ROLE OF NLP IN ERA OF 4IR

Understanding Social Media Beyond Texts

Eftekhar Hossain
Lecturer, CUET

Social Media Data

- The number of users on social media is huge
- One in-three people in the world use social media
- Important data source in both industry and academia

Social Media Implications

- Diverse applications in
  - Sales, Marketing
  - Disaster management,
  - Crime surveillance and Event detection.

Challenges

- A great number of users who update massive information every second
- Information is not only included in the short textual content
  - embedded in the images and videos

Objective

- ➢ Utilize Multimodal Data or Multiple modalities
  - Image
  - Text
  - Acoustic

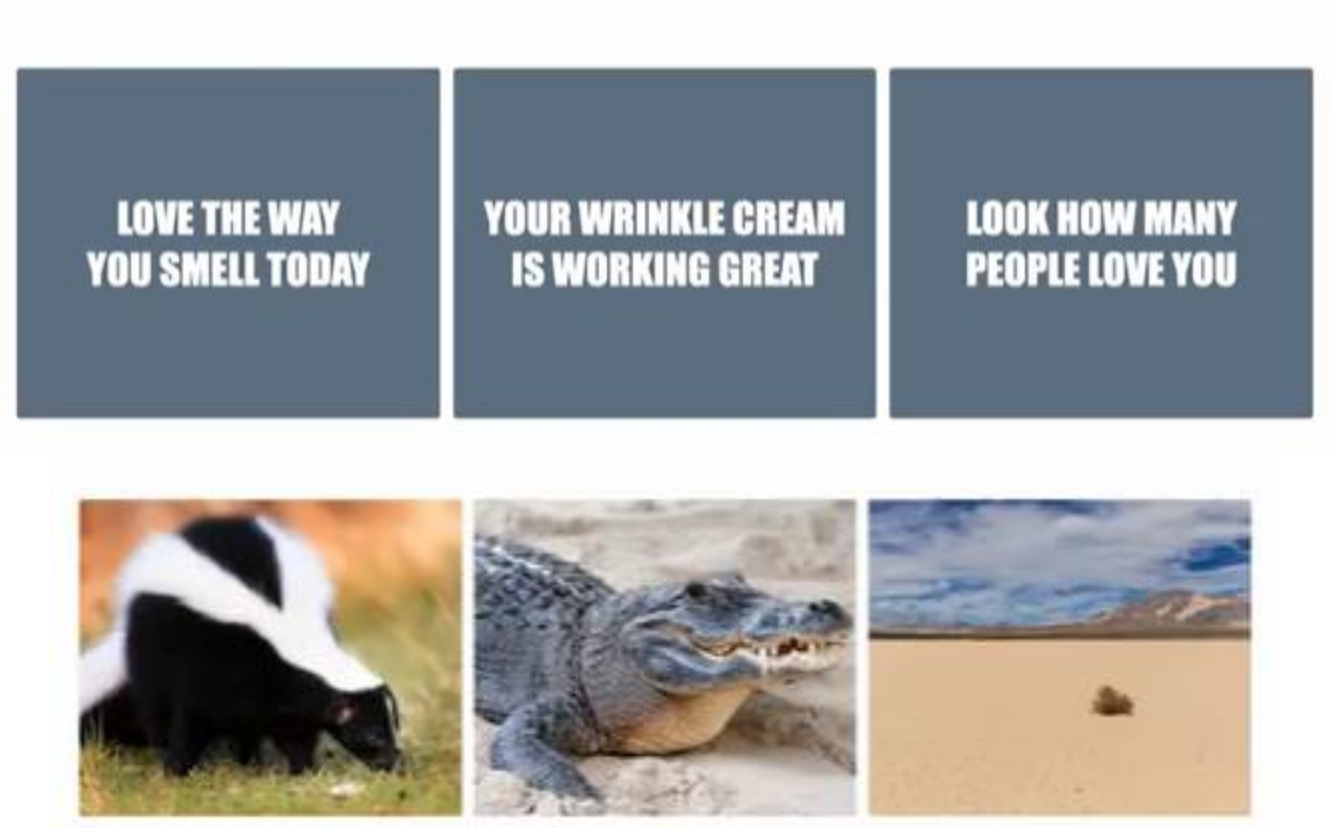
Harmful Content Detection

- Epidemic of online offensive and abusive behaviour
- Mode of communication transforming day by day
- Easier to deceive the surveillance Engine

Harmful Content Detection

- **Memes** can propagate information humorously or sarcastically
- Facebook Hateful Memes Challenge (2020)

Harmful Content Detection



Collected from Facebook AI research

Harmful Content Detection

Combine them meaning become harmful

Collected from Facebook AI research

Harmful Content Detection

Change the images meaning become harmless

Collected from Facebook AI research

Goal

- Effective tool for detecting Harmful content

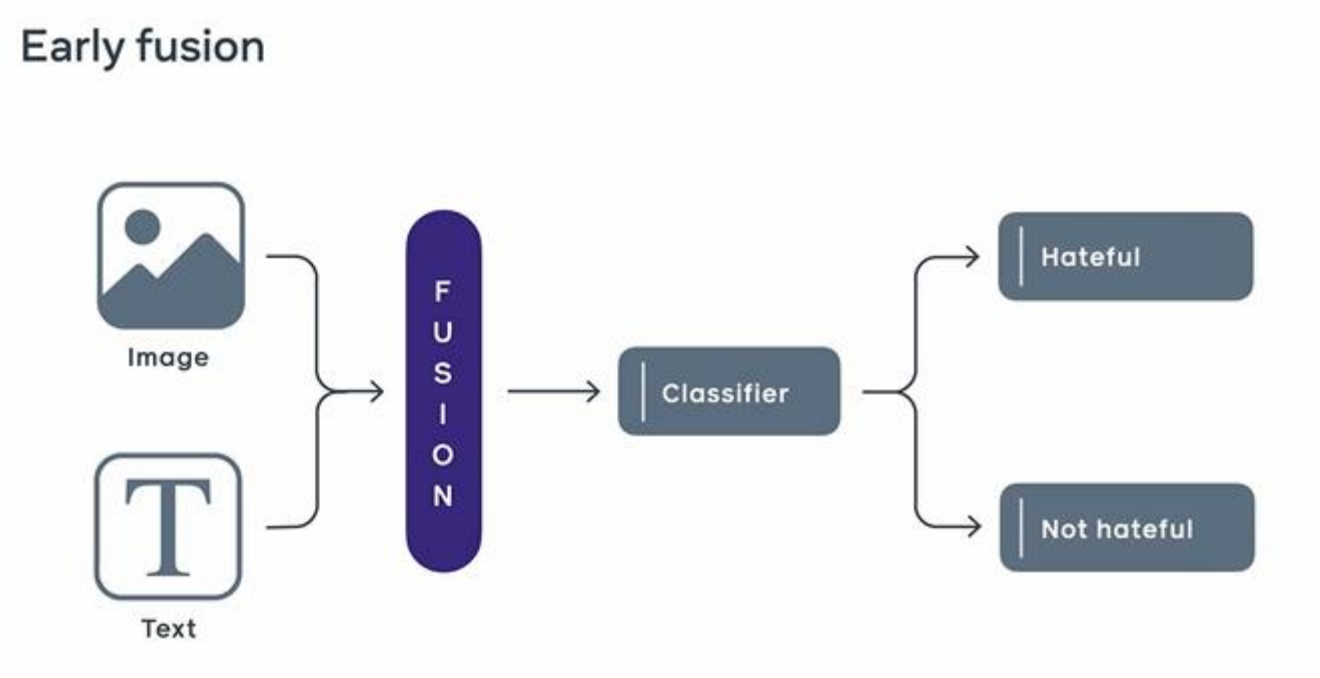
  When viewing a meme,

  - → we don’t think about the words and photo independently of each other;
  - → we understand the combined meaning together.

Challenging for Machines

- ➔ Can’t just analyze the text and the image separately.
- ➔ Must combine these different modalities and
- ➔ Understand how the meaning changes when they are presented together.

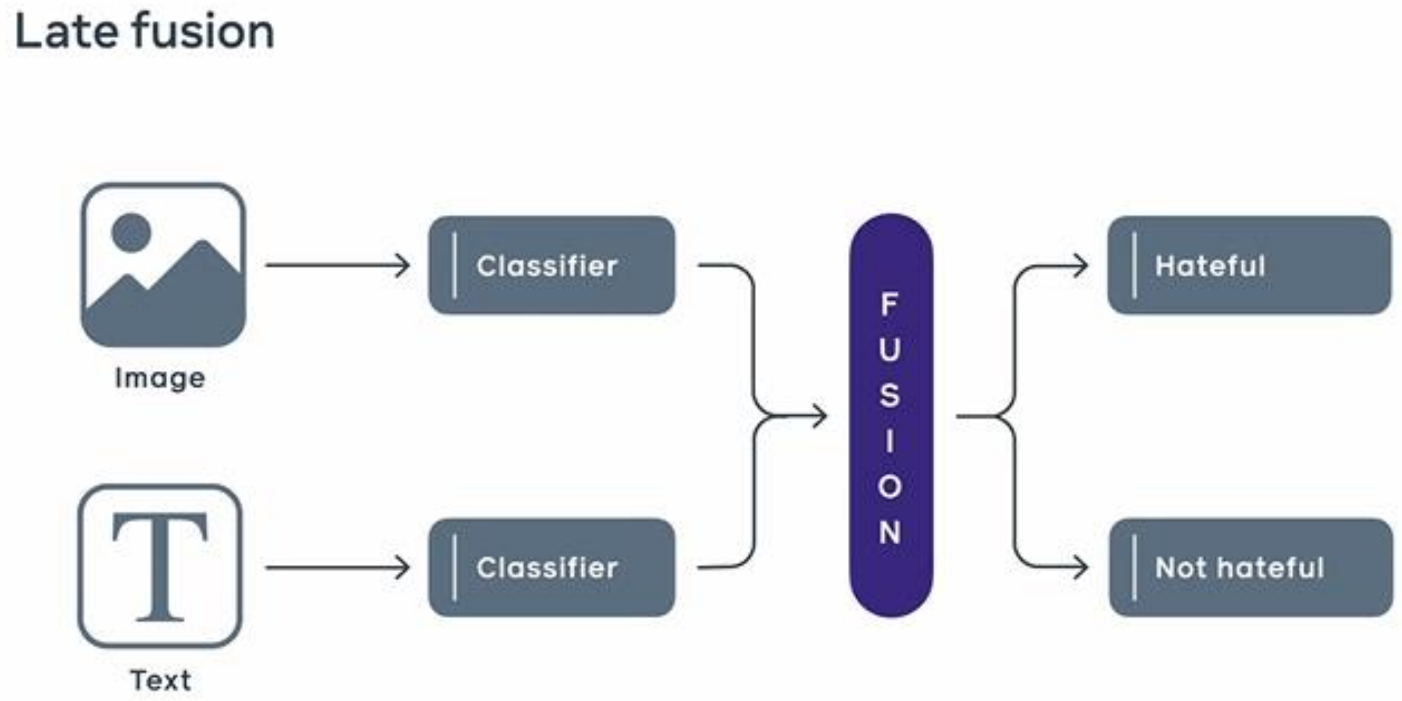
Multimodal AI



This approach enables the system to analyze the different modalities together

Collected from Facebook AI research

Multimodal AI



easier to build but less effective at understanding complex multimodal content

Collected from Facebook AI research

Multimodal AI (Paper-1)

Identification of Multilingual Offense and Troll from Social Media Memes using Weighted Ensemble of Multimodal Features (Hossain et. al, 2022) *[Journal Paper]*

Authors: ***Eftekhar Hossain**, Omar Sharif, Mohammed Moshiul Hoque, M. Ali Akber Dewan, Nazmul Siddique, Md. Azad Hossain*

Multimodal AI (Paper-1)

Identification of Multilingual Offense and Troll from Social Media Memes using Weighted Ensemble of Multimodal Features [Journal Paper]

(a) Offensive (b) Troll (c) Troll

Drawbacks of Previous Works

- ➔ Past studies considered only a single modality (image or text)
- ➔ Not explored the joint modelling of multimodal features
- ➔ As well as their counteractive unimodal features (i.e., image, text) to classify undesired memes
- ➔ No unified architecture for multilingual memes

Research Question

➔ How to develop a framework leveraging features from visual and textual modality to identify offense and troll from memes ?

Contributions

- Propose a model that exploits visual, textual and multimodal features of the multilingual memes.
- Investigate the multimodal decision fusion, and feature fusion approaches
- Employed an ensemble technique that automatically assigns appropriate weight to the participating modules

Description of the Task

- ➔ Develop a framework (***F***) to identify offense and troll from memes
- ➔ ***F*** analyzes a set of memes and $M = m_1, m_2, \ldots\ldots m_n$ categorize them as offense/troll *(c = 1)* or not *(c = 0)*
- ➔ Each meme consists of visual (***v***) and textual (***t***) information and the **F** utilize these information

Dataset

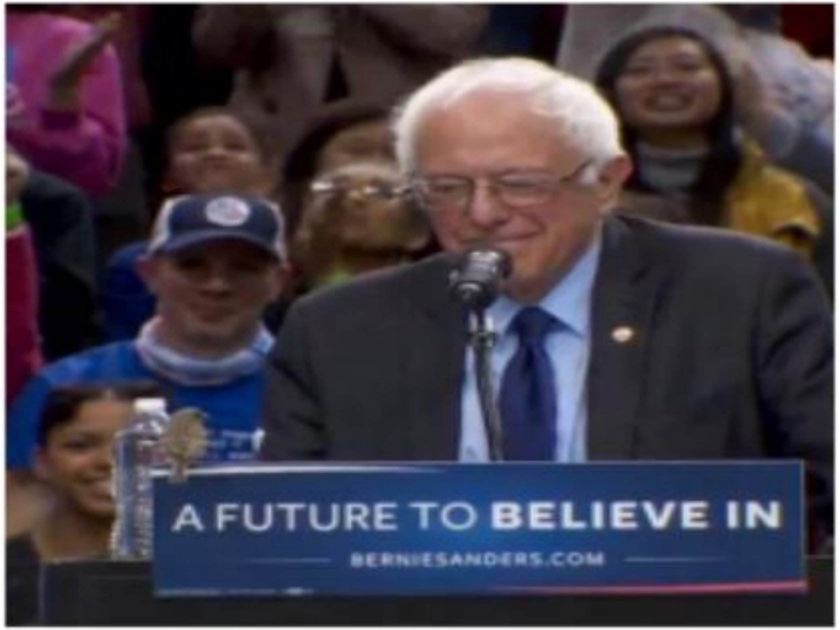
D1: MultiOFF (offense: 303, not-offense: 440)

D2: TamilMemes (Troll: 1677, Not-troll: 1290)

Offense: Demean social identity, harass targeted individuals, community or a minority group

Troll : Provoke, abuse or insult individuals, group, or a race

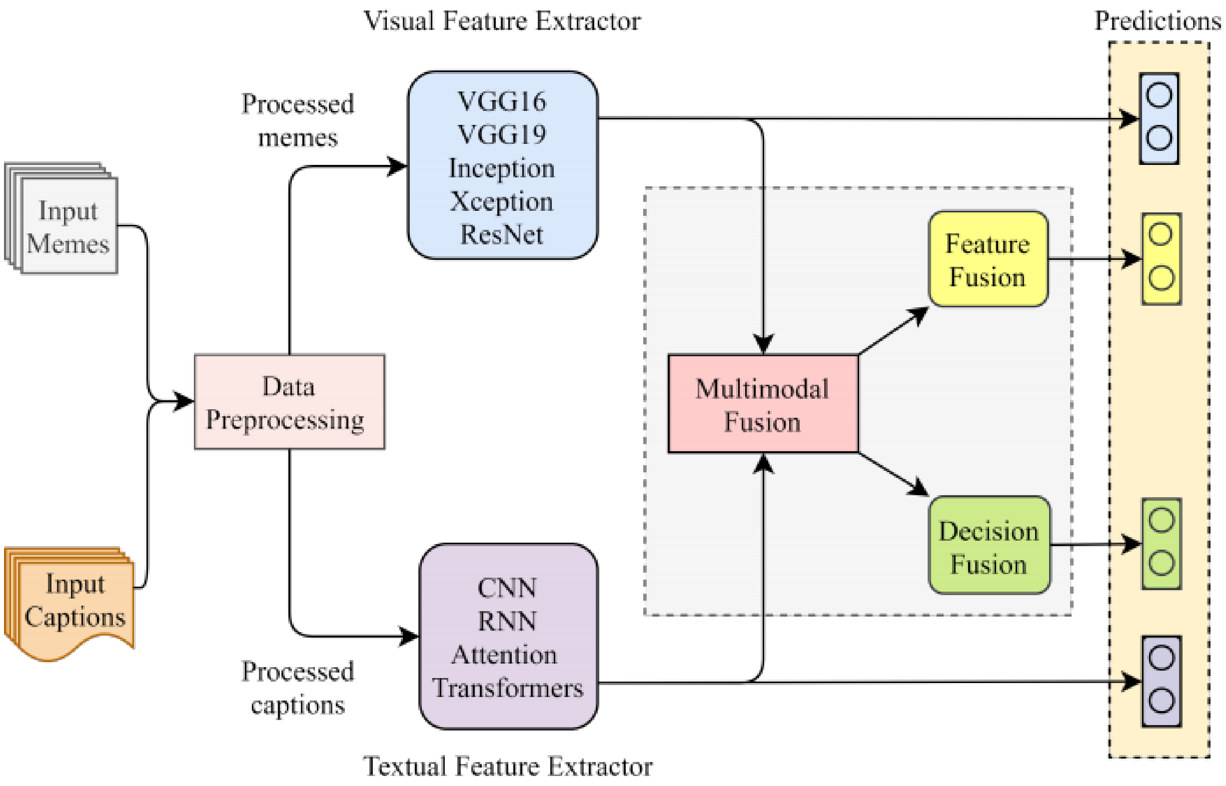
Dataset



Offensive

Not-offensive

Methodology



Abstract view of the multimodal offense and troll detection system

Methodology

Algorithm 1: Process of selecting best 3 visual and textual models

```
Input: Weighted f1-scores
Output: Best visual and textual models
Vf ← [vf1, vf2, ..., vfN] (Weighted f1 scores of visual models);
Tf ← [tf1, tf2, ..., tfM] (Weighted f1 scores of textual models);
Vm ← [];
Tm ← [];
sort(Vf, Vf + N);
sort(Tf, Tf + M);
//choosing best 3 visual and textual models
for i∈(1, 3) do
    Vm.append(Vf[i]);
    Tm.append(Tf[i]);
    i = i + 1;
end
```

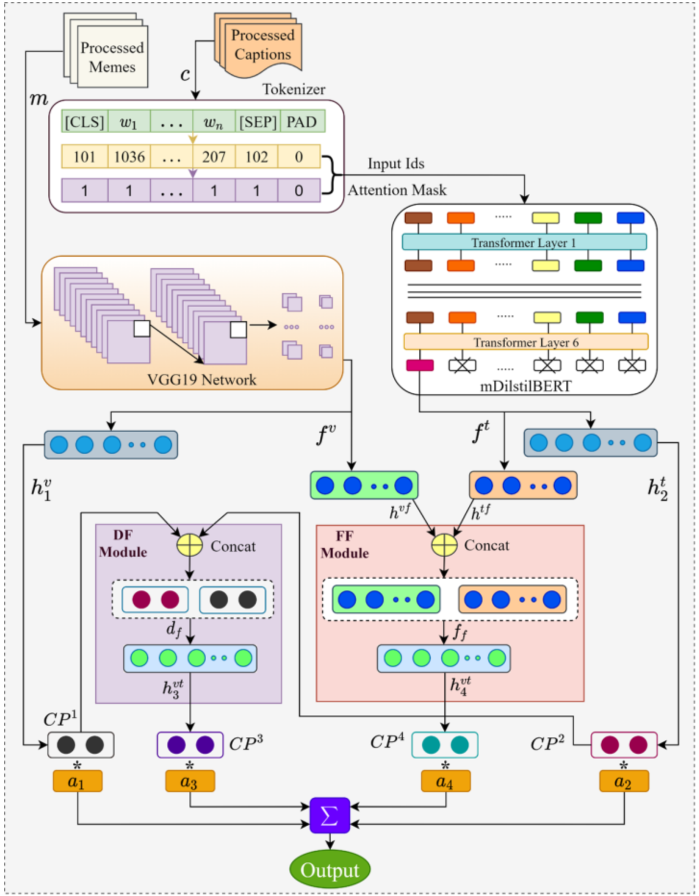
Methodology

- VGG16, VGG19, and ResNet50 are the best visual models

- m-BERT, m-DistilBERT, and XLM-R are the best textual models.

- Multimodal Models
  - we obtain a total of ((3x3)x 2) = 18 multimodal models where each fusion approach (i.e., decision, feature) contributed 9 different models.

Proposed Ensemble Technique



This approach exploits the strength of multiple models and tries to increase the overall system predictive accuracy

Algorithm 2: Process of the proposed weighted ensemble technique

Input: Class probabilities and Accuracy
Output: Predictions of the W-ensemble

$cp \leftarrow []$ (class probabilities);
$a \leftarrow []$ (accuracy);

$sum = []$ (weighted sum);
for $i \epsilon (1, m)$ **do**
  for $j \epsilon (1, l)$ **do**
    $sum[i] = sum[i] + (cp_i^j[] * a_j)$;
    $j = j + 1$;
  end
  $i = i + 1$;
end

$n\_sum = 0$;
for $j \epsilon (1, l)$ **do**
  $n\_sum = n\_sum + a_j$;
  $j = j + 1$;
end

$P = (sum/n\_sum)$ //normalized probabilities;
$E_p = \arg\max(P)$ // set of predictions;

Experiments and Results

Approach	Models	Dataset-1 (D1)				Dataset-2 (D2)			
		A	P	R	f_1-score	A	P	R	f_1-score
Visual	VGG16	0.577	0.581	0.577	0.579	0.596	0.572	0.596	0.502
	VGG19	0.610	0.621	0.610	**0.614**	0.575	0.536	0.575	**0.516**
	ResNet50	0.624	0.607	0.624	0.606	0.592	0.560	0.592	0.503
	InceptionV3	0.604	0.562	0.604	0.532	0.509	0.456	0.509	0.464
	Xception	0.503	0.493	0.503	0.497	0.572	0.506	0.572	0.478
Textual	CNN	0.510	0.502	0.510	0.506	0.559	0.523	0.559	0.518
	BiLSTM	0.530	0.487	0.530	0.496	0.595	0.568	0.595	0.530
	BiLSTM + CNN	0.590	0.556	0.590	0.550	0.595	0.569	0.595	0.536
	BiLSTM + Attention	0.597	0.568	0.597	0.564	0.548	0.509	0.548	0.507
	m-BERT	0.638	0.625	0.638	0.626	0.608	0.591	0.608	0.561
	m-DistilBERT	0.671	0.662	0.671	**0.654**	0.601	0.583	0.601	**0.573**
	XLM-R	0.591	0.573	0.591	0.576	0.601	0.578	0.601	0.556

Table 1: Performance comparison of visual and textual models on test set

Experiments and Results

Approach	Models		Dataset-1 (D1)				Dataset-2 (D2)			
			A	P	R	f_1-score	A	P	R	f_1-score
Decision Fusion	m-BERT +	VGG16	0.483	0.488	0.483	0.485	0.583	0.539	0.583	0.499
		VGG19	0.544	0.541	0.544	0.542	0.589	0.555	0.589	0.513
		ResNet50	0.577	0.558	0.577	0.562	0.513	0.532	0.513	0.517
	m-DBERT +	VGG16	0.537	0.523	0.537	0.528	0.601	0.579	0.601	0.547
		VGG19	0.591	0.628	0.591	**0.595**	0.582	0.583	0.582	**0.583**
		ResNet50	0.570	0.576	0.570	0.573	0.574	0.556	0.574	0.556
	XLM-R +	VGG16	0.497	0.523	0.497	0.503	0.592	0.579	0.592	0.579
		VGG19	0.497	0.528	0.497	0.502	0.567	0.559	0.567	0.567
		ResNet50	0.604	0.563	0.604	0.532	0.574	0.551	0.574	0.548
Feature Fusion	m-BERT +	VGG16	0.584	0.564	0.584	0.567	0.580	0.556	0.580	0.549
		VGG19	0.577	0.547	0.577	0.549	0.604	0.588	0.604	0.529
		ResNet50	0.584	0.567	0.584	0.570	0.568	0.511	0.568	0.489
	m-DBERT +	VGG16	0.604	0.592	0.604	0.595	0.589	0.563	0.589	0.546
		VGG19	0.685	0.681	0.685	**0.660**	0.591	0.568	0.591	**0.557**
		ResNet50	0.611	0.598	0.611	0.600	0.597	0.571	0.597	0.528
	XLM-R +	VGG16	0.570	0.582	0.570	0.574	0.586	0.539	0.586	0.487
		VGG19	0.530	0.524	0.527	0.502	0.568	0.518	0.568	0.499
		ResNet50	0.577	0.589	0.577	0.581	0.608	0.618	0.609	0.508

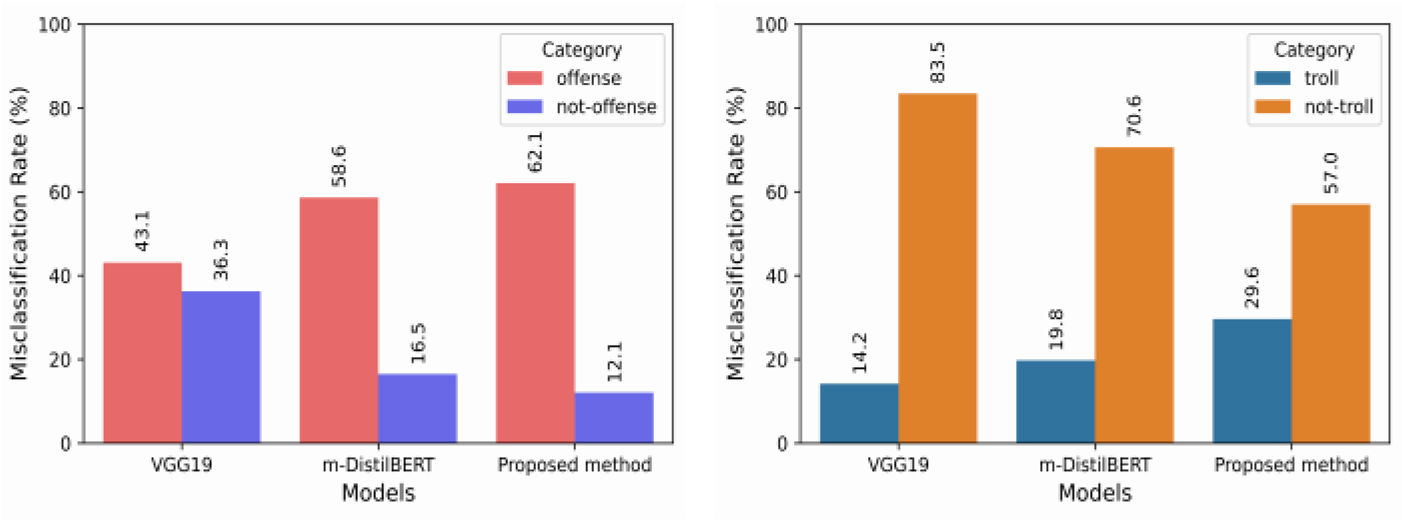
Table 2: Performance comparison of multimodal models on test set

Experiments and Results

Approach	Models	Dataset-1 (D1)				Dataset-2 (D2)			
		A	P	R	f_1-score	A	P	R	f_1-score
Average Ensemble	V + T	0.617	0.609	0.617	0.612	0.588	0.555	0.588	0.522
	V + DF	0.597	0.614	0.597	0.602	0.574	0.535	0.574	0.516
	V + FF	0.638	0.625	0.638	0.626	0.586	0.548	0.586	0.509
	T + DF	0.678	0.669	0.678	0.663	0.594	0.574	0.594	0.566
	T + FF	0.678	0.678	0.678	0.644	0.603	0.584	0.603	0.571
	DF + FF	0.678	0.673	0.678	0.651	0.594	0.573	0.594	0.563
	V + T + DF	0.570	0.565	0.570	0.567	0.585	0.556	0.585	0.540
	V + T + FF	0.678	0.669	0.678	0.665	0.592	0.566	0.592	0.546
	V + DF + FF	0.604	0.592	0.604	0.594	0.588	0.557	0.588	0.532
	T + DF + FF	0.655	0.656	0.655	0.654	0.601	0.583	0.601	0.573
	V + T + DF + FF	0.671	0.662	0.671	0.659	0.592	0.567	0.592	0.548
Weighted Ensemble	V + T	0.637	0.624	0.637	0.6232	0.583	0.551	0.583	0.5314
	V + DF	0.597	0.614	0.597	0.6019	0.574	0.535	0.574	0.5164
	V + FF	0.644	0.630	0.644	0.6133	0.593	0.564	0.592	0.5292
	T + DF	0.677	0.669	0.677	0.6627	0.594	0.573	0.593	0.5658
	T + FF	0.678	0.678	0.677	0.6444	0.597	0.576	0.596	0.5632
	DF + FF	0.671	0.663	0.671	0.6458	0.594	0.572	0.594	0.5625
	V + T + DF	0.597	0.590	0.597	0.5927	0.587	0.561	0.588	0.5457
	V + T + FF	0.677	0.669	0.677	0.6650	0.592	0.566	0.592	0.5460
	V + DF + FF	0.617	0.602	0.617	0.6041	0.592	0.565	0.592	0.5415
	T + DF + FF	0.685	0.686	0.685	0.6536	0.601	0.583	0.575	0.5734
	V + T + DF + FF	0.677	0.669	0.684	**0.6673**	0.583	0.587	0.585	**0.5859**

Table 3: Performance comparison of Ensemble techniques on test set

Error Analysis



(a) Dataset-1 (D1) (b) Dataset-2 (D2)

Fig 1. Proportion of misclassification among the classes of dataset-1 (D1) and dataset-2 (D2)

Error Analysis

(a) **Visual modality:** Offense (✓) **Text modality:** not-Offense (✗) **Proposed method:** Offense(✓)

(b) **Visual modality:** Offense (✗) **Text modality:** not-Offense (✓) **Proposed method:** not-Offense(✓)

(c) **Unimodal:** Not-offense (✗) **Proposed method:** Offense (✓)

(d) **Unimodal:** not-offense (✗) **Proposed method:** not-offense (✗)

Fig 2. Few correctly and misclassified examples predicted by the proposed and other approaches on the dataset-1

(a) **Visual modality:** Not-Troll(✓) **Text modality:** Troll (✗) **Proposed method:** Not-Troll(✓)

(b) **Visual modality:** Not-Troll (✗) **Text modality:** Troll (✓) **Proposed method:** Troll (✓)

(c) **Unimodal:** Not-Troll (✗) **Proposed method:** Troll (✓)

(d) **Unimodal:** Not-Troll (✗) **Proposed method:** Not-Troll (✗)

Fig 3. Few correctly and misclassified examples predicted by the proposed and other approaches on the dataset-2

Key Findings

model’s performance becomes biased towards a particular class (i.e., not-offense/not-troll) for both datasets

The possible reason of this

➔ extensive appearance of some strong words such as “Trump”, “Hilary”, “Bernie”, “Communist”, “Amala”, “Sayessha”, “boys”, “girls”, and “Anna”

➔ some world-famous person faces frequently appeared in the memes of both classes

Comparison

Techniques	Datasets	WF (%)
Suryawanshi et al. [13]	MultiOFF	54
Mishra et al. [103]	TamilMemes	30
Huang et al. [104]	TamilMemes	40
Hegde et al. [74]	TamilMemes	47
Manoj et al. [45]	TamilMemes	48
Que et al. [105]	TamilMemes	49
Bharathi et al. [106]	TamilMemes	50
Zichao et al. [73]	TamilMemes	55
Suryawanshi et al. [14]	TamilMemes	57
Proposed (weighted ensemble)	MultiOFF	66.73
	TamilMemes	58.59

Table 4: Comparative analysis of the proposed method with the existing state-of-the-art techniques

Conclusion

- Proposed technique outdoes the unimodal (i.e., image, text), multimodal, and average ensemble models with weighted f1-score of 66.73% (MultiOFF) and 58.59% (TamilMemes).

- Proposed technique outcomes are approximately 13% (in ‘MultiOFF’) and 1.69% (in ‘TamilMemes’) ahead compared to the current state of the art systems.

- Thus, results ensured the effectiveness of the proposed technique in detecting offensive and troll memes based on multimodal information.

Multimodal AI (Paper-2)

MemoSen: A Multimodal Dataset for Sentiment Analysis of Memes

[Language Resource and Evaluation Conference(LREC), 2022]

Authors: ***Eftekhar Hossain***, *Omar Sharif, Mohammed Moshiul Hoque*

Introduction

Sentiment analysis of memes has become a crucial research issue in low resource languages like Bengali.

Necessity

To mitigate the spread of negativity and understand the public expression towards an event or topic.

Scarcity of benchmark corpora in Bengali

Challenges

Challenging for the machines and humans for several reasons

- Memes are context dependent
- Visual and textual information are often disparate
- Embedded text is too short

Extracting the code-mixed and code switched text from the memes

Contribution

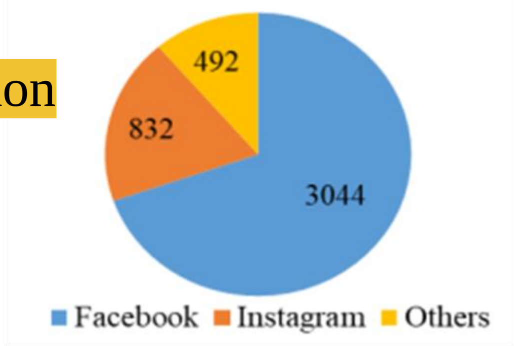
- ✓ Created the ***MemoSen***, a multimodal sentiment analysis dataset for Bengali
- ✓ Annotated with Positive, Negative, Neutral labels.

Performed extensive experiments with state-of the-art visual and textual and multimodal models.

MemoSen: A New Benchmark Dataset

Data Accumulation

Total - 4700 Memes Collected



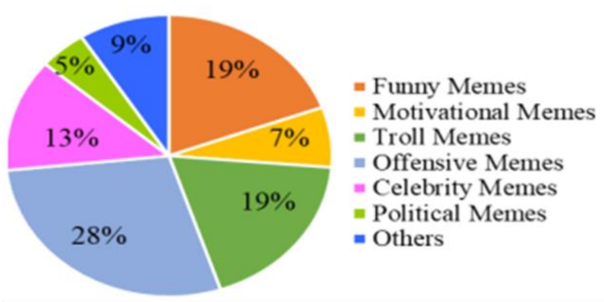


Figure 1. Source statistics of the MemoSen dataset

(a) memes without visual content

(b) memes without textual content

(c) memes with cartoons

(d) Non readable memes

Removed 332 Memes based on the above criteria

MemoSen: A New Benchmark Dataset

Data Annotation

Positive, Negative , Neutral

Positive → expresses affection, support, gratitude, accolade, and motivation

Negative → intends to denigrate, insult, disregard an entity based on its social, personal and organizational status

Neutral → intention of the memes can not infer as positive or negative

MemoSen: A New Benchmark Dataset

Process of Annotation

- ***MemoSen*** consists of 4368 memes.
- Considered memes with captions in Bengali, Bengali and English (code-mixed) or in Banglish (code-switched) manner.

✓Captions are manually extracted.

A mean kappa score of 0.674 is obtained between the three annotators

Algorithm 1: Sentiment label assigning process

Input: Set of memes with associated captions
Output: Dataset with sentiment annotation

$M \leftarrow \{m_1, m_2, ..., m_n\}$ (set of collected memes);
$MemoSen \leftarrow []$ (Multimodal sentiment dataset);
$SL \leftarrow []$ (final sentiment labels of the memes);
$L[n][2] \leftarrow \{x_1, x_2, .., x_m\}$ (initial labels);

for $m_i \epsilon M$ **do**
    $y_1 = L[i][1]$ (first annotator label);
    $y_2 = L[i][2]$ (second annotator label);
    if $(y_1 == y_2)$ **then**
        $MemoSen.append(m_i)$;
        $SL.append(y_1)$;
    else
        1. expert resolve the issue;
        2. decide final label and add it to 'MemoSen'
    end
    $i = i + 1$;
end

MemoSen: A New Benchmark Dataset

Data Samples

(a) meme shows affection (b) meme shows accolade (c) meme shows funny humor (d) insult a person (e) denigrate a group of celebrities (f) shows obscene content (g) memes with inherent sentiment (h) memes intention is incomprehensible

MemoSen: A New Benchmark Dataset

Dataset Distribution and Analysis

Class	Train	Test	Valid	Total
Positive	950	285	114	1349
Negative	2001	524	203	2728
Neutral	195	64	32	291

Table 1: Number of samples in train, test and validation set for each class

	Positive	Negative	Neutral
Positive	-	**0.355**	0.213
Negative	-	-	0.228

Table 2: Jaccard similarity of 400 most frequent words between each pair of classes

Methodology

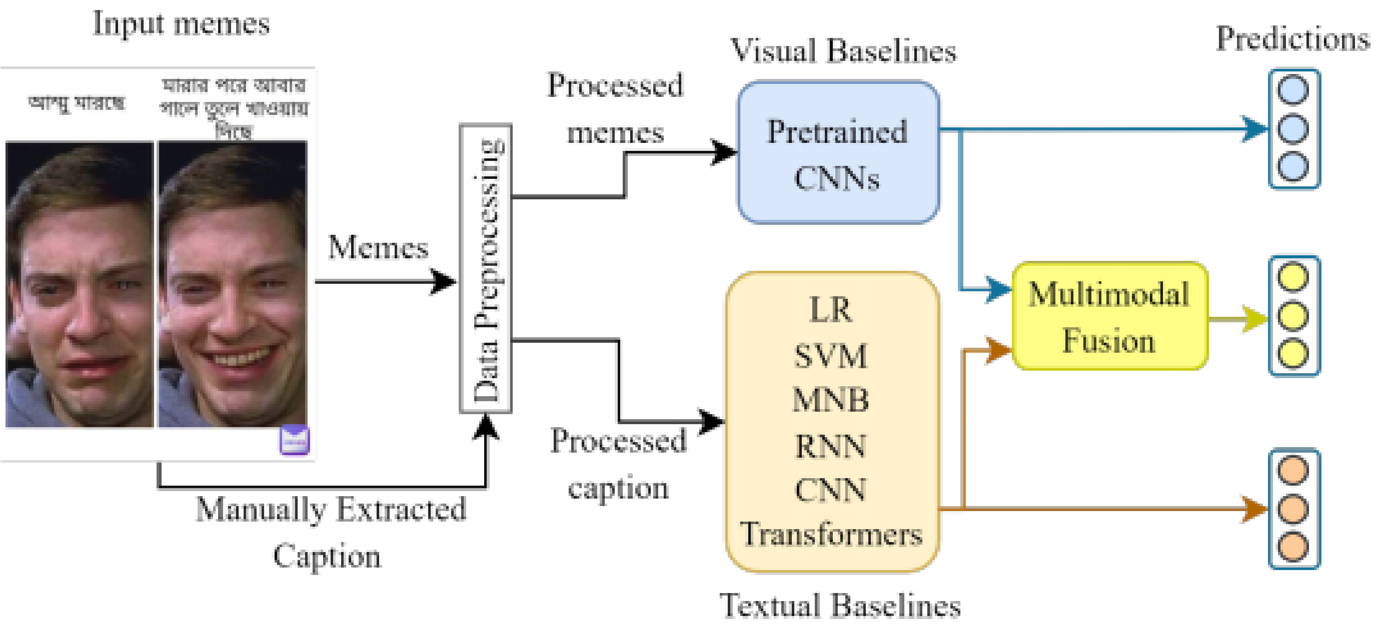


Fig 1. Abstract view of the Bengali meme sentiment classification system

Experiments and Results

Approach	Models	P	R	WF
Visual	Xception	0.587	0.615	0.579
	VGG19	0.588	0.543	0.563
	VGG16	0.582	0.571	0.559
	ResNet50	0.602	0.628	0.600
	DenseNet	0.585	0.609	0.594
Textual	LR	0.617	0.663	0.608
	MNB	0.643	0.663	0.628
	SVM	0.670	0.653	0.608
	BiLSTM (B)	0.587	0.604	0.594
	CNN (C)	0.605	0.600	0.594
	B+C	0.606	0.554	0.576
	MurIL	0.624	0.640	**0.631**
	Bangla-BERT	0.622	0.605	0.605
	XLM-R	0.360	0.600	0.450

Table 3: Performance comparison of visual and textual models on the test set.

		Models	P	R	WF
FF	R+	BiLSTM	0.625	0.633	0.626
		CNN	0.575	0.591	0.582
		BiLSTM+CNN	0.615	0.578	0.592
		MurIL	0.525	0.392	0.419
		Bangla-BERT	0.510	0.557	0.508
DF	R+	BiLSTM	0.644	0.631	0.635
		CNN	0.663	0.628	**0.643**
		BiLSTM+CNN	0.566	0.592	0.575
		MurIL	0.552	0.554	0.543
		Bangla-BERT	0.504	0.394	0.329

Table 4: Performance comparison of multimodal models on test set. Here, (+) sign denoted the aggregation of visual and textual models

Error Analysis

(a) **Visual Model:** Negative (✗)
Textual Model: Neutral (✗)
Multimodal Model: Positive (✓)

(b) **Visual Model:** Neutral (✗)
Textual Model: Positive (✗)
Multimodal Model: Negative (✓)

(c) **Visual Model:** Negative (✗)
Textual Model: Negative (✗)
Multimodal Model: Positive (✗)

Fig 2. Example memes where aggregation of the visual and textual modalities yield better predictions

Error Analysis

Model's performance is more biased towards negative class
Imbalanced dataset

Observations

- large number of words are overlapped between the classes
- the code-mixed and code-switched words
- the consistent visual features (i.e., familiar person faces) across the memes of the different classes

Conclusion

- ➔ We introduced MemoSen, a multimodal benchmark dataset.
- ➔ The evaluation exhibits that the integration of multimodal information significantly improves **(about 1.2%)** the meme sentiment classification

Multimodal AI (Paper-3)

A Deep Attentive Multimodal Learning Approach for Disaster Identification from Social Media Posts *[IEEE Access Journal, 2022]*

Authors: ***Eftekhar Hossain***, *Mohammed Moshiul Hoque, Enamul Hoque, Md Saiful Islam*

Multimodal AI (Paper-3)

#terriblefire
#plascobuilding
#nostalgia #tragedy
#buildingcollapse

Drawbacks of Previous Works

- ➔ While many studies have shown the effectiveness of combining text and image contents for disaster identification
- ➔ Most previous work focused on analyzing only the textual modality and/or applied traditional RNN or CNN which might lead to performance degradation in case of long input sequences.

Objective

➔ Develop an effective computational model for identifying disaster-related information by synergistically integrating features from visual and textual modalities.

Contribution

- ➔ Propose a multimodal architecture that utilizes ResNet50 and BiLSTM recurrent neural network with attention mechanism to classify the damage-related posts
- ➔ compare the performance of the proposed model with a set of existing unimodal (i.e., image, text) and multimodal techniques.

- ➔ Empirically evaluate the proposed model on a benchmark dataset and
- ➔ demonstrated how introducing attention could enhance the system performance through an intrinsic evaluation.

Contribution

- Propose a multimodal architecture that utilizes ResNet50 and BiLSTM recurrent neural network with attention mechanism to classify the damage-related posts
- compare the performance of the proposed model with a set of existing unimodal (i.e., image, text) and multimodal techniques.

- Empirically evaluate the proposed model on a benchmark dataset and demonstrated how introducing attention could enhance the system performance through an intrinsic evaluation.

Problem Formulation and Dataset

- ➔ Automatically classify disaster types such as floods, fires, earthquake etc. from social media posts
- ➔ Disaster Types:
    - ◆ Damage to infrastructure (DI)
    - ◆ Damage to nature (DN)
    - ◆ Fires (F)
    - ◆ Floods (Fl)
    - ◆ Human damage (HD)

Methodology

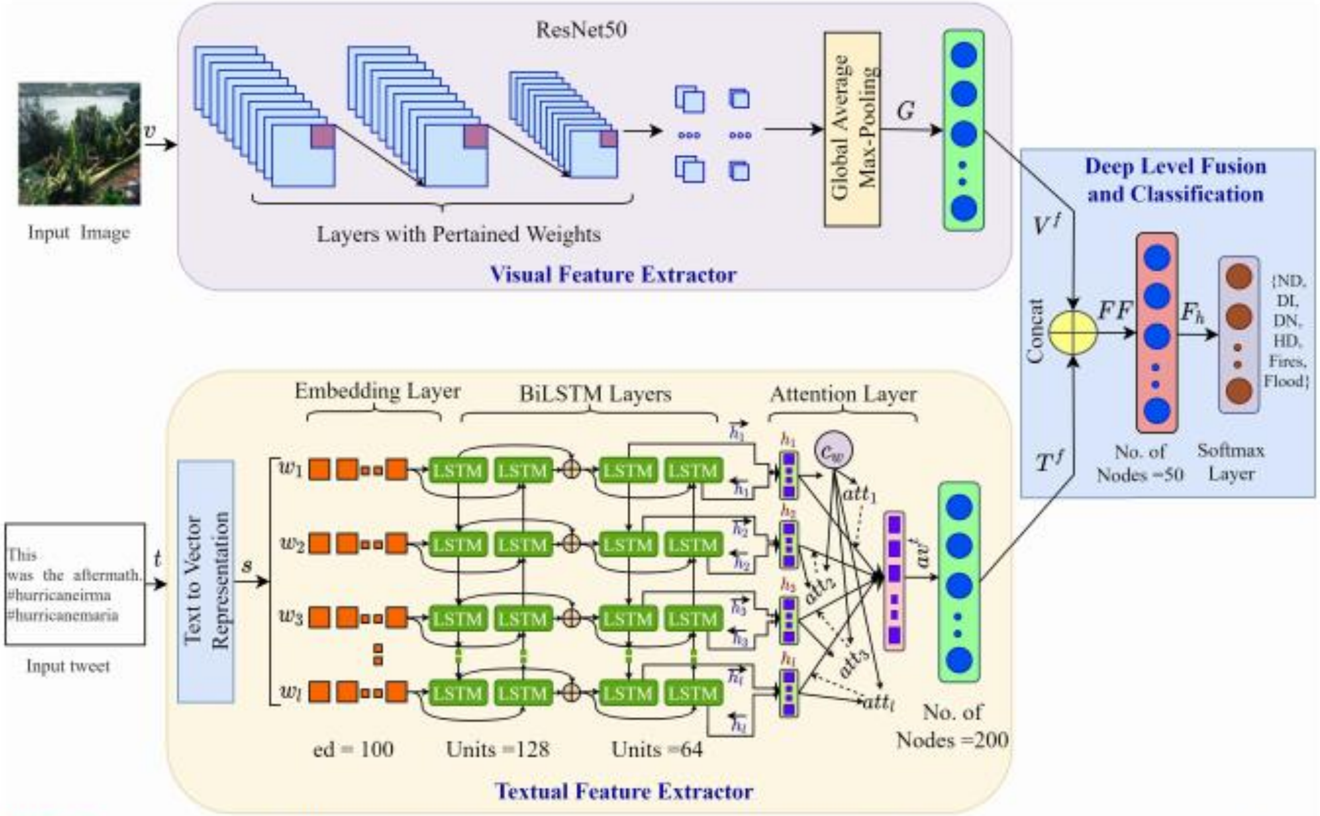


FIGURE 2. Our proposed multimodal architecture for disaster identification: the upper block represents the visual feature extractor module and the bottom block is the textual feature extractor module. Here, *v* and *t* indicates the preprocessed image and text respectively. The features extracted from the two modules are passed through the deep level fusion and classification layer to classify the sample.

Results

Approach	Models	P(%)	R(%)	WF(%)
Visual	VGG19 [49]	81.06	81.51	81.21
	Inception [50]	77.41	77.91	77.38
	ResNet50 [40]	81.88	81.51	81.63
Textual	BiLSTM	85.92	85.45	85.57
	CNNText	84.97	84.25	84.45
	BiLSTM+CNNText	85.54	84.42	84.70
	BiLSTM+Attention	89.14	88.87	88.75
Multimodal	VGG19+BiLSTM	81.98	76.20	78.14
	VGG19+CNNText	74.39	73.46	72.57
	VGG19+BiLSTM+CNNText	78.24	77.74	77.67
	VGG19+BiLSTM+Attention	89.54	89.38	89.19
	Inception+BiLSTM	82.21	74.48	77.01
	Inception+CNNText	79.66	79.10	78.28
	Inception+BiLSTM+CNNText	77.29	78.08	77.38
	Inception+BiLSTM+Attention	81.18	80.82	80.48
	ResNet50+BiLSTM	84.22	81.34	81.90
	ResNet50+CNNText	77.68	78.42	77.45
	ResNet50+BiLSTM+CNNText	80.30	79.62	79.84
	ResNet50+BiLSTM+Attention (**Proposed Method**)	93.35	93.15	**93.21**

Table 1: Performance comparison of different unimodal and multimodal models on the test set

Error Analysis

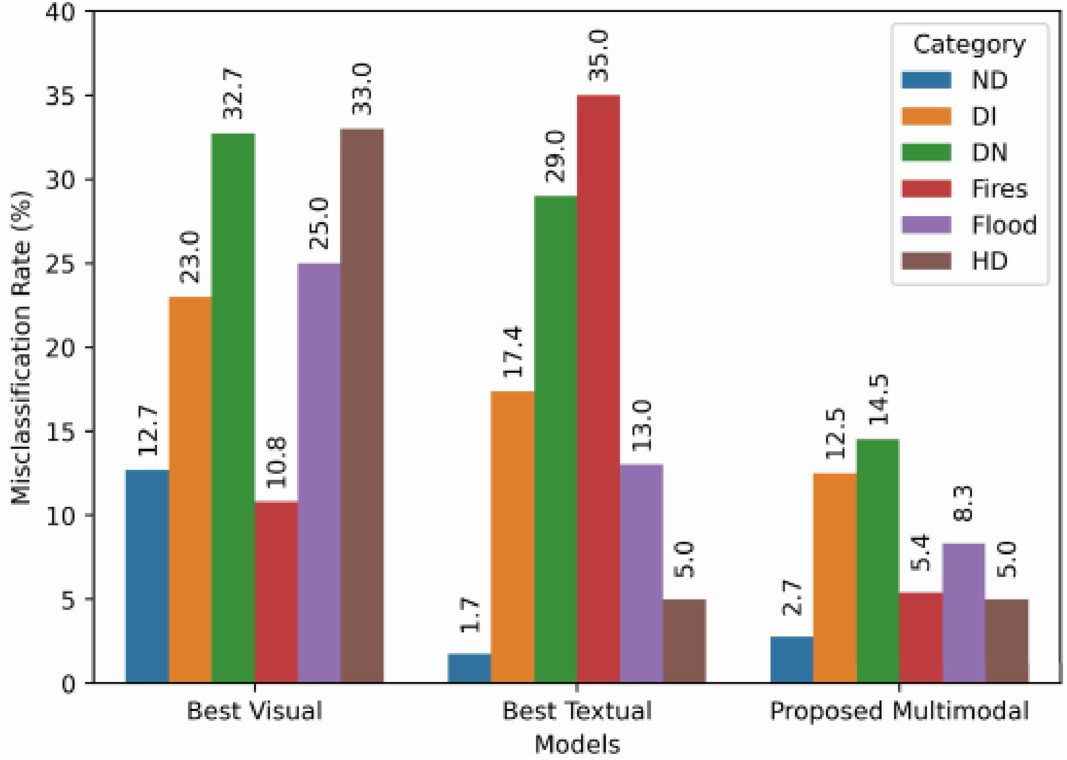


Fig 3: Error rate analysis of the individual classes with different approaches.

Error Analysis

Sample	Image	Tweet	Actual label	Predicted label
(1)		MooseMonday with my favorites! A couple #bullmoose from the weekend! #moose #wildlife #wildlifephotography #mammal #wilderness #wildernessculture	ND	**Visual Modality:** DN (✗) **Visual Modality:** DN (✗) **Proposed Multimodal:** ND(✓)
(2)		#sandy #youwhore massive #treebranch fell and took out two 8 foot sections of the fence in the pic.#fallentree #30ftdrop #sandydamage	DI	**Visual Modality:** DN (✗) **Textual Modality:** DN (✗) **Proposed Multimodal:** DI(✓)
(3)		Please curtail this hazardous 20+ year practice.#csi #uci #bordertown #newportbeach #mudslide #caution #landslide #smashingpumkins	DN	**Visual Modality:** DI(✗) **Textual Modality:** DI(✗) **Proposed Multimodal:** DN(✓)

Table 2: Example image and tweet text pairs where model aggregation of the input modalities produce better results

Intrinsic Performance Analysis

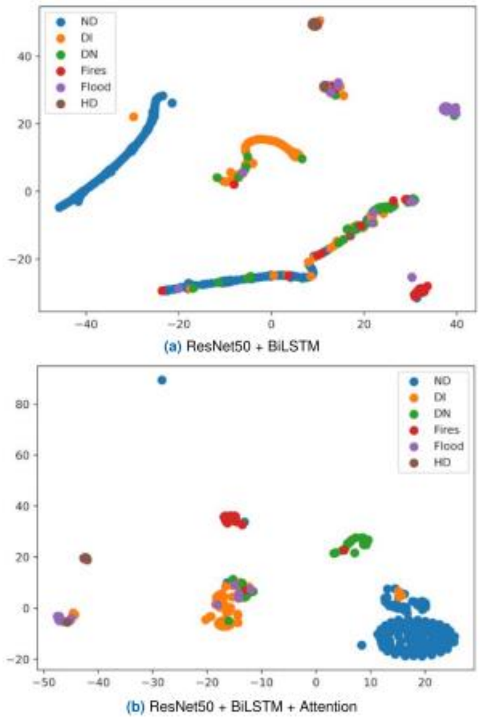


Fig 4. Scatter plots of test input features extracted by the multimodal models (a) without attention layer and (b) with attention layer

Comparison

Method	Modality	WF(%)
Mouzannar et al. [7]	Image+Text	92.14
Ferda et. al [8]	Image+Text	75.11
Kumar et. al [11]	Image+Text	77.84
Nguyen et al [29]	Image-only	75.17
Caragea et al. [21]	Text-only	75.23
Aipe et. al. [22]	Text-only	76.76
Yu et. al. [23]	Text-only	78.47
Xiao et. al [18]	Text-only	86.05
Proposed	**Image+Text**	**93.21**

Table 3. Results of comparison concerning WF-score

Conclusion

- ➔ presented a multimodal approach that can effectively learn from the image and text data.
- ➔ Proposed model outperforms the baseline unimodal and multimodal models by acquiring the highest weighted F1-score of 93.21%.
- ➔ Comparative analysis illustrated that the proposed method outcome is approximately 1% and 7% ahead of the existing start-of-the-art models.

Future Directions

- ➔ Multimodal Hate Speech Detection
- ➔ Multimodal Emotion Recognition
- ➔ Multimodal Event Detection
- ➔ Multimodal Humor or Sarcasm Detection

ThankYou