducing the zeroth dimension to one. After element-wise multiplying the final result with the permuted input, the output is then rotated to its original dimensions (the same as that of the input). After completing this for each of the three branches, the output is formed by averaging the three resulting tensors.

The further sub-section describes the textual branch via knowledge distillation-based attentional tokenization architecture.

4.2.2.2 Knowledge Distillation Based Attentional Caption (KDAC) Module

This module describes the textual branch that consists of a knowledge distillation base during the pre-training phase which is an approximated version of BERT, i.e., DistilBERT [82]. It is a smaller network used to approximate the whole output distributions of a big neural network once it has been trained. The advantage of using DistilBERT is that it tries to maintain as much performance as feasible while optimizing the training by scaling back BERT and speeding it up. Another advantage of using DistilBERT, in particular, is that it is 40% smaller, 60% faster, and 97% functionally equivalent to the original BERT-base model. As inputs, NLP models usually require numeric vectors, which frequently require transforming attributes like words into numbers. The BERT algorithm, which Google released in 2019, is a member of the class of NLP-based language models referred to as transformers and has tokenizers that convert sentences into numeric representations. It has been pre-trained on big text datasets, making it superior to the earlier word embedding methods like TF-IDF, Word2Vec, Glove, etc. As a result, the word embedding produced by DistilBERT is superior. The attention mechanism

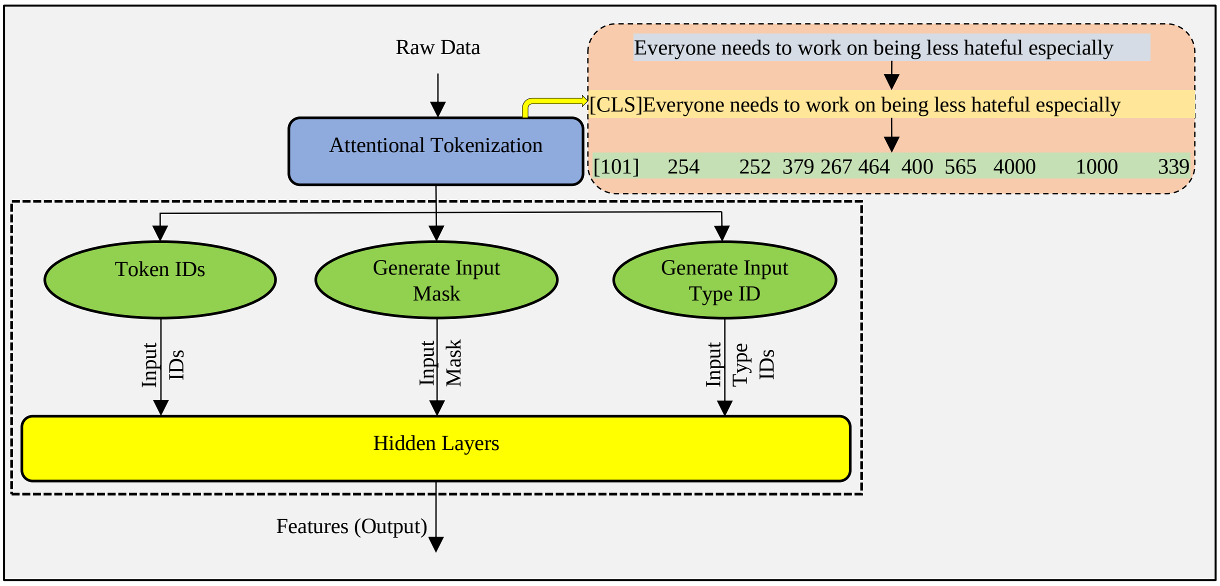


Fig. 4. 5 Knowledge Distillation Based Attentional Caption (KDAC) Architecture for Caption Branch: Attentional Tokenization Architecture

is used by the model to produce word embedding. It so collects connections for each word based on the terms on both sides of the sentence. Positionally encoded-word embedding keeps track of the sequence and arrangement of each word in the sentence. It produces high-quality contextualized or context-aware word embedding by passing through each encoder layer. In this experiment, the DistilBERT-cased model is employed (see **Fig. 4. 5)** with 12 attention heads, 768 hidden layers, and six transformer layers for pre-processing. After input had been tokenized, the texts were padded, the tokens were converted into input ids, and then the sequence of vectors was fed into the DistilBERT model. For the given caption, $\boldsymbol{C_j}$ with **n** words can be denoted as $\boldsymbol{C_j}$**= {**$\mathbf{w_{j1}.w_{j2}.w_{j3},\ldots\ldots\ldots,w_{jn}}$**}**. Each word $\mathbf{w_{jk}}$ is embedded in a vector representation. $\boldsymbol{v_j} \in \boldsymbol{R^d}$ is the d- dimension vector for $\mathbf{K^{th}}$ word. The final tokenization is denoted as $\boldsymbol{V}_{\mathbf{j}}$ **= {**$\mathbf{v_{j1}, v_{j2}, v_{j3}, \ldots\ldots\ldots, v_{jn}}$**}.** The output is generated as a set of 16 features from the text module.

4.2.2.3 Feature Fusion

Social media posts that convey a user's opinion are examples of multi-modal data. We have carried out a cross-model feature-level fusion for the detection of hate content. In multi-modal hate speech detection, by effectively creating the link across modalities, the right meanings of the memes must be guaranteed. Early fusion has produced impressive outcomes in a number of investigations, while late fusion has produced studies with good performance. The algorithm for the proposed work is given in **Table 4. 1**.

Table 4. 1 Algorithm for the proposed work

```
Algorithm 1: Multimodal Hate Speech Detection via Multi-Scale Visual Kernels and Knowledge Distillation
Attentional Caption Module
Aim: To learn a mapping function from the multi-modal training
Input: Captions and Visual Sets
Output: Hate Content Classification task as hate, or no-hate
    1.  All words in the caption content have been tokenized
    2.  Extract characteristics from the caption information
    3.  Extract characteristics from the visual information by Conv $1 \times 1$ - Selective Kernel – Conv $1 \times 1$ at
        each convolution path and multi-directional visual attention module
    4.  For N $\leftarrow$ 1 to Epochs
          Word to vector representations
          Caption feature extraction from caption content
          Feature mapping from visual modality
          Concatenation of features by summation or concatenation operator
          Output ϵ {Hate, No Hate} via fully connected layer and normalized exponential function.
          Compute the Loss and carry out back propagation;
    5.  End
```

In the proposed study, the main component of the suggested architecture is feature fusion, the combining of data from many layers. Basic techniques like summation or concatenation are frequently used to achieve this. Following feature fusion, the vectors are transferred to a softmax (normalized exponential function) layer for hate identification.

4.2.3 Experimental Setup

We conduct experiments on three hate image datasets for classification. We evaluate a pre-trained DistilBERT model and the SKResNext model in a combined way. We have included the results for fusion methods where 16 features from the image module and 16 features from the text module are concatenated to a total of 32 features. Then, 32 features are passed to a fully connected layer resulting in the binary classification as hate or non-hate. This section provides brief dataset descriptions on which the experimentation is being done, various classification metrics used for the evaluation of results, hardware requirements, pre-processing, and hyperparameter specifications.

4.2.3.1 Datasets

We have done our experimentation on three publicly available datasets specifically related to hate or offensive content. The first issue we found in these datasets is that in most images, the textual part and the visual information both imply different things. The second issue is related to the variation in the channel size of images.

4.2.3.1.1 Multimodal Memes Dataset (MultiOFF)

A collection of memes from multiple social platforms makes up the multimodal memes dataset (MultiOFF). The 2016 U.S. Presidential Election Event dataset was produced utilizing 743 memes that were built from a collection of manually annotated picture URLs and text that was present in the images. All the unrelated features, such as likes, upvotes, timestamps, etc., have already been removed at the time of dataset preparation. Furthermore, the dataset is then separated into three files: Training file, Testing File, and Validation File, respectively containing 445, 149, and 149 memes.

4.2.3.1.2 Hateful Memes Dataset

Dataset released by Facebook AI to identify the multimodal hate content over internet memes. A collection of about 10000 PNG images, further separated into training and testing files. The dataset is reviewed by three annotators classified as binary labels.

4.2.3.1.3 MMHS150K Dataset

MMHS150K dataset contains approximately 1,50,000 tweets collected from Twitter and annotated using the Amazon Turk crowdsourcing platform divided into six labels: No attacks, sexist, racist, religion-based, homophobic, or attacks to other communities.

4.2.3.2 Classification Metrics

Six distinct performance metrics, including accuracy, precision, recall, F1-Score, area under the curve (AUC), and Mathews Correlation Coefficient (MCC), were used to assess the two-class classification model. **Table 4. 2** shows the general formulas for the above-mentioned performance metrics, along with their ranges. Before going deep into the description of the six-performance metrics, a few significant terms associated with the performance are given:

True Positive (TP): It is described as a situation where the values for the predicted and actual outcomes are both positive.

True Negative (TN): It is described as a situation where the values for the predicted and actual outcomes are both negative.

False Positive (FP): A situation in which the predicted value is positive but the actual value is negative.

False Negative (FN): A situation in which the predicted value is negative but the actual value is positive.

Table 4. 2 General Formulas for Classification Metrics used to Evaluate the Performance

Metrics	Formula	Range
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	[0,1]
Precision	$\frac{TP}{TP + FP}$	[0,1]
Recall	$\frac{TP}{TP + FN}$	[0,1]
F1-Score	$\frac{(2 \times Recall \times Precision)}{Recall + Precision}$	[0,1]
AUC	-	[0,1]
MCC	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$	$[-1, +1]$

Based on the above definitions of the terms, following is the explanation for the above mentioned classification metrics:

- **Accuracy:** The number of accurate (TP + TN) predictions divided by the total number of predictions (TP + TN + FP + FN) is used to measure the model's accuracy.

- **Precision:** It is a metric used to assess a model's ability to accurately identify the positive class; it is calculated as the ratio of true positives to both true and false positives.
- **Recall:** Recall estimates how effectively the model separates the true positives and false negatives from all of the positive observations in the dataset.
- The definition of **F-Measure or F1-score** is the harmonic mean of recall and precision.
- **AUC:** As a measure of the model's performance over all potential classification thresholds, AUC is the measurement of the complete two-dimensional area under the curve.
- **Mathews Correaltion Coefficient (MCC):** The correlation between the observed and anticipated classes is measured by the Mathews Correaltion Coefficient (MCC). MCC with a value of '1' denotes positive correlation, whereas MCC with a value of '-1' denotes negative correlation.

4.2.3.3 Hardware

We have conducted the proposed experiment on two NVIDIA Titan RTX 24GB GPUs in parallel and the system memory used is 128GB.

4.2.3.4 Pre-processing

For the Pre-Processing of images, the images of size $(3 \times 256 \times 256)$ pixel values normalized to $[0,1]$. Four-channel images with dimensions $[x, y, z, w]$ are taken into consideration only after changing the dimension to randomly selected three channels such as $[x, y, z]$, $[x, y, w]$, $[y, z, w]$, $[x, z, w]$. For the pre-processing of text, an inbuilt DistilBERT preprocessing layer is utilized that tokenizes and packs inputs. **Table 4.** 3 shows the overall size of the datasets and the split ratio as approximately $80: 10: 10$ for training, testing and validation respectively.

Table 4. 3 Dataset Size and Split Ratio (Training set: Validation set: Testing set)

Dataset	**Size**	**Training set**	**Validation set**	**Testing set**
MultiOff	737	500	117	120
Hateful Memes	8496	6800	800	896
MMHS150K	140000	100000	20000	20000

4.2.3.5 Hyper parameter Specification

The hyper parameter settings for the experiments are done on the datasets are shown in **Table 4.** 4. The hyper parameters used are number of epoch, batch-sizes, optimizers, learning rate and decay rate.

Table 4. 4 Hyper parameter Specifications in terms of Number of Epochs, Batch-Size, Optimizer, Learning Rate, Linear Decay

Datasets	**Number of Epochs**	**Batch-Size**	**Optimizer**	**Learning Rate**	**Linear Decay**
MultiOff	40	4	Adam	0.0001	10%
Hateful Memes	20	16	Adam	0.001	20%
MMHS150K	10	32	SGD	0.001	50%

4.2.4 Results

This section shows the classification performance of our proposed architecture in terms of accuracy, precision, recall, F1-score, area under curve (AUC), and Mathews Correlation Coefficient (MCC) for each of the datasets. The performance metrics signifies that the proposed architecture shows the tremendous improvements in terms of accuracy and AUC with a value of 0.8078 and 0.7665 respectively in MMHS150K and an accuracy, precision, AUC

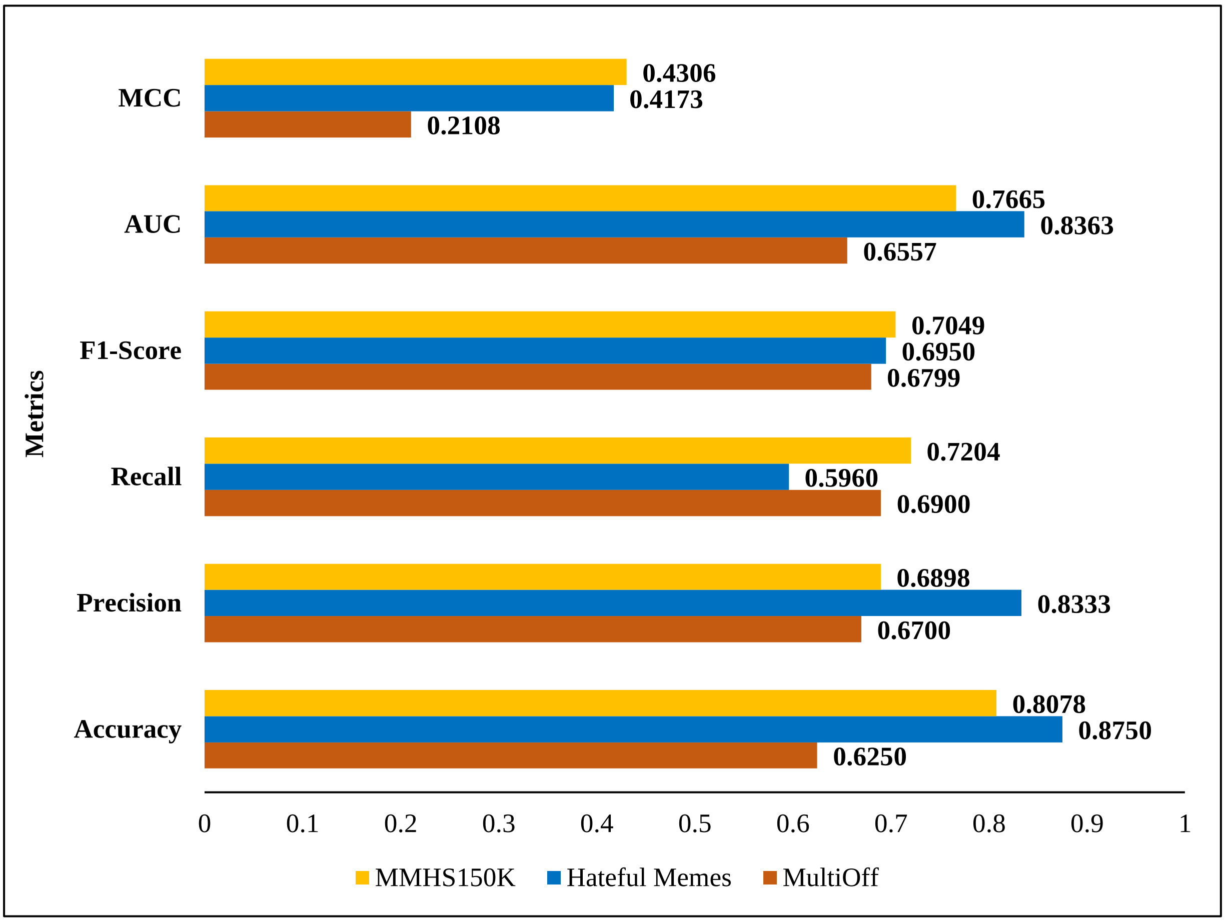


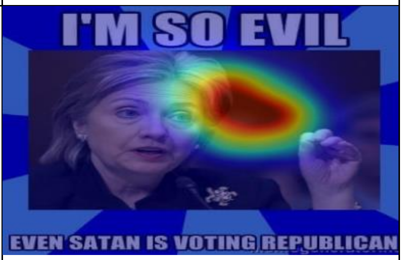
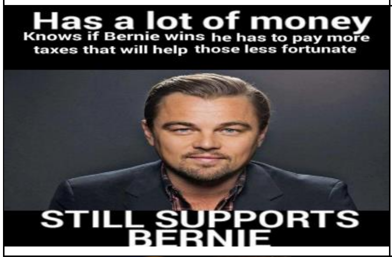
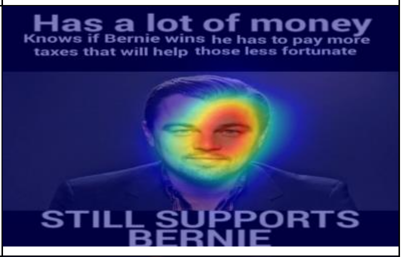
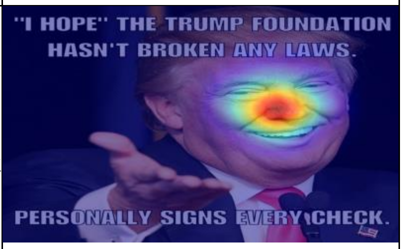
Fig. 4. 6 Classification Results in terms of Accuracy, Precision, Recall, F1-Score, AUC, and MCC

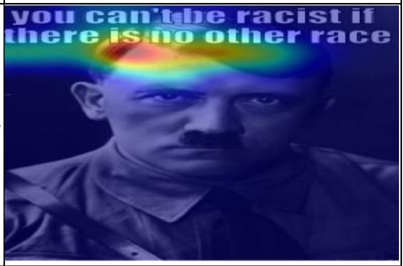
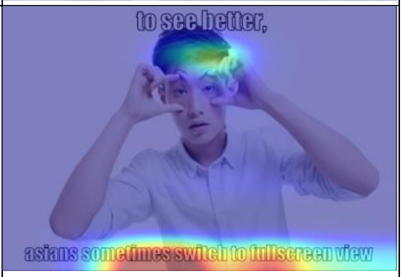
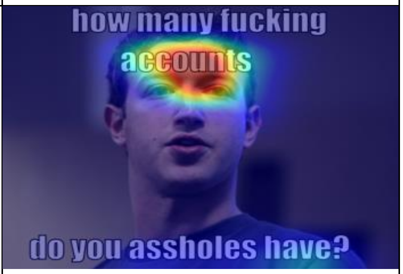
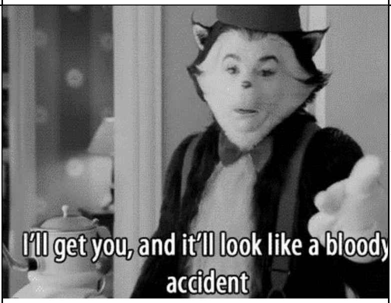
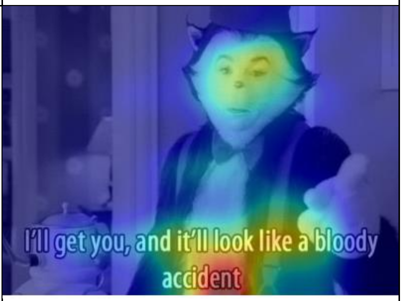
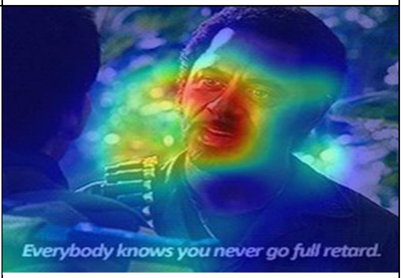
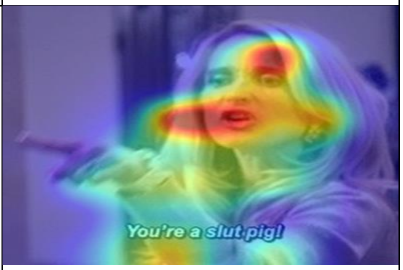
with 0.7607, 0.8333, 0.8261 respectively in hateful memes. For MultiOff dataset, accuracy and AUC have shown a significant improvement with scores 0.6250 and 0.6557 respectively. Precision, Recall and F1-scores with values 0.6700, 0.6900, and 0.6799 have surpassed the benchmark results for MultiOff dataset. We can visualize the classification results of all three datasets in **Fig. 4.** 6.

4.2.4.1 MSKAV Spatio-Region Focus

Due to the widespread usage of social media and digital platforms, consumers are now able to express their opinions through a variety of mediums. Memes are a very common way of sharing views now a day. As part of images and captions in memes, both contain a significant amount of information that can't be ignored. The spatial region represents the informative section of the image. Thus, it locates the relevant visual parts according to the visual attended features **Fig. 4.** 5 represents the visualization of activation mapping by attention method, i.e., LayerCAM [109], to locate the fine-grained localization of objects.

Table 4. 5 MSKAV Spatio-Region of Importance showing the Localization via LayerCAM

Datasets	**Memes**	**Caption**	**Region of focus for MSKAV branch**
MuliOff		I'm so evil Even Satan is voting republic	
	 Has a lot of money Knows if Bernie wins he has to pay more taxes that will help those less fortunate STILL SUPPORTS BERNIE	Has a lot of money Knows if Bernie wins he has to pay more taxes that will help those less fortunate Still supports Bernie	 Has a lot of money Knows if Bernie wins he has to pay more taxes that will help those less fortunate STILL SUPPORTS BERNIE
	"I HOPE" THE TRUMP FOUNDATION HASN'T BROKEN ANY LAWS. PERSONALLY SIGNS EVERY CHECK	I hope the Trump foundation hasn't broken any laws Personally signs every check	 "I HOPE" THE TRUMP FOUNDATION HASN'T BROKEN ANY LAWS. PERSONALLY SIGNS EVERY CHECK.

Datasets	Memes	Caption	Region of focus for MSKAV branch
Hateful Memes		You can't be racist if there is no other race	
	to see better, asians sometimes switch to fullscreen view	To see better, Asians sometimes switch to fullscreen view	 to see better, asians sometimes switch to fullscreen view
		How many fucking accounts Do you assholes have?	
MMHS150K		I'll get you, and it'll look like a bloody accident	
		Everybody knows you never go full retard	
		You're a slut pig	

The observations from the MSKAV Spatio region focus in images are as follows:

- Provides better object localization by focusing on the specific spatial region.

- The activation map by LayerCAM determines the impact of each region on the output of a model.

4.2.4.2 Baseline Comparison

In this section, we compare our results on MultiOff [87], Hateful memes [96], and MMHS150K [101] with their respective SOTA methods, as shown in **Table 4.** 6. Notably, there are very few implemented works on multi-modal hate content detection. The approaches compared in this section includes these few contributions [[87], [99], [96], [99], [100], [101]].

Table 4. 6 Comparison Table in terms of performance metrics on three benchmark datasets

Datasets	**Methods**	**Performance Metrics**				
		Accuracy	**Precision**	**Recall**	**F1-Score**	**AUC**
MultiOff	[87]	-	0.4000	0.6700	0.5000	-
	[99]	-	0.6450	0.6510	0.6460	-
	Proposed	-	**0.6700**	**0.6900**	**0.6799**	-
Hateful Memes	[96]	0.6947	-	-	-	0.7544
	[99]	0.7580	-	-	-	0.8280
	[100]	0.7108	0.7000	-	0.6900	0.7141
	Proposed	**0.8750**	**0.8333**	**-**	**0.6950**	**0.8383**
MMHS150K	[101]	0.6850	-	-	0.7040	0.7340
	Proposed	**0.8078**	-	-	**0.7049**	**0.7665**

The comparison of our proposed model with the given methods shows that learning and extracting features via multi-scale and multi-directional visual and caption branches improves the accuracy and AUC approximately by 15% and 0.64%, respectively, in the Hateful Memes dataset by 18% and 4% respectively in MMHS150K dataset.

4.2.4.3 Ablation Study

This section conducts ablation experiments to examine the usefulness of our proposed architecture. Numerous ablation studies are performed on MultiOff, Hateful memes, and MMHS150K datasets. Based on the proposed architecture, we have generated three cases for the conduction of ablation study. The first case considers only the KDAC branch denoted as "KDAC (Text Branch Only)." The second case eliminates the visual attention module from MSKAV branch, denoted as "MSKAV w/o Attention." The last case, denoted as "MSKAV with Attention." The outcomes of these ablation trials are denoted in **Table 4.** 7.

Table 4. 7 Ablation Trials consider Caption Branch Only, Visual Branch w/o attention, and Visual Branch with Attention

Methods	**Datasets**					
	MultiOff		**Hateful Memes**		**MMHS150K**	
	Accuracy	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
KDAC (Text Branch Only)	0.5667	0.3937	0.6585	0.5302	0.7437	0.4909
MSKAV w/o Attention	0.5333	0.5502	0.3750	0.4333	0.7500	0.3846
MSKAV with Attention	0.6117	0.6263	0.7607	0.8281	0.8025	0.6958
Proposed	**0.6250**	**0.6557**	**0.8750**	**0.8363**	**0.8078**	**0.7665**

The ablation study leads to the observations with the following conclusions:

- The proposed architecture, which is a dual-branch architecture, yields the best results on all three datasets.
- The removal of any of the branches results in inferior prediction results.
- The removal of an attentional layer from MSKAV is also producing low prediction results.

From the above observations, it can be concluded that the proposed architecture is essential and contributes to the overall performance in detecting hate content in memes.

4.2.4.4 Generalization Study

Despite the outstanding performance on multimodal hate content datasets, most recently developed methods mainly rely on in-depth design analysis within the confines of the dataset, neglecting the generalization ability that is necessary when techniques must examine examples from different domains or datasets. Consequently, in order to demonstrate the resilience of the model, this section presents a cross-dataset analysis rather than evaluating the quality of the suggested architecture solely on one dataset. As the model performance is evaluated on three publicly available datasets (MultiOff, Hateful memes, and MMHS150K), therefore, the proposed model is trained on the combination of two datasets and tested on the third dataset to show the generalizability of the architecture (**Table 4.** 8).

Table 4. 8 Depicts the generalization study when the combination of two datasets is used for training, and the third is used for testing

Train Dataset	**Test Dataset**	**ACC**	**P**	**R**	**F1**	**AUC**	**MCC**
MultiOff + MMHS150K	Hateful Memes	0.8045	0.7993	0.5711	0.6661	0.7789	0.4088
MultiOff + Hateful Memes	MMHS150K	0.7238	0.6123	0.6793	0.6440	0.6994	0.3558
MMHS150K + Hateful Memes	MultiOff	0.5997	0.6482	0.6384	0.6432	0.6189	0.1964

The graphs represent the accuracy (**Fig. 4.** 7) and AUC (**Fig. 4.** 8) comparisons for the same and cross-dataset evaluation.

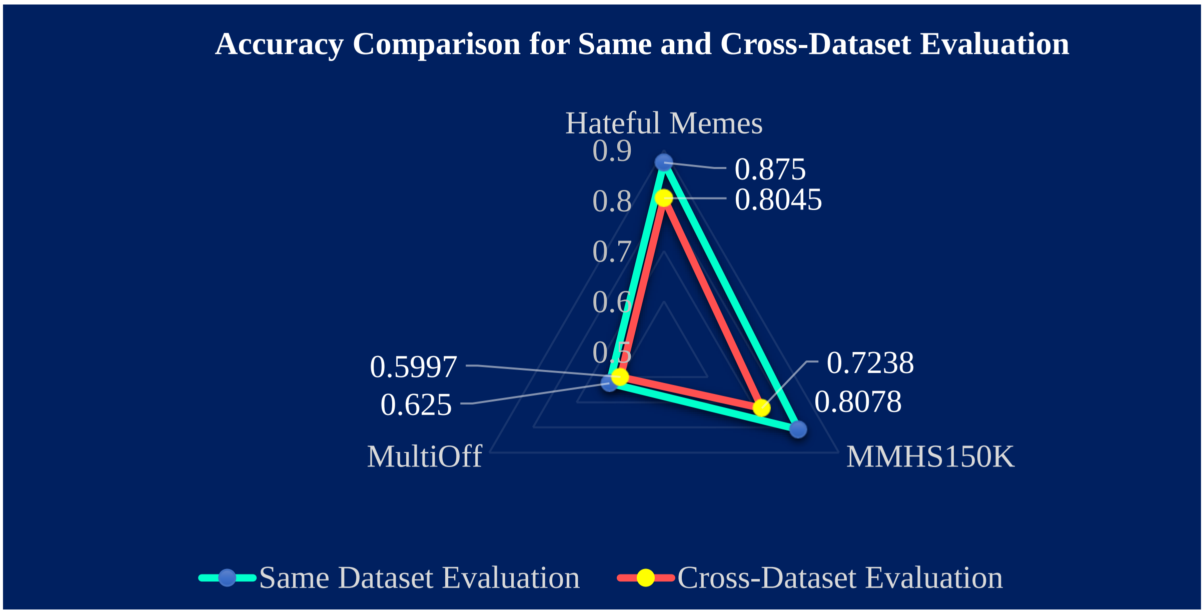


Fig. 4. 7 Accuracy Comparison for Same and Cross- dataset Evaluation

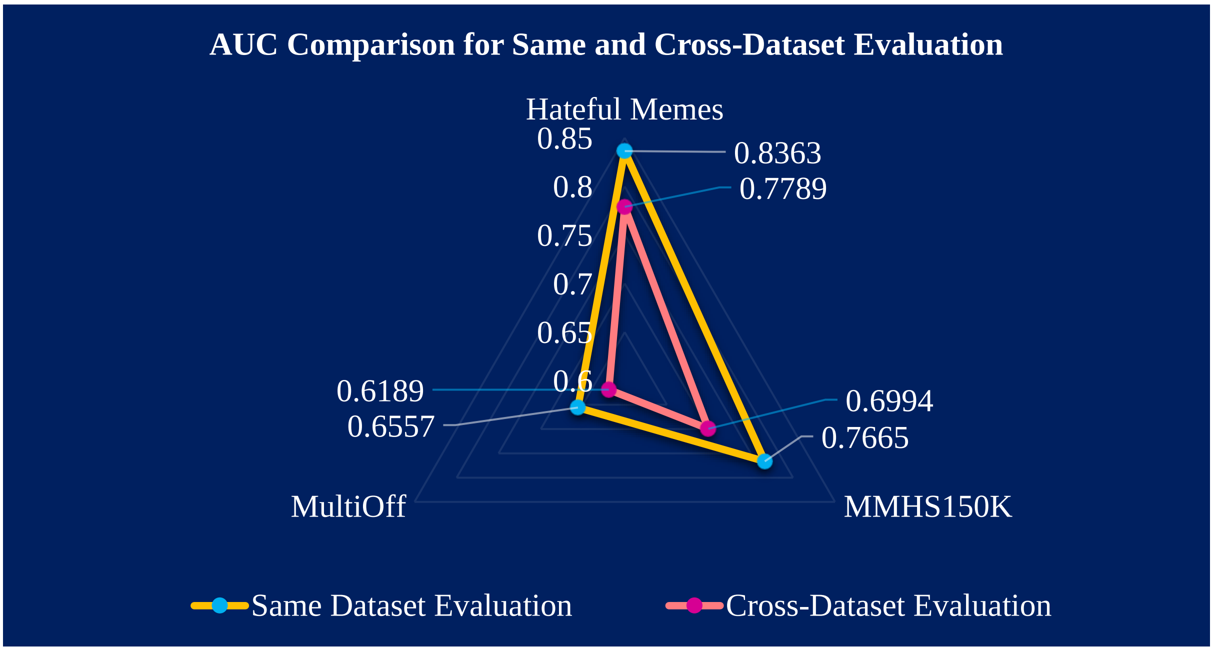


Fig. 4. 8 AUC Comparison for Same and Cross- dataset Evaluation

From the above-presented analysis, we can conclude that there is a slight degradation in the performance but not reduced too much when taken as a whole, which shows strong generalization in terms of the reliability and robustness of the proposed model.

4.2.5 Discussion

The proposed architecture intends to classify multimodal data into two classes: Hate and No-Hate. Although there exist many forms of hate but two major categories of hate can be classified as explicit and implicit hate. Accurate detection and classification of online hate is a difficult task. Implicit hate identification is indeed a challenging task as such content tends to have unusual syntax, polysemic words, and fewer markers of prejudice (slurs). The words as vectors and syntax representations as co-occurrence information can be combined to detect implicit hate content online. The other way for implicit hate identification in memes is to find the stronger correlation of visual features with deeper semantic contextualization of text through the inclusion of multimodal data. Also, to work on various forms of hate speech, one must possess a strong definition of hate speech, as there are many definitions of hate speech available. The presence of insulting terms in a text might not convey hate speech, according to some researchers, while it can be true for others. Afterward, it is very important to identify the right approach towards effective text mining, information retrieval, better feature exploration and feature selection, and applying non-linear models or weakly supervised learning approaches. Moreover, due to the inherent complexity of hate content detection, it becomes essential to differentiate among various forms of hate as it might be offensive for many people, and others can consider it as hate. The majority of the challenge is related to the quality of publicly available datasets. Thus far, the available multimodal datasets are not intended to perform hierarchical multi-label classification. The proposed model can follow two approaches on hierarchical multi-label datasets: a) training the model to initially predict all the irrelevant tweets, which learns to classify them to all the relevant target entities simultaneously, and b) training 'n' one vs. all binary classifier approach for each of the topics to encode the inter-label dependencies among co-occurring labels.

4.2.6 Conclusion

This study introduces a novel dual-branch network composed of knowledge distillation attention for extracting the essential information from the caption modality and multi-kernel attention for collecting pertinent information from the images. Extensive testing on three publicly accessible datasets showed that the suggested architecture outperformed baseline models, claiming better results in terms of accuracy and AUC scores. We also conducted numerous ablation trials on the datasets to conclude that the proposed architecture is contributing to the performance of hate content identification in memes. Thorough experimentation on multi-modal hate speech benchmarks MultiOff, Hateful Memes, and MMHS150K datasets achieved accuracy scores of 0.6250, 0.8750, and 0.8078, respectively. It also reaches impressive AUC scores of 0.6557, 0.8363, and 0.7665 on the three datasets, respectively, beating SOTA multi-modal hate speech identification models.

4.3 MHS-STMA: Multimodal Hate Speech Detection via Scalable Transformer- Based Multilevel Attention Framework

4.3.1 Abstract

Social media has a significant impact on people's lives. Hate speech on social media has emerged as one of society's most serious issues in recent years. Text and picture are two forms of multimodal data that are distributed within articles. Unimodal analysis has been the primary emphasis of earlier approaches. Additionally, when doing multimodal analysis, researchers neglect to preserve the distinctive qualities associated with each modality. To address these shortcomings, the present article suggests a scalable architecture for multimodal hate content detection called transformer-based multilevel attention (STMA). This architecture consists of three main parts: combined attention-based deep learning mechanism, a vision attention-mechanism encoder, and a caption attention-mechanism encoder. To identify hate content, each component uses various attention processes and handles multimodal data in a unique way. Several studies employing multiple assessment criteria on three hate speech datasets—Hateful memes, MultiOff, and MMHS150K— validate the suggested architecture's efficacy. The outcomes demonstrate that on all three datasets, the suggested strategy performs better than the baseline approaches.

4.3.2 Proposed Architecture

The details of the proposed architecture (**Fig. 4. 9**) are presented in this section.

4.3.2.1 Problem Definition

A set of multimodal samples $M = \{m_1, m_2, \dots, m_n\}$ is given, where each $m_i \in M$ has an image I_i with the corresponding target T_i and captions with w_i words. Attached to each T_i is a label y_i, which may be hate or no-hate. To achieve a uniform distribution of both modalities, we first eliminated those cases from the datasets that contain either caption or image data. Images and text undergo different preprocessing steps. The natural language toolkit (NLTK) package is used to preprocess text input. It assists in eliminating stop words and stemming and lemmatizing words to return them to their root form. Images are scaled and their mean is subtracted to achieve normalization. In addition, we have employed several data augmentation methods such as flipping, rotating, zooming, and so on to prevent the model from being overfit to the training set. Using the proposed STMA framework, our aim is to predict the proper label for the collection of unseen samples.

4.3.2.2 Patch Embeddings

Every image I_i is separated into smaller patches, and each one makes use of a 16×16 convolution with a stride of 16. The fixed-size patches from the batch of input photos with the shapes $(b, h, w, and\ c)$ are flattened to create the flat patches. We apply a trainable embedding vector of dimension d to these patches. This provides us with a linear embedding of the flattened patches in low dimensions. To obtain a consolidated representation of all the patches, a learnable token is prepended to the patch embeddings. After that, we include the positional embeddings so that the transformer model is fully aware of the image sequence. We are adding the spatial data associated with every patch in the series in this way.

4.3.2.3 Vision Attention-Mechanism Encoder

The transformer attention-based encoder receives the patched embeddings produced in the previous section and uses them to learn the abstract features. We have employed the Vision transformer as the foundational framework for the visual data. The following elements are essentially included in the encoder module: layer normalization (Norm), MLP, and MSA. Self-attention has the advantage of being able to extract information from the full visual globally. Consequently, the MSA block splits the inputs into numerous heads, each of which is capable of learning and comprehending the various facets of the input's abstract representation. All the heads' output is combined and sent to the MLP layer, which makes use of the GeLu nonlinearity. To cut down on the amount of time the network needs to train, layer normalization is applied before each layer. Additionally, residual connections are used to get around the issue of the vanishing gradient.

4.3.2.4 Text Attention-Mechanism Encoder

The bidirectional encoder (BERT) representation from Transformers [80] is used to encode the raw text sequences, once more making use of the attention mechanism. Token embeddings, segment embeddings, and position embeddings are combined to turn the text sequences into tokens. Token embeddings (Ti) provide the vocabulary IDs for each token, sentence embeddings (Si) aid in sentence differentiation, and position embeddings (Pi) show the word positions inside sentences. Every embedding layer is linked to the sublayers before it and comprises distinct MSA sublayers. The discriminative characteristics separating the text and image modalities are not learned by the multimodal analysis works now in use. It becomes

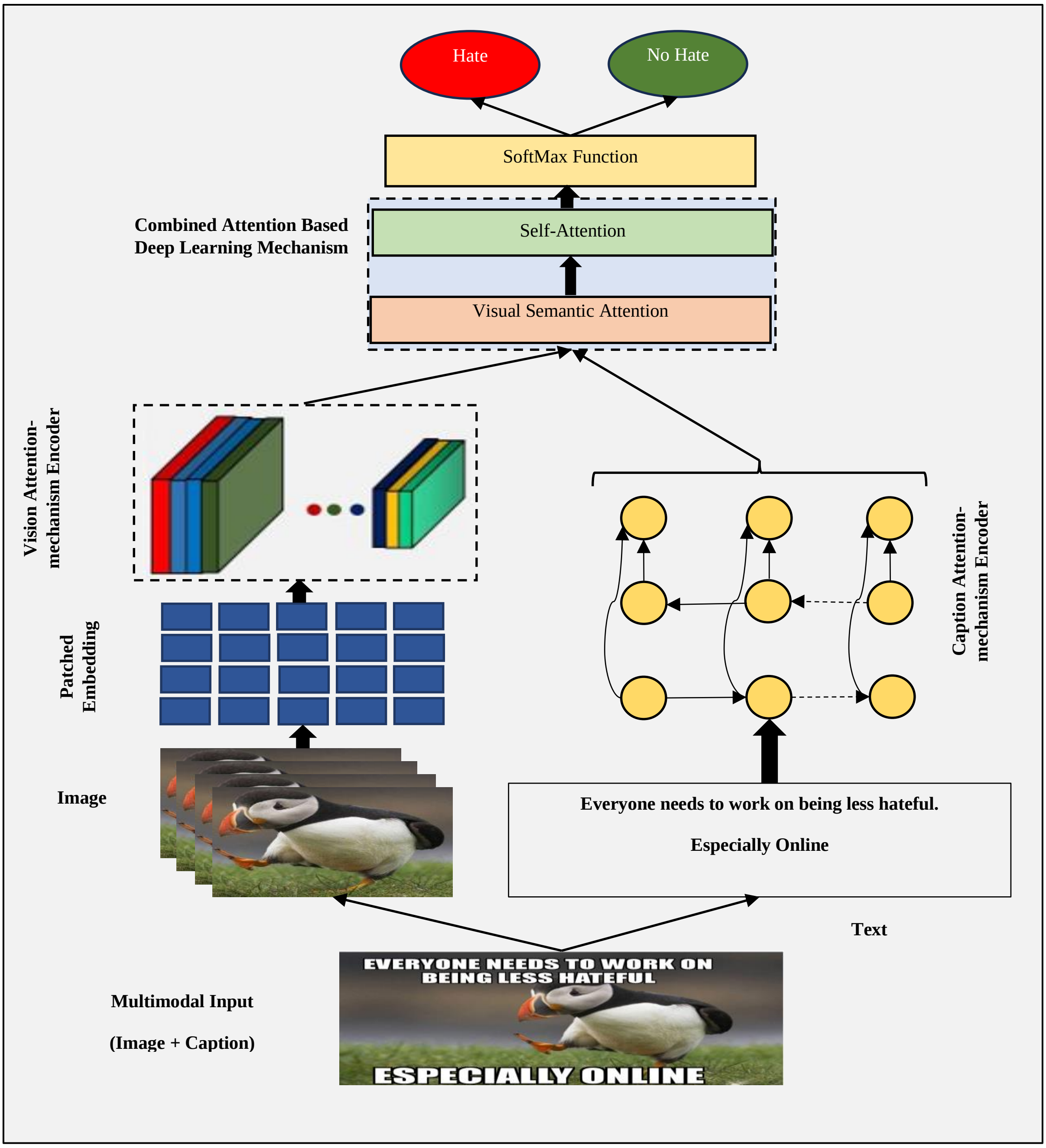

Hate
No Hate
SoftMax Function
Combined Attention Based Deep Learning Mechanism
Self-Attention
Visual Semantic Attention
Vision Attention-mechanism Encoder
Patched Embedding
Caption Attention-mechanism Encoder
Image
Everyone needs to work on being less hateful.
Especially Online
Text
Multimodal Input
(Image + Caption)
EVERYONE NEEDS TO WORK ON BEING LESS HATEFUL
ESPECIALLY ONLINE

Fig. 4. 9 Proposed STMA Architecture

essential to investigate the complementing information between the various modalities in multimodal feature learning. This will improve our model's overall performance.

4.3.2.5 Combined Attention-Based Deep Learning

Two modules are used in combined attention-based deep learning mechanism to accomplish this. Initial module is the visual semantic attention block, which creates multimodal features by extracting important picture aspects from attended text information. A self-attention block in the second module eliminates features from the multimodal data that aren't needed.

- The goal of the **visual semantic attention block** is to understand which image features to prioritize, using the words in the caption sequence. The visual semantic attention block receives an image-caption pair $\{I_i, C_i\}$ for the ith sample. Element-wise multiplication is utilized to combine two modalities to achieve this.
- Several modalities collaborate in the **self-attention block** to determine which feature should be prioritized and to calculate the attention of all the inputs in relation to one another. This is crucial since merging the modalities could produce a lot of unrelated features. The interaction between the multimodal elements—which include both text and image features—allows for the identification of the features that require additional attention. Because of this, the self-attention block will combine the attention of all the inputs with respect to one another, highlighting the various multimodal features based on their weights.

Finally, the SoftMax classifier receives the final features acquired and uses them for classification. A probabilistic activation function is called SoftMax. For every output label, it provides the likelihood that the label belongs to the class. The output chosen for the final class is the one with the highest probability. The algorithm for the proposed architecture is given in **Table 4.** 9.

Table 4. 9 Algorithm for the proposed STMA Architecture

```
Algorithm 1: Multimodal Hate Speech Detection via Scalable Transformer-Based Multilevel Attention Framework
Input: Set of multi-modal samples $M = \{m_1, m_2, \dots, m_n\}$. Each $m_i \in M$ contains captions with $w_i$ words and an image $I_i$ with an associated target $T_i$. Each $T_i$ is attached with label $l_i$.
Output: Hate Content Classification task as hate, or no-hate
6. Patch Embedding:
    • Split image $I_i$ into patches of $16 * 16$ convolution having stride 16.
    • To generate the embedding, multiply with the embedding vector.
    • Add positional embedding to create the patched embedding.
7. Vision Attention-mechanism Encoder
    • To understand the input's abstract representation, divide the input patches into several heads.
    • Combine all head outputs and pass them to the MLP layer which contains one hidden and an output layer.
8. Caption Attention-mechanism Encoder
    For a sequence of 'n' words:
        Encode the captions sequence by token, sentence and position embedding as:
            $E(f_i) = \{T_i + S_i + P_i\} \forall\ i = 1,2,\dots,n$
9. Combined Attention-based deep Learning mechanism
    • Pass the multimodal sample into the block of visual semantic attention.
    • Use self-attention to eliminate any characteristics that are unnecessary.
    • Utilize the SoftMax classifier to categorize the input sample as either hateful or not.
10. End
```

4.3.3 Experimental Setup

This part outlines the specific experimental parameters employed in this study to assess the efficacy of the proposed approach.

4.3.3.1 Dataset Description

The following multimodal datasets have been used to train and evaluate the efficacy of the proposed framework in detecting hateful memes.

4.3.3.1.1 Multi Modal Hate Speech Dataset (MMHS150K):

In [101], a multimodal hate speech dataset, MMHS150K, consisting of 150,000 tweets was created. Each tweet in the dataset includes both textual content and an accompanying image. Twitter API was utilized to collect real-time tweets. To guarantee the inclusion of both visual and textual information within all dataset instances, the authors eliminated the tweets containing textual images.

4.3.3.1.2 Multimodal Meme Dataset for Offensive Content (MultiOff):

The authors in [87] constructed a multimodal dataset consisting of 743 memes, categorized into offensive or non-offensive classes, by drawing upon the 2016 U.S. Presidential Election as a point of reference for identifying offensive content on social media.

4.3.3.1.3 Hateful Memes Challenge (HMC):

[110] introduced a challenging dataset for the identification of hate speech in memes. The dataset is constructed in such a manner that unimodal approaches fail to classify the memes accurately, and only multimodal frameworks can categorize them effectively. This is achieved by introducing confounder samples into the dataset, which makes it difficult to rely on a single modality.

4.3.3.2 Experimental Settings and Hyperparameters

The details regarding the experimental hyperparameter settings for the different datasets, including the number of epochs, batch size, initial learning rate, and optimizer are given in **Table 4.** 10.

Table 4. 10 Hyperparameters Settings

Dataset	Number of Epochs	Batch Size	Learning Rate	Optimizer
MMHS150K	10	32	0.001	Adam
MultiOff	40	8	0.001	Adam
HMC	20	16	0.001	Adam

4.3.3.3 Data Pre-Processing

This section outlines the pre-processing procedures implemented in the present experiment. The dimensions of all images are adjusted to a uniform size of 3×256×256. The pixel values undergo normalization, resulting in a range of [0,1].

4.3.3.4 Train, Validation, and Test Splits

This section contains the total number of samples in each of three datasets. The ratio of training, validation and testing sets is 8:1:1, respectively is shown in **Table 4. 11**.

Table 4. 11 Dataset size (total, training, testing and validation)

Dataset	**Size**	**Training Set**	**Validation Set**	**Testing Set**
MMHS150K	150000	120000	15000	15000
MultiOff	743	600	70	70
HMC	8496	6800	840	840

The experiments are conducted using two NVIDIA TITAN RTX GPUs, each with a memory capacity of 24 GB, operating simultaneously.

4.3.4 Results and Discussion

This section presents a comprehensive analysis of the results obtained.

4.3.4.1 Performance and Comparison against SOTA on Benchmark Datasets

The results of the suggested architecture on the MMHS150K, MultiOff, and HMC datasets are presented in this section. The accuracy, precision, recall, F1 score, and area under the curve values are displayed in **Table 4.** 12 along with the comparison against SOTA approaches. The proposed method for improved multimodal hate speech detection efficiently extracts important information from both textual and visual modalities. The accuracy values of 0.6509, 0.8790, and 0.8088 respectively obtained on MultiOff, HMC, and MMHS150K datasets show a significant improvement in performance. In comparison to earlier research, the AUC ratings of 0.6857, 0.8500 and 0.7840 also show a significant improvement in performance.

Table 4. 12 Performance and Comparison

	Ref	**Acc**	**P**	**R**	**F1**	**AUC**
MultiOff	[87]	-	0.4000	0.6600	0.5000	-
	[99]	-	0.6450	0.6510	0.6480	-
	Ours	**0.6509**	**0.6740**	**0.6940**	**0.6839**	**0.6857**

	Ref	Acc	P	R	F1	AUC
Hateful Memes	[110]	0.6947	-	-	-	0.7544
	[99]	0.7580	-	-	-	0.8280
	[111]	0.7650	-	-	-	0.8374
	[100]	0.7108	0.7000	-	0.6900	0.7141
	Ours	**0.8790**	**0.8348**	**0.6140**	**0.7678**	**0.8500**
MMHS150K	[101]	0.6850	-	-	0.7040	0.7340
	[112]	0.7143	-	-	0.7085	-
	[113]	-	-	-	-	0.7149
	[114]	-	0.6133	0.5134	0.5589	-
	[115]	0.7401	-	-	-	0.7634
	Ours	**0.8088**	**0.7108**	**0.7388**	**0.7246**	**0.7840**

4.3.4.2 Ablation Trials

To examine the impact of the individual components in our suggested architecture, we do ablation research in this part. We do the multi-modal analysis on all the datasets after first conducting the uni-modal analysis on the caption and vision data independently. **Table 4.** 13 provides a summary of the findings.

4.3.4.2.1 Uni-modal Analysis

The caption input goes through a caption attention-mechanism encoder, which is then followed by self-attention for the caption modality. The features that have been extracted are sent to the softmax layer for the last stage of classification. For the visual aspect, we create patched embeddings and send them to the visual attention-mechanism encoder module, then implementing the self-attention mechanism. The ultimate characteristics are passed straight to the softmax classifier. In both scenarios, the visual semantic attention block is removed because we are working with unimodal data exclusively.

4.3.4.2.2 Multi-modal Analysis

In multi-modal analysis, we assess the importance of each component by removing different elements from our proposed framework. The visual semantic block's multimodal features are sent to the softmax classifier without considering the self-attention block. Afterwards, we remove the visual-semantic attention block from the architecture, considering both the self-attention block and softmax layer.

The significance of integrating the semantic correlation between visual and caption features is evident in **Table 4.** 13. Next, the vision attention-mechanism encoder block is taken out, the patched embeddings are sent through the pretrained VGG-16 model, and combined-attention based deep learning mechanism is carried out. The findings clearly confirm the significance of our vision attention-focused encoder block in capturing the unchanged characteristics of the images. Ultimately, we disable the encoder that focuses on captions and observe that attention to the captions plays a vital role, as it highlights key words and assists in setting the context.

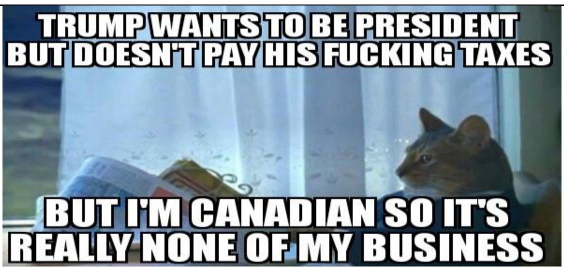
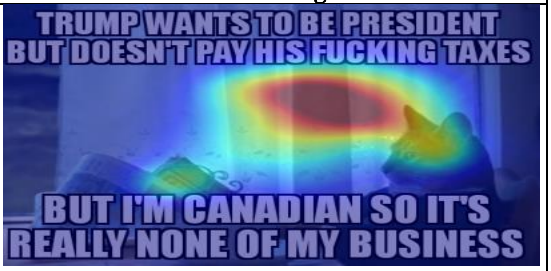
Table 4. 13 Ablation Scores

	Model	**Accuracy**		
		MultiOff	**Hateful Memes**	**MMHS150K**
Unimodal	Textual	0.5667	0.6585	0.7437
	Visual	0.5333	0.3750	0.7500
Multimodal	Without Visual Semantic Attention	0.5989	0.6900	0.7689
	Without Self Attention	0.5764	0.6756	0.7490
	Without Vision Attention-mechanism encoder	0.6091	0.7501	0.7736
	Without Caption Attention-mechanism encoder	0.6117	0.7607	0.8025
	Proposed	**0.6509**	**0.8790**	**0.8088**

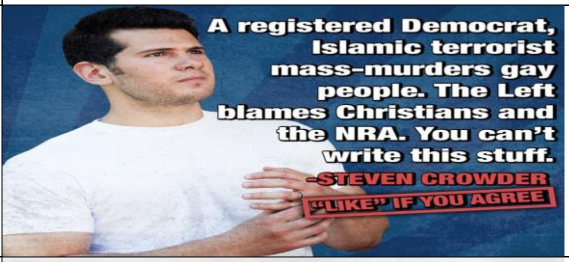
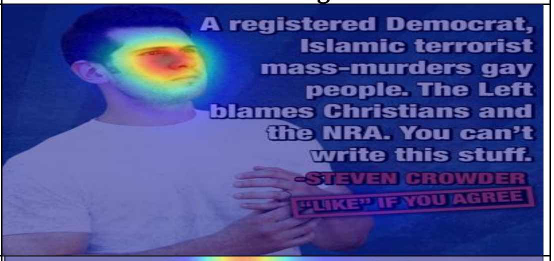
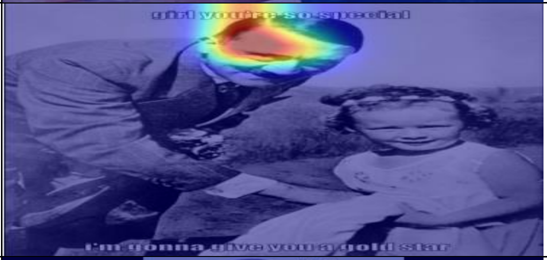
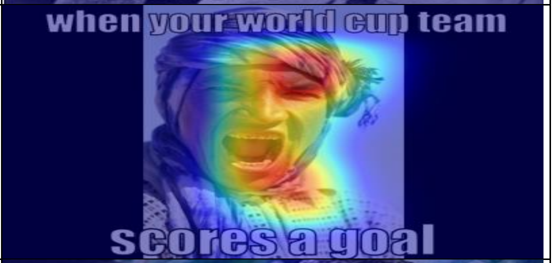
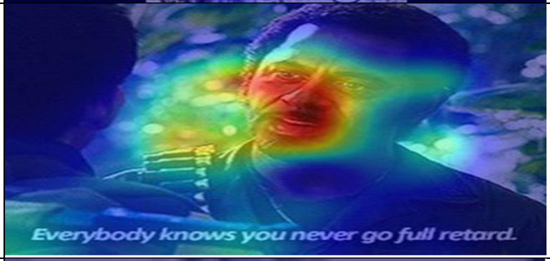
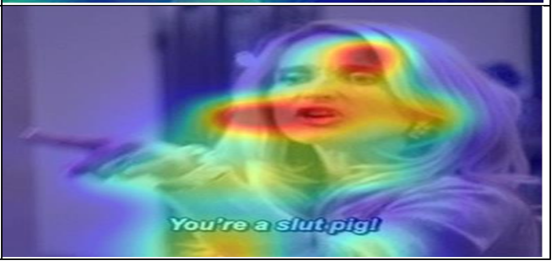
4.3.4.3 Qualitative Visualization

Memes' captions and visual portions both include a substantial quantity of information that is undeniable. The informative portion of the image is represented by the spatial region. It locates the pertinent visual components based on the visually attended elements. The activation mapping via attention approach, i.e., GradCAM, [116] is visualized in **Table 4.** 14 to find the fine-grained localization of objects. GradCAM requires a gradient to be present on a given layer to capture the target layer's attention.

Table 4. 14 Spatio-Region of Importance via GradCAM

	Memes	**Focused Region**
MultiOff		

TRUMP WANTS TO BE PRESIDENT BUT DOESN'T PAY HIS FUCKING TAXES
BUT I'M CANADIAN SO IT'S REALLY NONE OF MY BUSINESS
TRUMP WANTS TO BE PRESIDENT BUT DOESN'T PAY HIS FUCKING TAXES
BUT I'M CANADIAN SO IT'S REALLY NONE OF MY BUSINESS

	Memes	Focused Region
Hateful Memes	 A registered Democrat, Islamic terrorist mass-murders gay people. The Left blames Christians and the NRA. You can't write this stuff. -STEVEN CROWDER "LIKE" IF YOU AGREE	 A registered Democrat, Islamic terrorist mass-murders gay people. The Left blames Christians and the NRA. You can't write this stuff. -STEVEN CROWDER "LIKE" IF YOU AGREE
		
		
MMHS150K		
		

The observations from **Table 4.** 14 are:

- Improving object localization by concentrating on the designated spatial region.
- The GradCAM activation map ascertains the influence of each region on a model's output.

4.4 Conclusion

Social media platforms have facilitated various forms of communication, allowing for the extensive and rapid exchange of thoughts. These platforms attract millions of users who actively participate in the posts being circulated. This study offers a novel multimodal framework that can effectively filter out hateful memes. The proposed architecture comfortably outperforms the existing baselines by a significant margin, thus establishing the efficacy of the suggested methodology. The scarcity of research articles focused on investigating multimodal hate content detection indicates the extensive range of unexplored research opportunities. The remarkable performance of the proposed architecture motivates us to extend its application to other prominent multimodal areas, including sentiment analysis, sarcasm detection, and fake news detection.

4.5 Significant Outcomes of this Chapter

The significant outcomes of this chapter are as follows:

- Proposed a novel twofold branch architecture named multi-scale kernels for visuals and a knowledge distillation-based model for texts. A novel multi-branch attentional module is introduced to collect pertinent data using visual modality, and a branch of the attentional tokenizer for the caption modality is created to extract prominent features. The "multi-scale kernel attentive visual" (MSKAV) module uses an efficient multi-branch structure to extract discriminative visual features from memes. MSKAV contains multi-scale kernels to have receptive fields of multiple resolutions. Additionally, MSKAV incorporates multi-directional visual attention to highlight spatial regions of importance and the "knowledge distillation-based attentional caption" (KDAC) module has self-attention-based transformer architecture for efficient sequence learning from meme text.
- Proposed an STMA framework that effectively models the interactions between textual and non-textual characteristics in multi-modal data by combining the strength of attention processes at multiple levels. The suggested technique successfully captures the semantic connections between the textual and visual characteristics by a cross-attention mechanism. Additionally, the multihead attention (MHA) mechanism ios provided, which integrates data from various attention levels. To be more precise, the framework would employ several heads of attention to handle various components of the multimodal data. This would enable a broad variety of interactions between the textual and non-textual characteristics to be captured by the framework.

The following research works form the basis of this chapter:

- ❖ **Anusha Chhabra**, Dinesh Kumar Vishwakarma. "Multimodal Hate Speech Detection via Multi-Scale Visual Kernels and Knowledge Distillation Architecture." Published in

Engineering Applications of Artificial Intelligence (126), 2023, (Pub: Elsevier).

❖ **Anusha Chhabra**, Dinesh Kumar Vishwakarma. "MHS-STMA: Multimodal Hate Speech Detection via Scalable Transformer- Based Multilevel Attention Framework" **https://arxiv.org/submit/5841081/view**

CHAPTER 5
ROBUST ANALYSIS OF HATE CONTENT USING MULTIMODAL DATA

5.1Scope of this Chapter

This chapter studies the tradeoff between performance and computational complexity for different visual attention mechanisms in a face manipulation detection model. Specifically, five recently proposed visual attention models are integrated with a baseline deep learning model, and their relative performance and computational costs are evaluated. Experimental results clearly indicate that an increase in the computational cost of the visual attention mechanism does not necessarily predict a similar increase in the performance in detecting facial manipulation.

5.2Multimodal Hate Speech Identification in Memes Based on Hypergraph

5.2.1 Abstract

Since hateful memes are widely used in fields including sentiment mining, social media analytics, and electronic healthcare, they have attracted more attention in recent years. Previous studies have mostly concentrated on sequence learning and graph-based methods, but they have neglected the long-term dependencies within each modality as well as the high-order interactions between various modalities. This chapter presents a unique hypergraph-based approach for multimodal hate identification in memes (MHM-HGraph) to tackle these issuesand enhancing the robustness of the model. MHM-HGraph is a tool for extracting features from two modalities: visual and text. It builds intra-modal and inter-modal hypergraphs (Intra-HGraph and Inter-HGraph) using hyperedges, treating each modality utterance in a meme as a node. After then, hypergraph convolutional networks are used to update the hypergraphs, learning non-linear relationships through enhanced connections. The suggested model was assessed on three benchmark datasets: MultiOff, Hateful memes, and MMHS150K.

5.2.2 Methodology

In the task of multimodal hate speech analysis, MHM-HGraph model is proposed as shown in **Fig. 5.** 1. The proposed architecture comprises of individual modality feature extraction separately (Text and Vision), hypergraph network, and a prediction layer.

5.2.2.1 Problem Definition:

Each hate meme (M_i^T, M_i^V) is represented in two modalities: V for Visual and T for Text. The objective of MHM-HGraph is to learn the interaction across two modalities.

5.2.2.2 Single Modality Feature Extraction:

To extract features from the visual and text modalities, respectively, we use DenseNet and Text CNN. To process the elements of the visual modality and increase their representational capacity, we use fully connected networks (**Eqn. 5.1**). To extract contextual information from the meme for the text modality, we use a bidirectional LSTM network (**Eqn. 5.2**). The computational process for both the modalities can be represented as follows:

$$x_i^V = W_e^V M_i^V + b_i^V \tag{5.1}$$

$$x_i^T = LSTM(M_i^T, h_{i-1}^T, h_{i+1}^T) \tag{5.2}$$

Where, (M_i^T, M_i^V) represents the input for text and visual modalities, and (x_i^T, x_i^V) denotes the encoded output for text and visual modalities.

5.2.2.3 HyperGraph Construction:

We employ the retrieved features as input to build both intra- and inter-modal hypergraphs to capture the interactions between two modalities. Within each modality of intra-HGraph, two types of hyperedges are formed, indicated by the notation $(\in^T, \in^V)$. Every node in the same modality is linked to the Past P and Future F context nodes. Interaction between two modalities is referred to as an Inter-HGraph. The Intra-HGraph and the Inter-HGraph share the same

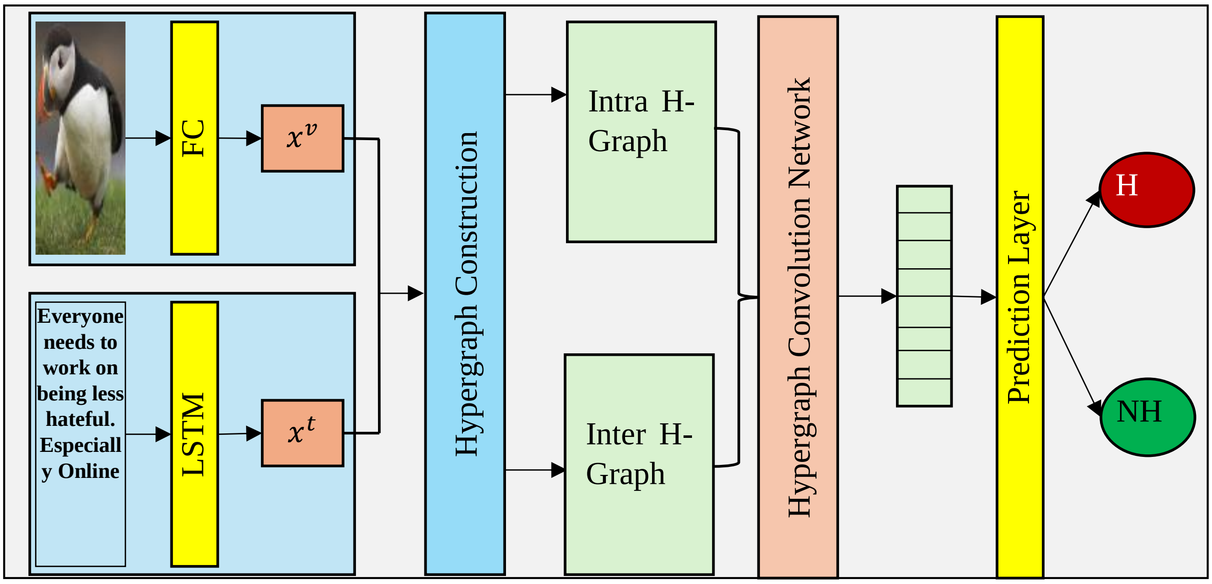

FC
x^v
Everyone needs to work on being less hateful. Especially Online
LSTM
x^t
Hypergraph Construction
Intra H-Graph
Inter H-Graph
Hypergraph Convolution Network
Prediction Layer
H
NH

Fig. 5. 1 Proposed Architecture

nodes. By creating inter modality hyperedges, we link each node to nodes that are part of several modalities ($\in^1,\ and\ \in^2$).

5.2.2.4 Hypergraph Convolution:

To effectively propagate information between vertices, the hypergraph convolution makes efficient use of local clustering structures and higher order interactions. The two stages of the hypergraph convolution method's multimodal hate speech detection procedure are information aggregation from vertices to hyperedges and information aggregation from hyperedges to vertices.

To be more precise, each vertex's data is combined to create the matching hyperedge, which produces a representation for every hyperedge. Subsequently, the hyperedges linked to every vertex are identified, and their data is combined into the vertex, producing a representation for every vertex. We combine the information from vertex $u_{i-1}^T, u_i^T, u_{i+1}^T$ to edge $\in^T$and the information from vertex $u_{i-1}^V, u_i^V, u_{i+1}^V$ to edge $\in^V$ in the Intra-HGraph.

We combine the information of vertex u_{i-1}^T, u_{i-1}^V to edge $\in^1$, the information of vertex u_i^T, u_i^V to edge $\in^2$, and the information of vertex u_{i+1}^T, u_{i+1}^V to edge $\in^3$in the Inter-HGraph. This procedure further improves the learning process by giving MHM-HGraph representations for every vertex in both intra-modal and cross-modal aspects.

5.2.3 Experiment Analysis

We first provide the datasets, implementation specifics, and performance metrics in this section. Next, we compare MHM-HGraph's effectiveness with several reliable baselines. Finally, a thorough analysis is carried out on three benchmark datasets: MultiOff [87], Hateful Memes [110], and MMHS150K [101].

5.2.3.1 Datasets

Our research has been conducted using three publicly accessible datasets that are particularly associated with hatred or offensive content. The first problem with these datasets is that most photos have multiple meanings implied by both the text and the visual information. The second problem has to do with the differences in image channel sizes.

5.2.3.1.1 MultiOff

The combination of memes from several social media networks makes the multimodal memes dataset (MultiOFF). The 2016 U.S. Presidential Election Event dataset consisted of 73 memes created from a set of manually annotated photo URLs and text discovered in the photos. All unnecessary attributes, such as likes, upvotes, timestamps, etc., were already removed during the dataset preparation procedure. Next, the dataset is split into three files: the Training file, Testing file, and Validation file. These files each contain 445, 149, and 149 memes.

5.2.3.1.2 Hateful Memes

Facebook AI published a dataset with the purpose of identifying multimodal hate content in online memes. roughly 10,000 PNG images in total, further divided into training and testing files. Three annotators who have binary label classifications examine the dataset.

5.2.3.1.3 MMHS150K

About 150,000 tweets were gathered from Twitter and annotated. The MMHS150K dataset is split up into six labels: No discriminatory, racial, homophobic, sexist, or insults on other communities.

5.2.3.2 Hardware and Implementation Details

Using NVIDIA Titan RTX 24GB GPUs running in parallel with 128GB of system memory, the suggested experiment was completed. Pre-processing involved setting the pixel values of images with dimensions of 3 x 256 x 256 to [0,1]. Only after the dimensions are altered to three randomly selected channels, such as [x,y,z], [x,y,w], [y,z,w], and [x,z,w], are images with four channels with dimensions [x,y,z,w] analyzed. Approximately 80:10:10 is the split ratio used for training, testing, and validation.

The learning rate, decay rate, batch sizes, optimizers, and epoch count are the hyperparameters (see **Table 5.** 1) that are employed.

Table 5. 1 Hyperparameter Settings

Datasets	**Linear Decay**	**Optimizer**	**Batch Size**	**Learning Rate**	**No. of Epochs**
MultiOff	10%	Adam	4	0.0001	40
Hateful Memes	20%	Adam	16	0.001	20
MMHS150K	50%	SGD	32	0.001	10

5.2.3.3 Performance Metrics

Accuracy, precision, recall, F1-Score, area under the curve (AUC), were the six different performance metrics that were employed. The general calculations and ranges for the performance metrics stated above are displayed in **Table 5.** 2.

Table 5. 2 Performance Metrics

Metrics	Formula	Range
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	[0,1]
Precision	$\frac{TP}{TP+FP}$	[0,1]
Recall	$\frac{TP}{TP+FN}$	[0,1]
F1-Score	$\frac{(2 \times Recall \times Precision)}{Recall + Precision}$	[0,1]
AUC	-	[0,1]

5.2.4 Result and Comparative Analysis

In this section, the outcomes of the proposed architecture on the MMHS150K, MultiOff, and HMC datasets are shown in **Table 5.** 3, considering accuracy (ACC), precision (P), recall (R), F1 score (F1), and area under curve (AUC) scores. This is demonstrated by the accuracy values of 0.8088, 0.6290, and 0.8790 achieved on the MMHS150K, MultiOff, and HMC datasets. The AUC ratings of 0.7840, 0.6659, and 0.8500 also indicate a major improvement in performance as compared to previous studies.

Table 5. 3 Classification Results

Dataset	Acc	P	R	F1	AUC
MultiOff	0.6290	0.6730	0.6930	0.6829	0.6659
Hateful Memes	0.8790	0.8348	0.6140	0.7678	0.8500
MMHS150K	0.8088	0.7108	0.7388	0.7246	0.7840

Fig. 5. 2 displays the graphical depiction of the classification results.

5.2.4.1 Comparison against SOTA Approaches

This section seeks to evaluate and contrast the effectiveness of the proposed framework with established benchmarks on the MMHS150K, MultiOff, and HMC datasets. **Table 5. 4** offers a comparison of the metric scores. The suggested framework clearly surpasses the previously accepted methodologies by a substantial margin. The recommended architecture beats the best performing baseline by 6.69% in accuracy and 1.86% in AUC for the MMHS150K dataset. With respect to the MultiOff dataset, the designed multimodal framework triumphs over the most effective SOTA technique by 2.7% in precision, 4% in recall, and 3.5% in F-1 score.

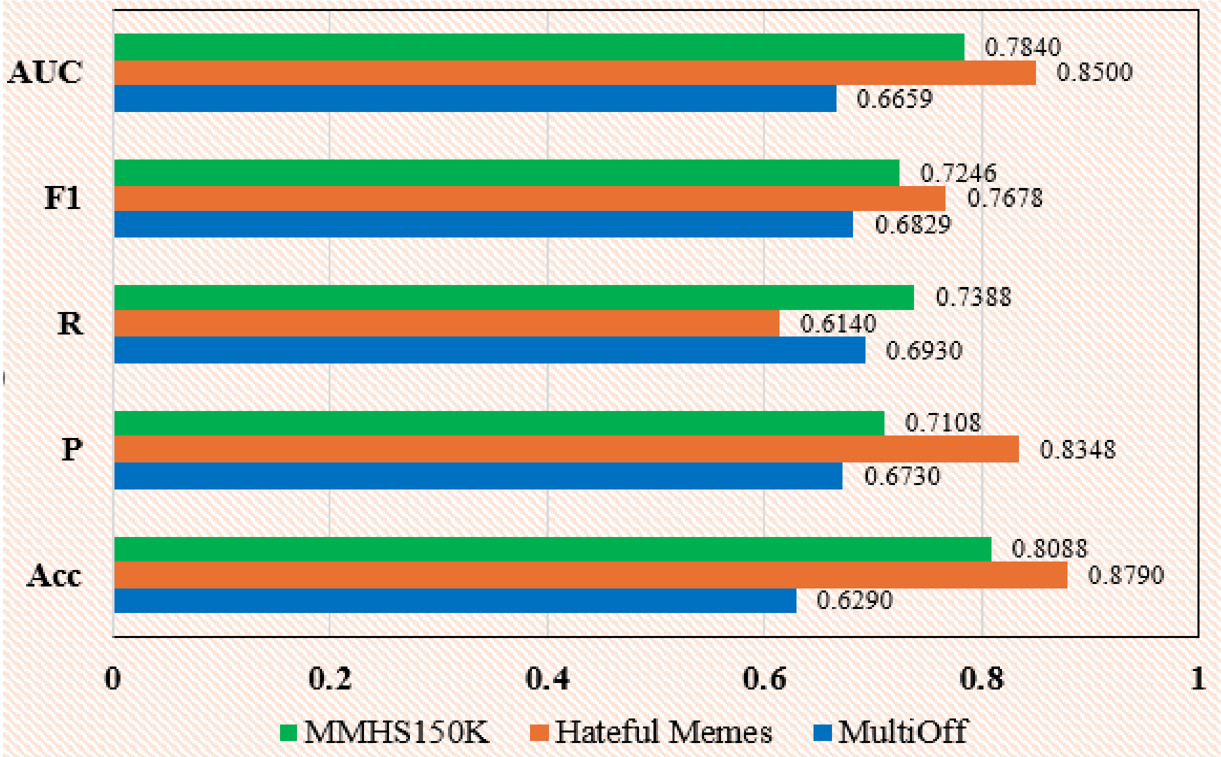


Fig. 5. 2 Classification Results

Likewise, the best baseline for the HMC dataset also falls short in comparison to the proposed framework by a significant margin of 10.6% in terms of accuracy and 1.3% in terms of AUC.

Table 5. 4 Comparison Table

	Ref	Acc	P	R	F1	AUC
MultiOff	[87]	-	0.4000	0.6600	0.5000	-
	[99]	-	0.6450	0.6510	0.6480	-
	Ours	**0.6290**	**0.6730**	**0.6930**	**0.6829**	**0.6659**
Hateful Memes	[110]	0.6947	-	-	-	0.7544
	[99]	0.7580	-	-	-	0.8280
	[111]	0.7650	-	-	-	0.8374
	Ours	**0.8790**	**0.8348**	**0.6140**	**0.7678**	**0.8500**
MMHS150K	[101]	0.6850	-	-	0.7040	0.7340
	[112]	0.7143	-	-	0.7085	-
	[115]	0.7401	-	-	-	0.7634
	Ours	**0.8088**	**0.7108**	**0.7388**	**0.7246**	**0.7840**

5.3 Conclusion

Social media platforms have facilitated various forms of communication, allowing for the extensive and rapid exchange of thoughts. These platforms attract millions of users who actively participate in the posts being circulated. Although these platforms have included social norms and protocols, it remains challenging to limit the propagation of undesirable posts that contain hate speech. Identifying hateful content from multimodal posts is a demanding endeavor. These posts may exhibit blatant expressions of hatred, or they may be influenced by the personal opinions of a particular user or community. Dependence on manual review causes a delay in the process, and the provocative content may persist online for an extended period. Therefore, it is crucial to establish effective robust systems capable of detecting hateful content on social networking sites without the need for human intervention. The proposed architecture comfortably outperforms the existing baselines by a significant margin, thus establishing the efficacy of the suggested methodology.

5.4 Significant Outcomes of this Chapter

The significant outcomes of this chapter are as follows:

- A novel robust approach MHM-HGraph is proposed to effectively capture the contextual dependencies within two modalities (Visual and Text).
- To better capture the underlying patterns within the data, this model makes use of hypergraph convolution layers to investigate the application of non-local information, identify high-order correlations on hypergraph, and exploit the “enhancement connection” to perform non-linear mapping on the features.

The following research works form the basis of this chapter:

❖ **Anusha Chhabra**, Dinesh Kumar Vishwakarma. “Multimodal Hate Speech Identification in Memes Based on Hypergraph.” **IEEE Conference: International Conferences on Signal Processing and Advance Research in Computing (SPARC-2024), Amity School of Engineering and Technology, Amity University, Lucknow, UP, India.** https://arxiv.org/submit/5841006/view

Chapter 6
Conclusion & Future Scope

6.1 Conclusion

This chapter concludes the research work done in this thesis. Overall, four novel machine and deep learning-based architectures are proposed for manipulation detection in multimedia content. The first two models are dedicated to the problem of textual hate content detection. The other two architectures are dedicated to the problem of multimodal hate content detection. The details are as follows:

- In the first approach, it is seen that reducing features using Truncated SVD along with hyper parameter tuning helped in increasing balanced accuracy and F1 score for algorithms like Logistic Regression, SVM and XGBoost when compared to the baseline results. Still, the proposed approach is lacking in handling uncertain or imprecise data.
- The second approach focuses on handling the uncertainty and vagueness in the data by implementing the fuzzy classifiers. An empirical evaluation of seven classifiers is presented for hate speech detection on two commonly used benchmarks of different data characteristics providing essential insights into their detection in terms of accuracy for their deployment in real-world applications. Fuzzy classifiers outperformed the other two classifiers out of the three.
- In the third research work, the proposed architecture is dedicated to multimodal hate content detection by presenting a dual-branch network composed of knowledge distillation attention for extracting the essential information from the caption modality and multi-kernel attention for collecting pertinent information from the images. Extensive testing on three publicly accessible datasets showed that the suggested architecture outperformed baseline models, claiming better results in terms of accuracy and AUC scores. We also conducted numerous ablation trials on the datasets to conclude that the proposed architecture is contributing to the performance of hate content identification in memes.
- In the last approach, the proposed model “MHS-STMA” explored the problem of learning complementary information between multimodal data. The architecture utilizes transformers for capturing the dependencies and relationships between the elements in a sequence. The proposed architecture also utilizes attention mechanisms at multiple levels and focuses on crucial regions in the images based on the attended textual features. We also employ self-attention at the end to remove any redundancy from the multimodal data. The experimental

results conducted on three popular datasets show that our method performs efficiently.

6.2 Future Scope

In recent years, extensive research has been conducted to detect manipulation in multimedia content. While the performance has consistently improved in detecting or localizing these manipulations, several promising research directions need to be addressed.

- **Explainable AI:** Deep learning has mostly been used as a black-box tool where the model predicts the manipulation. However, interpreting why the model predicts the given output remains a mystery. Some tools help to understand the relative significance of the learned weights. These include plotting the class activation maps (CAMs). More research work needs to be done in this direction to improve the explainability of deep models.
- **Robustness to Adversarial Attacks:** While the performance of deep-learning models has consistently increased in detecting hateful contents, recent studies indicate that these models are highly prone to adversarial attacks. Introducing noise in the input pixel values can easily vary the predictions of a trained model. Improving the robustness of deep-learning models against adversarial attacks is a crucial future work direction.
- **Deployment:** While the theoretical research has gained leaps and bounds in detecting hate content, deploying these deep-learning models remains a challenge, given their high computational costs. More research work needs to be dedicated towards deployment issues for the end-user in the form of an application or web-based framework. The increasing capacity of recent hardware facilitates the use of such heavy computational models on mobile devices.
- **Multi-modal Approaches:** Multi-modal approaches have performed better than single-modality models due to the complementary feature of learning from multiple modalities. However, this comes at the additional computational cost of having multi-branch architectures with more parameters than a single-branch model. More research needs to be done to consistently use the benefits of the multi-modal approaches without significantly adding to the associated computational cost.

References

[1] B. Dean, "'Social Network Usage & Growth Statistics: How Many People Use Social Media in 2020?,'.," [Online]. [Online]. Available: https://backlinko.com/social-media-users.

[2] M. Bose, "https://www.thequint.com/news/politics/senior-bjp-leaders-giving-india-a-free-tutorial-in-hate-speech#read-more," The Quint.

[3] R. E. Brannigan, J. L. Moss, and J. Wren, "The conversation," Fertility and Sterility. [Online]. Available: https://theconversation.com/hate-speech-is-still-easy-to-find-on-social-media-106020

[4] M. Suster, "Business Insider," Amazon's Game-Changing Cloud Was Built By Some Guys In South Africa. [Online]. Available: https://www.businessinsider.com/736-of-all-statistics-are-made-up-2010-2?r=US&IR=T%0Ahttp://www.businessinsider.com/amazons-game-changing-cloud-was-built-by-some-guys-in-south-africa-2012-3

[5] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," *Soc. 2017 - 5th Int. Work. Nat. Lang. Process. Soc. Media, Proc. Work. AFNLP SIG Soc.*, no. 2012, pp. 1–10, 2017, doi: 10.18653/v1/w17-1101.

[6] R. Cohen-Almagor, "Freedom of Expression v. Social Responsibility: Holocaust Denial in Canada," *J. Mass Media Ethics Explor. Quest. Media Moral.*, vol. 28, no. 1, pp. 42–56, 2013, doi: 10.1080/08900523.2012.746119.

[7] R. Delgado and J. Stefancic, "Images of the Outsider in American Law and Culture: Can Free Expression Remedy Deeply Inscribed Social Ills?," *Fail. Revolutions*, vol. 77, no. 6, pp. 3–21, 2019, doi: 10.4324/9780429037627-2.

[8] Techterms.com, "Facebook Definition." [Online]. Available: http://www.techterms.com/definition/facebook

[9] Youtube, "YouTube hate policy." [Online]. Available: https://support.google.com/youtube/answer/2801939?hl=en

[10] Facebook, "What does facebook consider hate speech?" [Online]. Available: https://www.facebook.com/help/135402139904490

[11] J. T. Nockleby, "Hate Speech," *In Encyclopedia of the American Constitution (2nd ed., edited by Leonard W. Levy, Kenneth L. Karst et al., New York: Macmillan, 2000)*. pp. 1277–1279, 2000.

[12] Twitter, "Twitter_Hate Definition [online]." [Online]. Available: https://support.twitter.com/articles/

[13] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros, "Hate Speech Dataset from a White Supremacy Forum," pp. 11–20, 2019, doi: 10.18653/v1/w18-5102.

[14] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Comput. Surv.*, vol. 51, no. 4, 2018, doi: 10.1145/3232676.

[15] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," *Proc. - 2012 ASE/IEEE Int. Conf. Privacy, Secur. Risk Trust 2012 ASE/IEEE Int. Conf. Soc. Comput. Soc. 2012*, pp. 71–80, 2012, doi: 10.1109/SocialCom-PASSAT.2012.55.

[16] N. Thompson, *Equality, Diversity and Social Justice*, Sixth. PALGRAVE MACMILLAN, 2016. doi: 10.1007/978-1-137-58666-7_2.

[17] R. Guermazi, M. Hammami, and A. Ben Hamadou, “Using a semi-automatic keyword dictionary for improving violent web site filtering,” *Proc. - Int. Conf. Signal Image Technol. Internet Based Syst. SITIS 2007*, pp. 337–344, 2007, doi: 10.1109/SITIS.2007.137.

[18] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, “Abusive language detection in online user content,” *25th Int. World Wide Web Conf. WWW 2016*, pp. 145–153, 2016, doi: 10.1145/2872427.2883062.

[19] Google and Jigsaw, “Perspective API.” [Online]. Available: https://perspectiveapi.com

[20] Asia Centre, “Hate speech in Southeast Asia. New forms, old rules,” 2020. [Online]. Available: https://asiacentre.org/wp-content/uploads/2020/07/Hate-Speech-in-Southeast-Asia-New-Forms-Old-Rules.pdf

[21] S. Lomborg and A. Bechmann, “Using APIs for Data Collection on Social Media,” *Inf. Soc.*, vol. 30, no. 4, pp. 256–265, 2014, doi: 10.1080/01972243.2014.915276.

[22] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, “Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines,” *IEEE Intell. Syst.*, vol. 33, no. 6, pp. 17–25, 2018, doi: 10.1109/MIS.2018.2882362.

[23] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, “Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis,” *2016 IEEE 16th Int. Conf. Data Min.*, pp. 439–448, 2017, doi: 10.1109/icdm.2016.0055.

[24] S. Poria, E. Cambria, N. Howard, G. Bin Huang, and A. Hussain, “Fusing audio, visual and textual clues for sentiment analysis from multimodal content,” *Neurocomputing*, vol. 174, pp. 50–59, 2016, doi: 10.1016/j.neucom.2015.01.095.

[25] A. E. T. Niu, S. Zhu, L. Pang, “Sentiment analysis on multi-view social data,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, p. 9517, 2016, doi: http://dx.doi.org/10.1007/978-3-319-27674-8_2.

[26] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, “Image–text sentiment analysis via deep multimodal attentive fusion,” *Knowledge-Based Syst.*, vol. 167, pp. 26–37, 2019, doi: 10.1016/j.knosys.2019.01.019.

[27] H. Ma, J. Wang, L. Qian, and H. Lin, “HAN-ReGRU: hierarchical attention network with residual gated recurrent unit for emotion recognition in conversation,” *Neural Comput. Appl.*, vol. 33, no. 7, pp. 2685–2703, 2021, doi: 10.1007/s00521-020-05063-7.

[28] S. Poria, E. Cambria, and A. Gelbukh, “Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis,” *Conf. Proc. - EMNLP 2015 Conf. Empir. Methods Nat. Lang. Process.*, no. September, pp. 2539–2544, 2015, doi: 10.18653/v1/d15-1303.

[29] O. Araque and C. A. Iglesias, “An Ensemble Method for Radicalization and Hate Speech Detection Online Empowered by Sentic Computing,” *Cognit. Comput.*, pp. 48–61, 2022, doi: 10.1007/s12559-021-09845-6.

[30] P. Burnap and M. L. Williams, “Us and them: identifying cyber hate on Twitter across multiple protected characteristics,” *EPJ Data Sci.*, vol. 5, no. 1, 2016, doi: 10.1140/epjds/s13688-016-0072-6.

[31] I. Kwok and Y. Wang, “Locate the hate: Detecting tweets against blacks,” in *Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI 2013*, 2013, pp. 1621–1622.

[32] F. Husain and O. Uzuner, “Investigating the Effect of Preprocessing Arabic Text on Offensive Language and Hate Speech Detection,” *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 21, no. 4, pp. 1–20, 2022, doi: 10.1145/3501398.

[33] A. G. Chowdhury, “ARHNet - Leveraging Community Interaction For Detection Of Religious Hate Speech In Arabic,” *Proc. 57th Annu. Meet. te Assoc. Comput. Linguist.*, pp. 273–280, 2019.

[34] Z. Waseem and D. Hovy, “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter,” pp. 88–93, 2016, doi: 10.18653/v1/n16-2013.

[35] S. Liu and T. Forss, “Combining N-gram based similarity analysis with sentiment analysis in web content classification,” *KDIR 2014 - Proc. Int. Conf. Knowl. Discov. Inf. Retr.*, pp. 530–537, 2014, doi: 10.5220/0005170305300537.

[36] E. Greevy and A. F. Smeaton, “Classifying racist texts using a support vector machine,” *Proc. Sheff. SIGIR - Twenty-Seventh Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, no. January 2004, pp. 468–469, 2004, doi: 10.1145/1008992.1009074.

[37] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” *Proc. 11th Int. Conf. Web Soc. Media, ICWSM 2017*, pp. 512–515, 2017.

[38] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” *26th Int. World Wide Web Conf. 2017, WWW 2017 Companion*, no. 2, pp. 759–760, 2017, doi: 10.1145/3041021.3054223.

[39] E. Katona, J. Buda, and F. Bolonyai, “Using N-grams and Statistical Features to Identify Hate Speech Spreaders on Twitter,” *CEUR Workshop Proc.*, 2021.

[40] Y. Mehdad and J. Tetreault, “Do Characters Abuse More Than Words?,” in *SIGDIAL 2016 - 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, 2016, pp. 299–303. doi: 10.18653/v1/w16-3638.

[41] Z. Miao *et al.*, “Detecting Offensive Language Based on Graph Attention Networks and Fusion Features,” *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 1, pp. 1493–1505, 2023, doi: 10.1109/TCSS.2023.3250502.

[42] H. Mulki, C. B. Ali, H. Haddad, and I. Babao, “Tw-StAR at SemEval-2019 Task 5 : N-gram embeddings for Hate Speech Detection in Multilingual Tweets,” *Proc. 13th Int. Work. Semant. Eval.*, pp. 503–507, 2019.

[43] S. N. Group, “Stanford NLP Group.” [Online]. Available: https://nlp.stanford.edu/

[44] C. Wang, M. Day, and C. Wu, “Political Hate Speech Detection and Lexicon Building : A Study in Taiwan,” *IEEE Access*, vol. 10, pp. 44337–44346, 2022, doi: 10.1109/ACCESS.2022.3160712.

[45] S. Liu and T. Forss, “New classification models for detecting hate and violence web content,” *IC3K 2015 - Proc. 7th Int. Jt. Conf. Knowl. Discov. Knowl. Eng. Knowl. Manag.*, vol. 1, pp. 487–495, 2015, doi: 10.5220/0005636704870495.

[46] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, “A lexicon-based approach for hate speech detection,” *Int. J. Multimed. Ubiquitous Eng.*, vol. 10, no. 4, pp. 215–230, 2015, doi: 10.14257/ijmue.2015.10.4.21.

[47] S. Agarwal and A. Sureka, “Characterizing Linguistic Attributes for Automatic Classification of Intent Based Racist/Radicalized Posts on Tumblr Micro-Blogging Website,” 2017, [Online]. Available: http://arxiv.org/abs/1701.04931

[48] F. Del Vigna, A. Cimino, F. Dell’Orletta, M. Petrocchi, and M. Tesconi, “Hate me, hate me not: Hate speech detection on Facebook,” *CEUR Workshop Proc.*, vol. 1816, pp. 86–95, 2017.

[49] M. Z. Ali, S. Rauf, K. Javed, and S. Hussain, “Improving Hate Speech Detection of Urdu Tweets Using Sentiment Analysis,” *IEEE Access*, vol. 9, pp. 84296–84305, 2021, doi: 10.1109/ACCESS.2021.3087827.

[50] C. Baydogan and B. Alatas, “Sentiment analysis in social networks using social spider optimization algorithm,” *Teh. Vjesn.*, vol. 28, no. 6, pp. 1943–1951, 2021, doi: 10.17559/TV-20200614172445.

[51] J. Pablo and J. Jiménez, “Topic modelling of racist and xenophobic YouTube comments . Analyzing hate speech against migrants and refugees spread through YouTube in Spanish,” *TEEM’21 Ninth Int. Conf. Technol. Ecosyst. Enhancing Multicult.*, pp. 456–460, 2021.

[52] H. Liu, W. Alorainy, P. Burnap, and M. L. Williams, “Fuzzy multi-task learning for hate speech type identification,” *Web Conf. 2019 - Proc. World Wide Web Conf. WWW 2019*, pp. 3006–3012, 2019, doi: 10.1145/3308558.3313546.

[53] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, no. 4–5, pp. 993–1022, 2003, doi: 10.1016/b978-0-12-411519-4.00006-9.

[54] V. Mujadia, “IIIT-Hyderabad at HASOC 2019 : Hate Speech Detection,” *CEUR Workshop Proc.*, 2019.

[55] P. Kumar and K. Varalakshami, “Hate Speech Detection using Text and Image Tweets Based On Bi-directional Long Short-Term Memory,” *2021 Int. Conf. Disruptive Technol. Multi-Disciplinary Res. Appl.*, pp. 158–162, 2021.

[56] K. Dinakar, R. Reichart, and H. Lieberman, “Modeling the detection of textual cyberbullying,” *AAAI Work. - Tech. Rep.*, vol. WS-11-02, pp. 11–17, 2011.

[57] B. L. Gaydhani Aditya, Doma Vikrant, Kendre Shrikant, “Detecting Hate Speech and Offensive Language onTwitter using Machine Learning: An N-gram andTFIDF based Approach,” in *IEEE International Advance Computing Conference 2018*, 2018.

[58] G. Gambino, R. Pirrone, and D. Ingegneria, “CHILab @ HaSpeeDe 2 : Enhancing Hate Speech Detection with Part-of-Speech Tagging,” *CEUR Workshop Proc.*, 2020.

[59] E. Erizal and C. Setianingsih, “Hate Speech Detection in Indonesian Language on Instagram Comment Section Using Maximum Entropy Classification Method,” *2019 Int. Conf. Inf. Commun. Technol.*, pp. 533–538, 2019.

[60] M. Bilal, A. Khan, S. Jan, and S. Musa, “Context-Aware Deep Learning Model for Detection of Roman Urdu Hate Speech on Social Media Platform,” *IEEE Access*, vol. 10, no. September, pp. 121133–121151, 2022, doi: 10.1109/ACCESS.2022.3216375.

[61] X. Zhou *et al.*, “Hate Speech Detection based on Sentiment Knowledge Sharing,” *Proc. 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Jt. Conf. Nat. Lang. Process.*, pp. 7158–7166, 2021.

[62] F. M. Plaza-del-Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, “Comparing pre-trained language models for Spanish hate speech detection,” *Expert Syst. Appl.*, vol. 166, no. March 2020, p. 114120, 2021, doi: 10.1016/j.eswa.2020.114120.

[63] W. Warner and J. Hirschberg, “Detecting hate speech on the world wide web,” in *Proceeding LSM ’12 Proceedings of the Second Workshop on Language in Social Media*, 2012, pp. 19–26. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390374.2390377

[64] Y. Haralambous and P. Lenca, “Text classification using association rules, dependency pruning and hyperonymization,” *CEUR Workshop Proc.*, vol. 1202, pp. 65–80, 2014.

[65] S. Abro, S. Shaikh, and Z. Ali, “Automatic Hate Speech Detection using Machine Learning : A Comparative Study,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 484–491, 2020.

[66] A. Pandey and D. K. Vishwakarma, “VABDC-Net: A framework for Visual-Caption Sentiment Recognition via spatio-depth visual attention and bi-directional caption processing,” *Knowledge-Based Syst.*, vol. 269, p. 110515, 2023, doi: 10.1016/j.knosys.2023.110515.

[67] A. Martín and D. Camacho, “Recent advances on effective and efficient deep learning-based solutions,” *Neural Comput. Appl.*, vol. 34, no. 13, pp. 10205–10210, 2022, doi: 10.1007/s00521-022-07344-9.

[68] X. Yu, N. Ma, L. Zheng, L. Wang, and K. Wang, “Developments and Applications of Artificial Intelligence in Music Education,” *MDPI Technol.*, 2023, doi: https://doi.org/10.3390/technologies11020042.

[69] A. Sharma, A. Kabra, and M. Jain, “Ceasing hate with MoH: Hate Speech Detection in Hindi–English code-switched language,” *Inf. Process. Manag.*, vol. 59, no. 1, 2022.

[70] A. M. Ali, F. A. Ghaleb, M. S. Mohammed, F. J. Alsolami, and A. I. Khan, “Web-Informed-Augmented Fake News Detection Model Using Stacked Layers of Convolutional Neural Network and Deep Autoencoder,” *Mathematics*, vol. 11, no. 9, 2023, doi: 10.3390/math11091992.

[71] A. M. Ali, F. A. Ghaleb, B. A. S. Al-Rimy, F. J. Alsolami, and A. I. Khan, “Deep Ensemble Fake News Detection Model Using Sequential Deep Learning Technique,” *Sensors*, vol. 22, no. 18, 2022, doi: 10.3390/s22186970.

[72] A. Chhabra and D. K. Vishwakarma, “A literature survey on multimodal and multilingual automatic hate speech identification,” *Multimed. Syst.*, no. 0123456789, 2023, doi: 10.1007/s00530-023-01051-8.

[73] A. Chhabra and D. Kumar, “A Truncated SVD Framework for Online Hate Speech Detection on the ETHOS Dataset,” pp. 1–4, 2023.

[74] A. Chhabra and D. K. Vishwakarma, “Fuzzy and Machine learning Classifiers for Hate Content Detection : A Comparative Analysis,” pp. 22–25, 2022.

[75] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, “Hate speech detection with comment embeddings,” in *WWW 2015 Companion - Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 29–30. doi: 10.1145/2740908.2742760.

[76] H. Watanabe, M. Bouazizi, and T. Ohtsuki, “Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection,” *IEEE Access*, vol. 6, pp. 13825–13835, 2018, doi: 10.1109/ACCESS.2018.2806394.

[77] M. Zampieri *et al.*, “SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020),” *Proc. Int. Work. Semant. Eval.*, no. OffensEval, 2020.

[78] P. Fortuna, J. Soler-Company, and L. Wanner, “How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?,” *Inf. Process. Manag.*, 2021.

[79] X. Shi, X. Liu, C. Xu, Y. Huang, F. Chen, and S. Zhu, “Cross-lingual offensive speech identification with transfer learning for low-resource languages,” *Comput. Electr. Eng.*, vol. 101, no. April, p. 108005, 2022, doi: 10.1016/j.compeleceng.2022.108005.

[80] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[81] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” in *ICLR*, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[82] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” *neurIPS*, pp. 2–6, 2019, [Online]. Available:

http://arxiv.org/abs/1910.01108

[83] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," *Adv. Neural Inf. Process. Syst.*, vol. 32, no. NeurIPS, pp. 1–18, 2019.

[84] R. Ali, U. Farooq, U. Arshad, W. Shahzad, and M. Omer, "Computer Speech & Language Hate speech detection on Twitter using transfer learning," *Comput. Speech Lang.*, vol. 74, no. February, p. 101365, 2022, doi: 10.1016/j.csl.2022.101365.

[85] T. Wullach, A. Adler, and E. Minkov, "Character-level HyperNetworks for Hate Speech Detection," *Expert Syst. Appl.*, vol. 205, no. May, p. 117571, 2022, doi: 10.1016/j.eswa.2022.117571.

[86] G. Rajput, N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "Hate Speech Detection Using Static BERT Embeddings," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13147 LNCS, no. March, pp. 67–77, 2021, doi: 10.1007/978-3-030-93620-4_6.

[87] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, and P. Buitelaar, "Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text," *Proc. Second Work. Trolling, Aggress. Cyberbullying*, vol. 2020-Decem, no. May, pp. 32–41, 2020, [Online]. Available: https://www.aclweb.org/anthology/2020.trac-1.6

[88] M. Zhang *et al.*, "A Review of SOH Prediction of Li-Ion Batteries Based on Data-Driven Algorithms," *Energies*, vol. 16, no. 7, 2023, doi: 10.3390/en16073167.

[89] M. Zhang *et al.*, "Electrochemical Impedance Spectroscopy: A New Chapter in the Fast and Accurate Estimation of the State of Health for Lithium-Ion Batteries," *Energies*, vol. 16, no. 4, pp. 1–16, 2023, doi: 10.3390/en16041599.

[90] W. A. Arentz and B. Olstad, "Classifying offensive sites based on image content," *Comput. Vis. Image Underst.*, vol. 94, no. 1–3, pp. 295–310, 2004, doi: 10.1016/j.cviu.2003.10.007.

[91] N. B. P. Kakumanu, S. Makrogiannis, "A survey of skin-color modeling and detection methods," *Pattern Recognit.*, vol. 40, no. 3, pp. 1106–1122, 2007.

[92] C. Tian, X. Zhang, W. Wei, and X. Gao, "Color pornographic image detection based on color-saliency preserved mixture deformable part model," *Multimed. Tools Appl.*, vol. 77, no. 6, pp. 6629–6645, 2018, doi: 10.1007/s11042-017-4576-2.

[93] S. Gandhi *et al.*, "Scalable detection of offensive and non-compliant content / logo in product images," *Proc. - 2020 IEEE Winter Conf. Appl. Comput. Vision, WACV 2020*, pp. 2236–2245, 2020, doi: 10.1109/WACV45572.2020.9093454.

[94] W. Hu, O. Wu, Z. Chen, Z. Fu, and S. Maybank, "Recognition of pornographic Web pages by classifying texts and images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1019–1034, 2007, doi: 10.1109/TPAMI.2007.1133.

[95] A. Bajaj and D. K. Vishwakarma, *A state-of-the-art review on adversarial machine learning in image classification*, no. 0123456789. Springer US, 2023. doi: 10.1007/s11042-023-15883-z.

[96] D. Kiela *et al.*, "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes," *arXiv:2005.04790v3 [cs.AI]*, pp. 1–17, 2021, [Online]. Available: http://arxiv.org/abs/2005.04790

[97] J. Lu, D. Batra, D. Parikh, and Stefan Lee1, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada*, 2019.

[98] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual Captions: A Cleaned,

Hypernymed, Image Alt-text Dataset For Automatic Image Captioning," *ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 1, pp. 2556–2565, 2018.

[99] R. K. W. Lee, R. Cao, Z. Fan, J. Jiang, and W. H. Chong, *Disentangling Hate in Online Memes*, vol. 1, no. 1. Association for Computing Machinery, 2021. doi: 10.1145/3474085.3475625.

[100] A. Bhat, V. Varshney, V. Bajlotra, and V. Gupta, "Detection of hatefulness in Memes using unimodal and multimodal techniques," in *Proceedings of the Sixth International Conference on Intelligent Computing and Control Systems (ICICCS 2022)*, IEEE, 2022, pp. 65–73.

[101] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Exploring hate speech detection in multimodal publications," *Proc. - 2020 IEEE Winter Conf. Appl. Comput. Vision, WACV 2020*, pp. 1459–1467, 2020, doi: 10.1109/WACV45572.2020.9093414.

[102] D. S. Chauhan, G. V. Singh, A. Arora, A. Ekbal, and P. Bhattacharyya, "An emoji-aware multitask framework for multimodal sarcasm detection," *Knowledge-Based Syst.*, vol. 257, p. 109924, 2022, doi: 10.1016/j.knosys.2022.109924.

[103] I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, "ETHOS: an online hate speech detection dataset," *Complex Intell. Syst.*, 2022, doi: 10.1007/s40747-021-00608-2.

[104] M. ElSherief *et al.*, "Latent Hatred: A Benchmark for Understanding Implicit Hate Speech," pp. 345–363, 2021, doi: 10.18653/v1/2021.emnlp-main.29.

[105] A. Baruah, F. A. Barbhuiya, and K. Dey, "ABARUAH at SemEval-2019 task 5: Bi-directional LSTM for hate speech detection," *NAACL HLT 2019 - Int. Work. Semant. Eval. SemEval 2019, Proc. 13th Work.*, no. 2017, pp. 371–376, 2019, doi: 10.18653/v1/s19-2065.

[106] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 5987–5995, 2017, doi: 10.1109/CVPR.2017.634.

[107] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 510–519, 2019, doi: 10.1109/CVPR.2019.00060.

[108] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," *Proc. - 2021 IEEE Winter Conf. Appl. Comput. Vision, WACV 2021*, pp. 3138–3147, 2021, doi: 10.1109/WACV48630.2021.00318.

[109] P. T. Jiang, C. Bin Zhang, Q. Hou, M. M. Cheng, and Y. Wei, "LayerCAM: Exploring hierarchical class activation maps for localization," *IEEE Trans. Image Process.*, vol. 30, no. June, pp. 5875–5888, 2021, doi: 10.1109/TIP.2021.3089943.

[110] D. Kiela *et al.*, "The hateful memes challenge: Detecting hate speech in multimodal memes," *Adv. Neural Inf. Process. Syst.*, vol. 2020-Decem, no. NeurIPS, pp. 1–14, 2020.

[111] C. Yang, F. Zhu, G. Liu, J. Han, and S. Hu, "Multimodal Hate Speech Detection via Cross-Domain Knowledge Transfer," *MM 2022 - Proc. 30th ACM Int. Conf. Multimed.*, pp. 4505–4514, 2022, doi: 10.1145/3503161.3548255.

[112] K. L. Tsun-hin Cheung, "Crossmodal bipolar attention for multimodal classification on social media," *Neurocomputing*, vol. 514, pp. 1–14, 2022.

[113] D. Kumar, N. Kumar, and S. Mishra, "QUARC: Quaternion multi-modal fusion architecture for hate speech classification," *Proc. - 2021 IEEE Int. Conf. Big Data Smart Comput. BigComp 2021*, vol. 0, pp. 346–349, 2021, doi: 10.1109/BigComp51126.2021.00075.

[114] G. Sahu, R. Cohen, and O. Vechtomova, "Towards A Multi-agent System for Online Hate Speech Detection," *Proc. 20th Int. Conf. Auton. Agents Multiagent Syst.*, 2021.

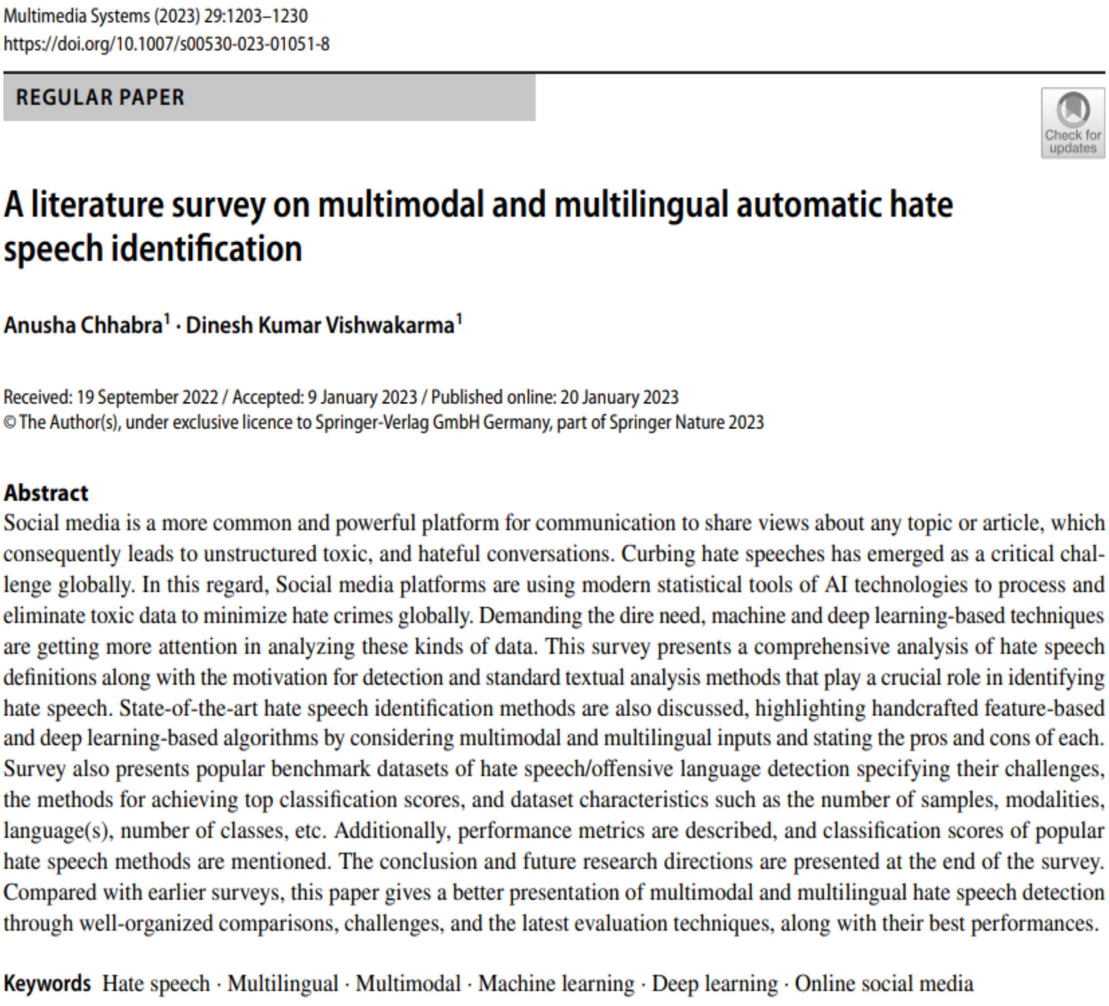
[115] N. Prasad, S. Saha, and P. Bhattacharyya, “A Multimodal Classification of Noisy Hate Speech using Character Level Embedding and Attention,” *Proc. Int. Jt. Conf. Neural Networks*, vol. 2021-July, pp. 1–8, 2021, doi: 10.1109/IJCNN52387.2021.9533371.

[116] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.

PROOF OF PUBLICATIONS

SCIE Journal Paper 1:

Anusha Chhabra, Dinesh Kumar Vishwakarma. "A Literature Survey on Multimodal and Multilingual Automatic Hate Speech Identification." Published in **Multimedia Systems**, 2023, (Pub: Springer): **DOI:** https://doi.org/10.1007/s00530-023-01051-8.



Multimedia Systems (2023) 29:1203–1230
https://doi.org/10.1007/s00530-023-01051-8

REGULAR PAPER

A literature survey on multimodal and multilingual automatic hate speech identification

Anusha Chhabra[1] · Dinesh Kumar Vishwakarma[1]

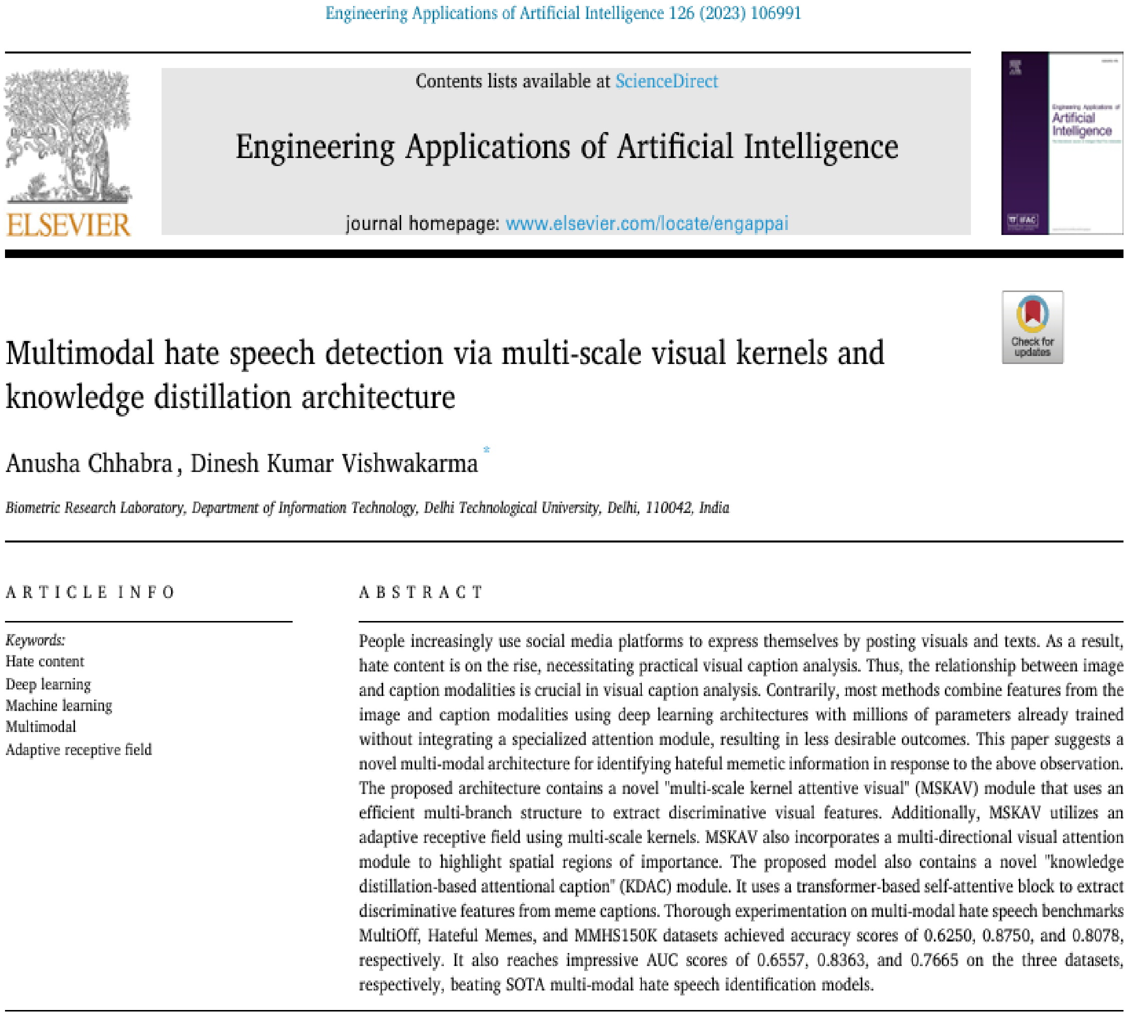
Received: 19 September 2022 / Accepted: 9 January 2023 / Published online: 20 January 2023

Abstract
Social media is a more common and powerful platform for communication to share views about any topic or article, which consequently leads to unstructured toxic, and hateful conversations. Curbing hate speeches has emerged as a critical challenge globally. In this regard, Social media platforms are using modern statistical tools of AI technologies to process and eliminate toxic data to minimize hate crimes globally. Demanding the dire need, machine and deep learning-based techniques are getting more attention in analyzing these kinds of data. This survey presents a comprehensive analysis of hate speech definitions along with the motivation for detection and standard textual analysis methods that play a crucial role in identifying hate speech. State-of-the-art hate speech identification methods are also discussed, highlighting handcrafted feature-based and deep learning-based algorithms by considering multimodal and multilingual inputs and stating the pros and cons of each. Survey also presents popular benchmark datasets of hate speech/offensive language detection specifying their challenges, the methods for achieving top classification scores, and dataset characteristics such as the number of samples, modalities, language(s), number of classes, etc. Additionally, performance metrics are described, and classification scores of popular hate speech methods are mentioned. The conclusion and future research directions are presented at the end of the survey. Compared with earlier surveys, this paper gives a better presentation of multimodal and multilingual hate speech detection through well-organized comparisons, challenges, and the latest evaluation techniques, along with their best performances.

Keywords Hate speech · Multilingual · Multimodal · Machine learning · Deep learning · Online social media

SCIE Journal Paper 2:

Anusha Chhabra, Dinesh Kumar Vishwakarma. "Multimodal Hate Speech Detection via Multi-Scale Visual Kernels and Knowledge Distillation Architecture." Published in **Engineering Applications of Artificial Intelligence** (126), 2023, (Pub: Elsevier), **DOI:** https://doi.org/10.1016/j.engappai.2023.106991.



Engineering Applications of Artificial Intelligence 126 (2023) 106991

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Multimodal hate speech detection via multi-scale visual kernels and knowledge distillation architecture

Anusha Chhabra, Dinesh Kumar Vishwakarma *

Biometric Research Laboratory, Department of Information Technology, Delhi Technological University, Delhi, 110042, India

ARTICLE INFO

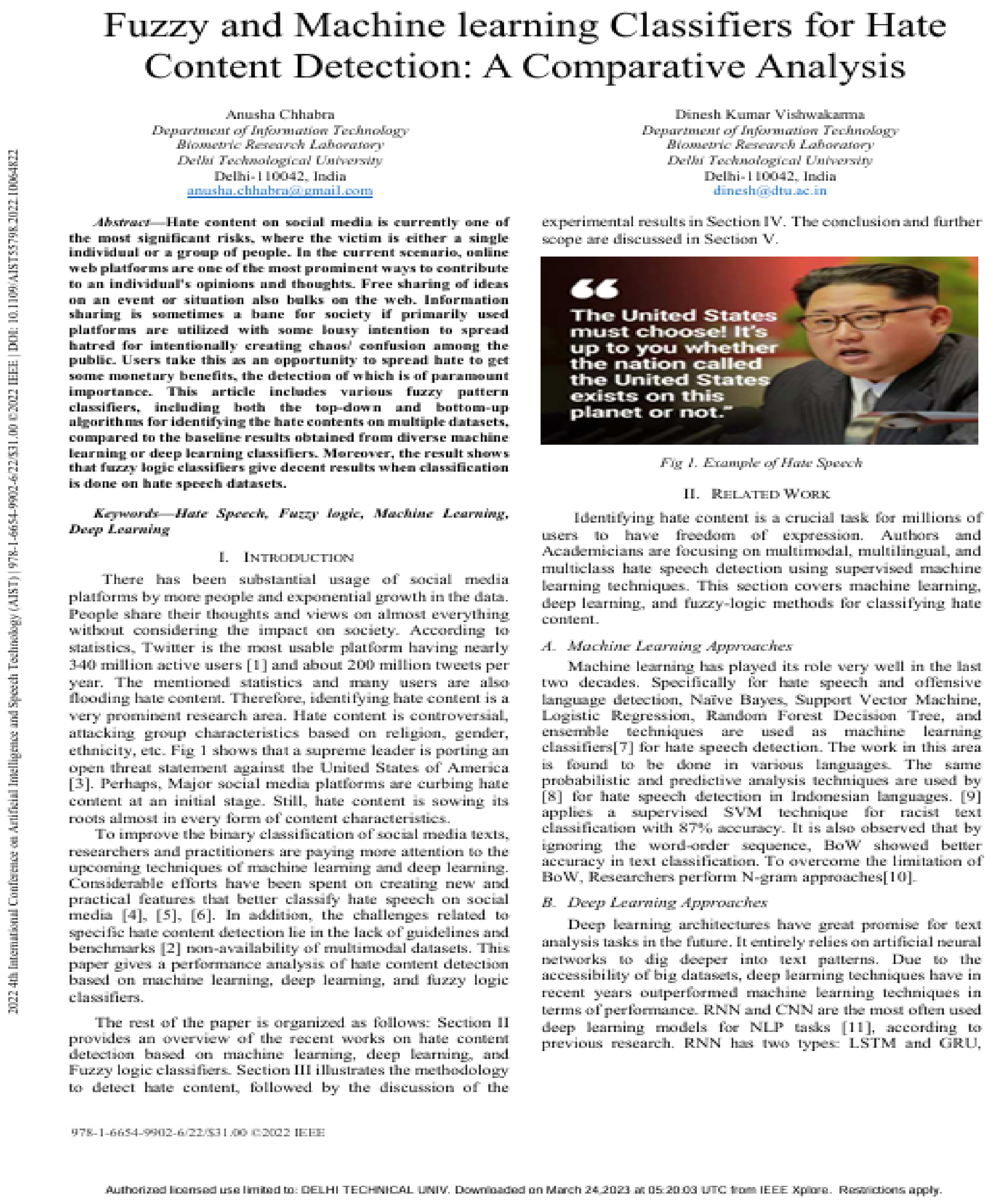
Keywords:
Hate content
Deep learning
Machine learning
Multimodal
Adaptive receptive field

ABSTRACT

People increasingly use social media platforms to express themselves by posting visuals and texts. As a result, hate content is on the rise, necessitating practical visual caption analysis. Thus, the relationship between image and caption modalities is crucial in visual caption analysis. Contrarily, most methods combine features from the image and caption modalities using deep learning architectures with millions of parameters already trained without integrating a specialized attention module, resulting in less desirable outcomes. This paper suggests a novel multi-modal architecture for identifying hateful memetic information in response to the above observation. The proposed architecture contains a novel "multi-scale kernel attentive visual" (MSKAV) module that uses an efficient multi-branch structure to extract discriminative visual features. Additionally, MSKAV utilizes an adaptive receptive field using multi-scale kernels. MSKAV also incorporates a multi-directional visual attention module to highlight spatial regions of importance. The proposed model also contains a novel "knowledge distillation-based attentional caption" (KDAC) module. It uses a transformer-based self-attentive block to extract discriminative features from meme captions. Thorough experimentation on multi-modal hate speech benchmarks MultiOff, Hateful Memes, and MMHS150K datasets achieved accuracy scores of 0.6250, 0.8750, and 0.8078, respectively. It also reaches impressive AUC scores of 0.6557, 0.8363, and 0.7665 on the three datasets, respectively, beating SOTA multi-modal hate speech identification models.

Conference Paper 1:

Anusha Chhabra, Dinesh Kumar Vishwakarma. “Fuzzy and Machine Learning Classifiers for Hate Content Detection: A Comparative Analysis.” **IEEE Conference: 4th International Conference on Artificial Intelligence and Speech Technology (AIST)**, *IGDTUW, Delhi. (2022) [**Presented]**,* **DOI: 10.1109/AIST55798.2022.10064822**



Fuzzy and Machine learning Classifiers for Hate Content Detection: A Comparative Analysis

Anusha Chhabra
Department of Information Technology
Biometric Research Laboratory
Delhi Technological University
Delhi-110042, India
anusha.chhabra@gmail.com

Dinesh Kumar Vishwakarma
Department of Information Technology
Biometric Research Laboratory
Delhi Technological University
Delhi-110042, India
dinesh@dtu.ac.in

2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST) | 978-1-6654-9902-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/AIST55798.2022.10064822

***Abstract*—Hate content on social media is currently one of the most significant risks, where the victim is either a single individual or a group of people. In the current scenario, online web platforms are one of the most prominent ways to contribute to an individual's opinions and thoughts. Free sharing of ideas on an event or situation also bulks on the web. Information sharing is sometimes a bane for society if primarily used platforms are utilized with some lousy intention to spread hatred for intentionally creating chaos/ confusion among the public. Users take this as an opportunity to spread hate to get some monetary benefits, the detection of which is of paramount importance. This article includes various fuzzy pattern classifiers, including both the top-down and bottom-up algorithms for identifying the hate contents on multiple datasets, compared to the baseline results obtained from diverse machine learning or deep learning classifiers. Moreover, the result shows that fuzzy logic classifiers give decent results when classification is done on hate speech datasets.**

Keywords*—Hate Speech, Fuzzy logic, Machine Learning, Deep Learning

I. Introduction

There has been substantial usage of social media platforms by more people and exponential growth in the data. People share their thoughts and views on almost everything without considering the impact on society. According to statistics, Twitter is the most usable platform having nearly 340 million active users [1] and about 200 million tweets per year. The mentioned statistics and many users are also flooding hate content. Therefore, identifying hate content is a very prominent research area. Hate content is controversial, attacking group characteristics based on religion, gender, ethnicity, etc. Fig 1 shows that a supreme leader is porting an open threat statement against the United States of America [3]. Perhaps, Major social media platforms are curbing hate content at an initial stage. Still, hate content is sowing its roots almost in every form of content characteristics.

To improve the binary classification of social media texts, researchers and practitioners are paying more attention to the upcoming techniques of machine learning and deep learning. Considerable efforts have been spent on creating new and practical features that better classify hate speech on social media [4], [5], [6]. In addition, the challenges related to specific hate content detection lie in the lack of guidelines and benchmarks [2] non-availability of multimodal datasets. This paper gives a performance analysis of hate content detection based on machine learning, deep learning, and fuzzy logic classifiers.

The rest of the paper is organized as follows: Section II provides an overview of the recent works on hate content detection based on machine learning, deep learning, and Fuzzy logic classifiers. Section III illustrates the methodology to detect hate content, followed by the discussion of the experimental results in Section IV. The conclusion and further scope are discussed in Section V.

“The United States must choose! It's up to you whether the nation called the United States exists on this planet or not.”

Fig 1. Example of Hate Speech

II. Related Work

Identifying hate content is a crucial task for millions of users to have freedom of expression. Authors and Academicians are focusing on multimodal, multilingual, and multiclass hate speech detection using supervised machine learning techniques. This section covers machine learning, deep learning, and fuzzy-logic methods for classifying hate content.

A. Machine Learning Approaches

Machine learning has played its role very well in the last two decades. Specifically for hate speech and offensive language detection, Naïve Bayes, Support Vector Machine, Logistic Regression, Random Forest Decision Tree, and ensemble techniques are used as machine learning classifiers[7] for hate speech detection. The work in this area is found to be done in various languages. The same probabilistic and predictive analysis techniques are used by [8] for hate speech detection in Indonesian languages. [9] applies a supervised SVM technique for racist text classification with 87% accuracy. It is also observed that by ignoring the word-order sequence, BoW showed better accuracy in text classification. To overcome the limitation of BoW, Researchers perform N-gram approaches[10].

B. Deep Learning Approaches

Deep learning architectures have great promise for text analysis tasks in the future. It entirely relies on artificial neural networks to dig deeper into text patterns. Due to the accessibility of big datasets, deep learning techniques have in recent years outperformed machine learning techniques in terms of performance. RNN and CNN are the most often used deep learning models for NLP tasks [11], according to previous research. RNN has two types: LSTM and GRU,

978-1-6654-9902-6/22/$31.00 ©2022 IEEE

Conference Paper 2:

Anusha Chhabra, Dinesh Kumar Vishwakarma. “A Truncated SVD Framework for Online Hate Speech Detection on the ETHOS Dataset.” **IEEE Conference: 4th International Conference on Innovative Trends in Information Technology (ICIIIT)**, *IIIT Kottayam, Kerala. (2023) [**Presented**]*, **DOI: 10.1109/ICITIIT57246.2023.10068574**

A Truncated SVD Framework for Online Hate Speech Detection on the ETHOS Dataset

Anusha Chhabra
Department of Information Technology
Biometric Research Laboratory
Delhi Technological University
Delhi-110042, India
anusha.chhabra@gmail.com

Dinesh Kumar Vishwakarma
Department of Information Technology
Biometric Research Laboratory
Delhi Technological University
Delhi-110042, India
dinesh@dtu.ac.in

2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT) | 978-1-6654-9414-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICITIIT57246.2023.10068574

***Abstract*—Hate content on social media is currently one of the most significant risks, where the victim is either a single individual or a group of people. In the current scenario, online web platforms are one of the most prominent ways to contribute to an individual's opinions and thoughts. Free sharing of ideas on an event or situation also bulks on the web. Information sharing is sometimes a bane for society if primarily used platforms are utilized with some lousy intention to spread hatred for intentionally creating chaos/ confusion among the public. Users take this as an opportunity to spread hate to get some monetary benefits, the detection of which is of paramount importance. This article utilizes the concept of truncated singular value decomposition (SVD) for detecting hate content on the ETHOS (Binary-Label) dataset. Compared with the baseline results, our framework has performed better in various machine learning algorithms like SVM, Logistic Regression, XGBoost, and Random Forest.**

Keywords—Hate Speech, Machine Learning, SVD, Binary-label Classification, TF-IDF

I. Introduction

There has been substantial usage of social media platforms by more people and exponential growth in the data. People share their thoughts and views on almost everything without considering the impact on society. According to statistics, Twitter is the most usable platform having nearly 340 million active users [1] and about 200 million tweets per year. The mentioned statistics and many users are also flooding hate content. Therefore, identifying hate content is a very prominent research area. Hate content can be defined as controversial, attacking group characteristics based on religion, gender, ethnicity, etc. Fig 1 shows that a leader is porting a divisive statement targeting those who raise their voices against CAA, NRC, and NPR [2]. Perhaps, Major social media platforms are curbing hate content at an initial stage. Still, hate content is sowing its roots almost in every form of content characteristics.

To improve the binary classification of social media texts, researchers and practitioners are paying more attention to the upcoming techniques of machine learning and deep learning. Considerable efforts have been spent on creating new and practical features that better classify hate speech on social media [3], [4], [5]. In addition, the challenges related to specific hate content detection lie in the need for more guidelines, benchmarks [6], and the non-availability of multimodal datasets. This paper presents a framework for identifying hate content on the ETHOS dataset.

The Major contributions of this manuscript are:

- Training the models on one dataset and cross-validation is done on another dataset which is approximately 24 times greater.
- To show the vulnerability of a small dataset with another related large dataset.

The rest of the paper is organized as follows: Section II provides an overview of the recent works on hate content detection using unsupervised machine learning approaches.

Section III illustrates the framework to detect hate content, followed by the discussion of the experimental results in Section IV. The conclusion and further scope are discussed in Section V.

"If you raise slogans against
PM Modi or CM Yogi Adityanath,
I will bury you alive."
- Raghuraj Singh, BJP Leader

Fig 1. Example of Hate Speech

II. Related Work

Identifying hate content is crucial for millions of users to have freedom of expression. Authors and Academicians are focusing on multimodal, multilingual, and multiclass hate speech detection using supervised, unsupervised and semi-supervised machine learning techniques.

Machine learning has played its role very well in the last two decades. Specifically for hate speech and offensive language detection, Naïve Bayes, Support Vector Machine, Logistic Regression, Random Forest Decision Tree, and ensemble techniques are used as machine learning classifiers[7] for hate speech detection. The work in this area is found to be done in various languages. The same probabilistic and predictive analysis techniques are used by [8] for hate speech detection in Indonesian languages. [9] applies a supervised SVM technique for racist text classification. It is also observed that by ignoring the word-order sequence, BoW showed better accuracy in text classification. To overcome the limitation of BoW, Researchers perform N-gram approaches[10]. Manual labeling of large data is a time-consuming task that leads to the requirement of an unsupervised method. It takes advantage of detecting hate speech in a huge stream of data. Authors in [11] used Kohonen maps for the detection of cyberbullying, claiming an accuracy of 72%. PCA is also another class to

978-1-6654-9414-4/23/$31.00 ©2023 IEEE

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis: Design and Development of Framework for Detection of Hate Content

Total Pages:

Name of the Scholar: Anusha Chhabra

Supervisor: Prof. Dinesh Kumar Vishwakarma

Department: Information Technology

This is to report that the above thesis was scanned for similarity detection. Process and outcome are given below:

Software used: Turnitin   Similarity Index: 9%   Word Count: 23,679 Words

Date: 09/12/2024

Candidate's Signature   **Signature of Supervisor**

Author Biography

Anusha Chhabra received Bachelor in Engineering degree in 2007 from Maharishi Dayanand University, Rohtak, Haryana and Master of Technology degree in 2011 from Guru Gobind Singh Indraprastha University, Delhi, India. She is currently pursuing Ph.D. degree from the Delhi Technological University, India.The topic of her doctoral dissertation is Design and Development of a Framework for Detection of Hate Content. Her research interests include machine learning approaches and deep-learning based computer vision problems etc.