# *I didn't mean what I wrote!* Exploring Multimodality for Sarcasm Detection

Suyash Sangwan
*Department of CSE*
*IIT Patna*
Patna, India
suyash.mtmc17@iitp.ac.in

Md Shad Akhtar
*Department of CSE*
*IIIT Delhi*
New Delhi, India
shad.akhtar@iiitd.ac.in

Pranati Behera
*Department of CSE*
*IIT Patna*
Patna, India
1811mc10@iitp.ac.in

Asif Ekbal
*Department of CSE*
*IIT Patna*
Patna, India
asif@iitp.ac.in

***Abstract*—Sarcasm detection is, inherently, a non-trivial problem where people express negative sentiment using positive insinuation words. Traditional approaches, in general, rely on the textual information to detect the incongruity between the surface meaning and the actual meaning. However, textual information is not always sufficient, and, often, other sources of information (e.g., visual) provides an important clue for sarcasm detection. In this paper, we propose an effective method based on deep learning that utilizes both textual and visual information for multi-modal sarcasm detection. Our proposed approach is based on the recurrent neural network that aims to exploit the interaction among the input modalities for the prediction. Experimental results suggest that the incorporation of visual modalities plays a decisive role in performance improvement.**

***Index Terms*—Sarcasm detection, Multimodal sarcasm detection, Deep Learning,**

## I. Introduction

Interest in Natural Language Processing (NLP) has grown many-folds during the past decade and a majority of these are centered around the analysis of social media posts to solve the problems like sentiment analysis [1], emotion detection [2], sarcasm detection [3], irony detection [4], metaphor detection [5] etc. Accuracy and robustness of NLP models are often affected by untruthful sentiments that are often of sarcastic in nature. For example, a sentence like *"So thrilled to be on call for work the entire weekend !"* could be naively classified as a sentence with a high positive sentiment. However, it is actually the negative sentiment that is clearly implied through sarcasm. Sarcasm is a sharp, bitter, or cutting expression or remark; a bitter jibe or taunt. Most noticeable in spoken word, sarcasm is mainly distinguished by the inflection with which it is spoken and is largely context-dependent. Sarcasm may employ ambivalence and is often confused with irony and satire. Irony describes the situations that are strange or funny because the things happen in a way that seems to be the opposite of what we expected. For example, a policeman violating a law. Satire means making fun of people by imitating them in ways that expose their stupidity or flaws. As with satire, sarcasm depends on the listener or reader to be in on the joke. Hence, irony employed in the service of mocking or attacking someone is 'Sarcasm'. Saying *"Oh, you're soooo clever !"* with sarcasm means the target is really just a dunderhead.

Most of the existing approaches to date have considered sarcasm detection task primarily as the 'text categorization' problem, where sarcasm is detected using lexical indicators (like interjections and intensifiers), emojis, hashtags, or the presence of incongruity in the input, etc. However, the text-only approaches may not be sufficient to infer the statements as sarcasm. For example, the sentence '*Loved walking on this beautiful spring day.*' seems to have a positive sentiment (or, non-sarcastic). However, if we also consider the visual information, as depicted in Figure 1b, it is evident that the sentence is actually a sarcasm, where the user has expressed his/her displeasure over the snowy weather. The incongruity between the visual and textual modalities implies the statement as sarcasm, and the systems that can handle such cases may provide more accurate predictions. In some cases, the pointer to the sarcasm comes from an individual modality, such as the textual modality (e.g., '*I love being ignored*') or the visual modality (e.g., example in Figure 1c), while in other cases, both the modalities contribute towards the sarcasm detection (e.g., Figures 1a, 1b, 1e). In Figure 1, we present a few sarcastic examples where the transcripts (or textual information) alone are incapable of detecting sarcasm.

Therefore, motivated by the above analyses, we propose to incorporate the textual and visual modalities in an RNN-based system for the multi-modal sarcasm detection in Instagram[1] posts. We aim to learn the interaction between the textual ($T$) and visual ($V$) modalities, and interaction between the textual ($T$) and transcripts ($T_N$) extracted[2] from the visual information in our proposed architecture. Since the visual and transcripts always overlap with the textual modality, i.e., textual information is always present and moreover, when we experiment with unimodality, i.e., using either only text or visual, we found that text yields better predictions. Our main motivation is, therefore, to exploit information from the visual modality and to verify whether it assists text for the final prediction. We introduce a gating mechanism that obtains a weighted representation of 'V' through the interaction between 'T' and 'V', and a weighted representation of '$T_N$' through the

[1]https://instagram.com

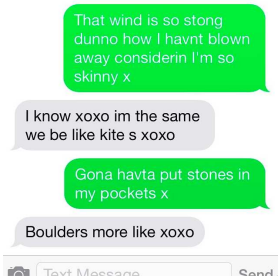
[2]We use Google OCR (Optical Character Recognition) for extracting the text from the visual modality.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

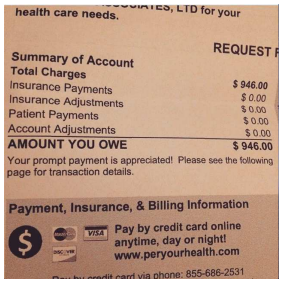
(a) Text: *Someone is excited for sweater season.*

(b) Text: *Loved walking on this beautiful spring day.*


(c) Text: *#1Quote #Happy #Fun #LoL #LaughOutLoud #Quote*




(d) Text: *This is why we get on so well.*




(e) Text: *Life is just going great this year so far for me.*

Fig. 1: Few Sarcastic examples involving multimodal information. For each case, the transcripts are not sufficient to infer the statements as sarcasm.

interaction between 'T' and '$T_N$'. We evaluate our proposed approach on two multi-modal sarcasm detection datasets that justify the incorporation of textual and visual information for the improved sarcasm detection. We summarize our main contributions as follows:

- To the best of our knowledge, this is the very first systematic attempt to identify sarcasm using both text and image using deep neural networks. The evaluation shows that our proposed technique achieves good performance improvement when image is considered along with text.
- We prepare a multimodal gold standard dataset of 1600 Instagram posts by annotating with the sarcasm class. After removing the hashtags these are annotated manually, first using only textual modality and then using both text and image modalities.
- We extract embedded text from the images using Google's OCR (Optical Character Recognition) and use these extracted features in building a better sarcasm classifier.

## II. Related Work

Sarcasm is an important problem in Natural Language Processing (NLP) and Text Mining. Compared to the other related tasks like sentiment analysis, sarcasm is more difficult because it is mainly distinguished by the inflection with which it is spoken and is largely context-dependent. Multi-modality is, nowadays, gaining attention to build robust systems by utilizing information from more than one sources. Our literature survey reveals there has been almost no effort to build a multi-modal sarcasm detector *except* the one proposed in [6]. In our work, we use deep-neural network framework and use multimodality, i.e., both text and image to detect sarcasm.

Below we present some of the works for text-based sarcasm detection. Bharti et al. [7] proposed an approach for parsing-based lexicon generation algorithm and another approach to detect sarcasm based on the occurrence of the interjection word and then combined both of these approaches to detect sarcasm. Social-media platforms like Twitter, Instagram, Tumblr, etc. are used commonly for collecting data. Liebrecht et al. [3] and Reyes et al. [4] used the hashtag based techniques to collect sarcastic data from different social media platforms. Liebrecht et al. [3] worked on a set of 3.3 million Dutch tweets and used n-gram features like unigram, bigram, and trigram and showed that sarcasm is often signaled by hyperbole, using intensifiers and exclamations.

Till date, almost all the sarcasm detection approaches treat this task mainly as 'text categorization' problem. Lunando et al. [8] used naive Bayes classifier and Support Vector Machine (SVM), and introduced two additional features, i.e., negativity information, and the number of interjection words, to detect sarcasm. Gonzalez et al. [9] used '#sarcasm' and '#sarcastic' hashtags to accumulate sarcastic tweets from Twitter and used Logistic Regression classifier, and SVM with SMO (Sequential Mining Optimization) and investigated the impact of lexical and pragmatic factors on machine learning effectiveness for identifying sarcastic utterances.

Contextual information plays an important role in sarcasm detection. Bamman et al. [10] showed that by including extra-linguistic information from the context of an utterance such

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

as properties of the author, the audience, and the immediate communicative environment, the gains in accuracy can be achieved compared to the purely linguistic features. Our current work differentiates from these existing works in the sense that we use image as a contextual cue to the text, and use these for detecting sarcasm. We use the fusion of image and text both because only text does not always provide the complete description and image may assist in predicting the correct label. Our thorough analysis to text and image reveal that, in many cases, image modality contradicts to the text modality, and hence depicts the presence of sarcasm.

## III. Proposed Methodology

In this section, we describe our proposed methodology as shown in Figure 2, where we aim to leverage the multi-modal or contextual information. For text modality, we use RNN (Recurrent Neural Network) based framework. The proposed model takes multi-modal *(text, image, and transcript extracted from image)* information as an input and processes it. For text modality, after pre-processing, each word and emoji are represented as a 100-dimensional feature vector using GloVe. Each word of the final pre-processed text is then applied to a separate bi-directional Gated Recurrent Unit (GRU). Bi-GRU is used for capturing the contextual information and only one output representation for all the input words is obtained at the end. Similarly if the text is present in its corresponding image modality, the text extracted from it using OCR (Optical Character Recognition) is also applied word-wise to separate bidirectional GRUs, and one output representation for all the input words is obtained at the end, else a (1*100) dimensional zero vector is appended for that modality. Image features are extracted using 'VGG-16' which has been trained on 'Imagenet' [11] and a (1*4096) dimensional feature vector is obtained for each image input. Features extracted from all the three modalities are applied to a fully-connected layer to make the output dimension same.

### A. Gating Mechanism

It is true that all the modalities cannot contribute equally to the final prediction. Moreover, image modality and transcript extracted from the image cannot always help text in better predictions due to the presence of noise. For example, if only text is written in image modality as shown in Figure 1c, then the features extracted using VGG-16 are not of much importance. Hence, we apply a gating mechanism on image modality to decide the weight of image modality (I) w.r.t. textual modality (T). So at first, we concatenate text and image modality features and then apply a dense layer to the concatenation. Now, we pass the concatenated feature vector ($x$) through a Sigmoid which gives the value ($x'$) which represents the weight of the image modality w.r.t. text modality and finally that weight is multiplied to the entire image feature vector (I) giving ($I_T$) = $x'I$, which helps in either passing or suppressing the image features depending on their role in final prediction. Equations for image gating w.r.t. text are:-

$$x = fully\_connected([T, I]) \quad (1)$$

$$x' = sigmoid(x) \quad (2)$$

$$I_T = x'I \quad (3)$$

Similarly, noise may also be present in the transcript extracted from the image modality. For example, in Figure 1e, OCR extracts complete text but it is not completely useful for sarcasm detection. Hence, we need to calculate the weight of text extracted from the image modality w.r.t. the text modality as above. Equations for gating ($T_N$) w.r.t. T are:-

$$y = fully\_connected([T, T_N)]) \quad (4)$$

$$y' = sigmoid(y) \quad (5)$$

$$(T_N)_T = y'T_N \quad (6)$$

Finally, we concatenate both of these gated representations of I and $T_N$ with text modality (T) along with the residual connections of the modalities for the final prediction. We append residual connections of the modalities to boost the gradient flow to the lower layers. This concatenation is then passed through sigmoid layer for final prediction where we use 0.5 as a threshold value.

## IV. Dataset, Experiments, and Analysis

In this section, we describe the dataset used for our experiments, experimental methodology, input features, pre-processing steps, results, error analysis and finally state-of-the-art models that we use to compare our results with.

### A. Dataset

Instagram is a micro-blogging platform extensively used to express thoughts, reviews or current events and convey information in the form of short texts with the help of images. The relevant context of the posts is often specified with the use of hashtags. We evaluate our proposed approach on two Instagram datasets described below:

- **Silver-Standard Dataset:** We compile a multimodal sarcasm detection dataset with approximately 20K Instagram posts obtained from [6] and [12]. We ensure the even distribution of *sarcastic* and *non-sarcastic* posts in the dataset, i.e., approximately 10K sarcastic and 10K non-sarcastic posts. The posts are classified as *sarcastic* or *non-sarcastic* based on the hashtags *#sarcastic* or *#sarcasm*, i.e., posts containing these hashtags are labeled as *sarcastic*, while the posts without these hashtags are labelled as *non-sarcastic*. We further clean up the dataset by removing the hashtags '*#sarcastic*' and '*#sarcasm*' from the posts.

- **Gold-Standard Dataset:** We prepare a gold-standard dataset where we randomly select 1600 positive examples (i.e., denoting sarcasm) from the datasets that we describe above. Then we ask the different annotators to annotate these examples using only text modality and then based

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

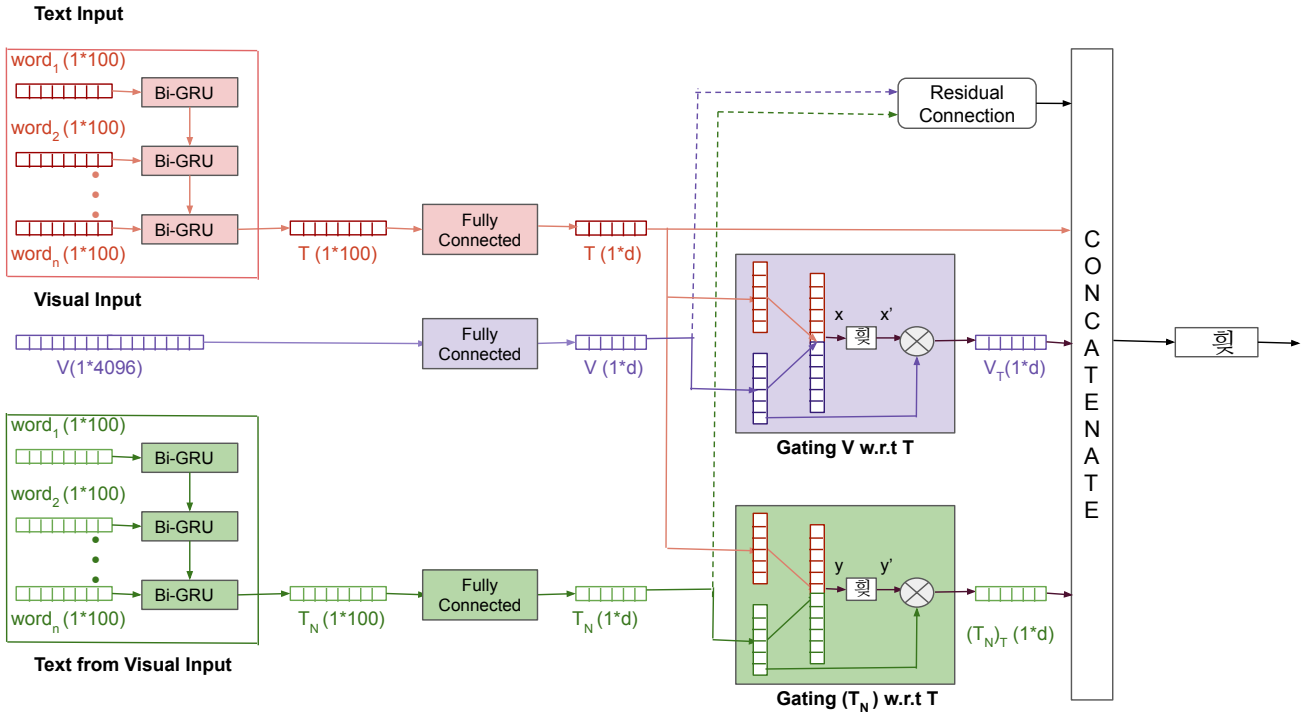



Fig. 2: Overall Architecture of the proposed framework where '$T$', '$I$', and '$T_N$' represents textual, image, and transcript extracted from image modality, respectively and '$I_T$', '$(T_N)_T$' represents I w.r.t. T and $T_N$ w.r.t. T respectively.

on the consensus, these examples are labeled. In the second step, we provide both text and image modalities to the annotators for all the 1600 instances and ask them to re-annotate. We observe many such examples which are sarcastic according to the author but are non-sarcastic according to the reader. For example, the text *'Lets do this. #idowhatiwant #idatemyself #sarcastic #humor'* has #sarcastic, therefore its label is sarcastic according to the author. But before giving it to the annotators, we remove the hashtags that we used to crawl the data, hence text becomes *'Lets do this. #idowhatiwant #idatemyself #humor'*, which is non-sarcastic.

We compute the Fleiss' kappa [13] for the above metrics to measure inter-rater consistency. We obtain the kappa score of 0.81, indicating substantial agreement.

For the experiments, we perform 5-fold cross-validation on the silver-standard dataset, while for the gold-standard dataset we train on random 1,000 posts and evaluate on the remaining 600. Since the number of samples in the gold-standard dataset is few, and, in general, we need a good amount of instances to effectively train a neural network, we adopt a transfer learning approach where we utilize the learned weights of the silver-standard dataset for initialization of the network during gold-standard training. The datasets can be found at http://www.iitp.ac.in/~ai-nlp-ml/resources.html.

### B. Feature Extraction

We extract textual features using 'GloVe' embeddings [14], and the image features using 'VGG-16' [15] which has been trained on the 'Imagenet'. Further, we extract the transcript in the image through Google OCR (Optical Character Recognition). These transcripts are, then, converted into 100-dimension GloVe embeddings for the experiments. The number of image features extracted from VGG-16 is 4096.

### C. Experiments

We evaluate our proposed approach on the Instagram dataset. We use the Python-based Keras library for its implementation. For evaluation, we compute the accuracy value to measure the performance of the model. We use the grid search to find the optimal hyper-parameters for our experiment. We use Bi-GRUs with 300 neurons each. We set dropout to 0.3, batch size to 100 and the number of epochs to 20. We use ReLu as an activation function and Adam as an optimizer. We use Sigmoid for sarcasm classification where we choose a threshold value of 0.5, i.e., we consider the input as sarcastic if the output at Sigmoid is above 0.5 and binary cross-entropy as the loss function.

We evaluate our proposed model with all input combinations, i.e., unimodal (only text modality (T), only image modality (I)), bi-modal(a combination of text and image modality (T+I)) and tri-modal (combination of text, image and transcript (T+I+$T_N$)) as shown in Table I. For consistency, we

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

| Modality | Without Gating | | Gating w.r.t Text | |
|---|---|---|---|---|
| | #Correct Predictions | Accuracy | #Correct Predictions | Accuracy |
| T | 3217 | 80.42% | - | - |
| I | 2928 | 73.20% | - | - |
| T+I | 3291 | 82.28% | 3355 | 83.87% |
| T+I+$T_N$ | 3325 | 83.12% | 3368 | 84.22% |

TABLE I: Results on the silver-standard dataset. Here, T represents Text, I represents image and $T_N$ represents transcript extracted from the image modality.

| Transfer Learning | Gating w.r.t. Text | T | | T+I | | T+I+$T_N$ | |
|---|---|---|---|---|---|---|---|
| | | #Correct Predictions | Accuracy | #Correct Predictions | Accuracy | #Correct Predictions | Accuracy |
| × | × | 360 | 60.00% | 382 | 63.67% | 390 | 65.00% |
| ✓ | × | 397 | 66.17% | 413 | 68.83% | 421 | 70.17% |
| × | ✓ | - | - | 398 | 66.30% | 403 | 67.17% |
| ✓ | ✓ | - | - | 420 | 70.00% | 429 | 71.5% |

TABLE II: Results on the Gold-standard dataset. Here, T represents Text, I represents image and $T_N$ represents transcript extracted from the image modality.

use the same hyper-parameters for training all the models. We obtain the best results when we use gating and all the three modalities altogether.

Evaluation results on the gold-standard dataset are shown in Table II. In the first experiment, we do not consider either gating or transfer learning. In the second experiment, only transfer learning is considered. In the third experiment, only gating is considered and in the fourth experiment, we consider both gating and transfer learning. The results show that using only transfer learning provides better performance than using only gating, and the best performance is obtained when all the three modalities are used along with transfer learning and gating mechanism.

For the silver standard dataset having *(having 20k posts)*, we obtain the accuracy values of 80.42%, 73.20%, and 84.22%, respectively, using only text, only image, and their combination *(i.e. textual, image and transcript)*. In gold-standard dataset, using only text, text and image, and their combination *(i.e. textual with image and transcript)* yield the accuracy values of 66.17%, 70% and 71.50%, respectively.

In all the experiments, we perform statistical significance t-test [16], which reveals that the improvement in the proposed model (using T+I+$T_N$) over the model with (T+I) and model with T only are significant [3].

### D. Baseline

As a baseline, we train a neural network with the textual features that comprise of lexical, pragmatic and linguistic incongruity. The neural network has the following configurations: No. of hidden layers: 2, No. of neurons in the first layer: 150, No. of neurons in the second layer: 20, Number of epochs: 30, Batch size: 30. Here, 'sigmoid' is used for final classification, and 'binary cross-entropy' is used as the loss function, with 'Adam' as an optimizer.

The features used are described in further details as below:

1) **Lexical features:** We use unigrams in order to extract the lexical information from the tweets. A dictionary is created using the training corpus. Each unique word is mapped onto a particular identifier (denoted as ID), and we use these ID numbers as the features. The value corresponding to each such feature number is the frequency of occurrence of that particular word in the tweet for which we generate the feature values. The dictionary would be large, owing to the vocabulary available in the corpus. The tweet would contain only a few words from this large vocabulary. Hence, the feature vector would naturally contain a lot of 0's corresponding to the words that do not appear in the tweet but are present in the dictionary. The IDs with values (frequency) 0 can be discarded since we look for the presence of words prevalent in sarcastic tweets which can be a potentially important indicator while the absence of words conveys no information.
2) **Pragmatic features:** We extract features like the number of punctuation marks in the tweet, number of capitalized letters in the tweet, number of emoticons in the tweet *(emoticons can be captured using regular expressions and UTF-8 encoding)*, and number of slang laughter expressions in the tweet *(like lolz, rofl, lmao, etc. which can also be captured using regular expressions)*
3) **Linguistic incongruity:** It suggests that sarcasm expressions consist of a positive sentiment contrasted with a negative situation or vice-versa, e.g., *'I love being ignored.'* We use 'SentiStrength' [4] tool to obtain the polarity of each word and complete sentence where -1 to -5 indicate negative polarity, 0 indicates neutral, and +1 to +5 indicate positive polarity. Then we extract the features like the largest positive or negative subsequence

[3]t-test conducted at 5% (0.05) significance level

[4]http://sentistrength.wlv.ac.uk/

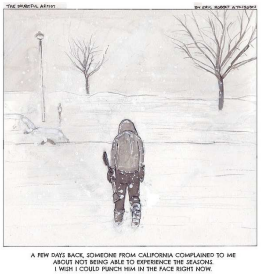
978-1-7281-6926-2/20/$31.00 ©2020 IEEE

(a) Text: *Loved walking on this beautiful spring day.*

(b) Text: *A very foggy morning in Poway.*

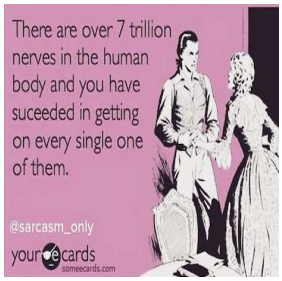
(c) Text: *Someone is excited for sweater season.*

(d) Text: *Smoking a cig while walking through the smoke dispensed by the grenade.*



(e) Text: *This is my favorite time of year*


(f) Text: *Yes!!! Whole reason I refuse to hold an account with Fakebook!*




(g) Text: *A few people come to mind when I see this.. I think you know a few of them.*

Fig. 3: Examples showing critical cases from different experimental setups

| | | Examples from Figure 3 | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | (a) | (b) | (c) | (d) | (e) | (f) | (g) |
| Actual | | Sar | Sar | Sar | Non-Sar | Sar | Sar | Sar |
| Predicted | T | Non-Sar | Non-Sar | Non-Sar | Sar | Non-Sar | Non-Sar | Non-Sar |
| | T+I | Sar | Sar | Sar | Non-Sar | Non-Sar | Non-Sar | Non-Sar |
| | T+I+$T_N$ | Sar | Sar | Sar | Non-Sar | Sar | Sar | Sar |

TABLE III: Predictions obtained by different experimental setups for the examples shown in Figure 3 (Here, T represents text, I represents image and $T_N$ represents transcript extracted from image modality)

of words, number of words with positive and negative polarity, number of sentiment incongruities *(A single numeric feature value which gives the count of the number of times a positive word is followed by a negative word and vice-versa.)* and lexical polarity *(where the overall polarity of the sentence is calculated)*.

To further explain how SentiStrength works, let us consider the following example:

***Input:*** *Wow I'm shocked.*

***Output:*** *Wow[2] I'm[0] shocked[-2]*

The baseline model yields recall, precision and F-measure of 53.2%, 78.95%, and 63.57%, respectively.

### *E. Detailed Analysis of the Results*

In this section, we closely analyse the outputs of our model in order to study the effect of different modalities. We exhibit a few cases to show where image modality helps the text modality for accurate prediction, and a few cases to show where transcript extracted from the image plays an important role in the final prediction. For example, the text modality of Figure 3a seems to be non-sarcastic due to presence of the words like *'loved'*, *'beautiful'*, but its image modality depicts *'snow'* which contradicts with the textual modality (*'spring'*), hence depicting sarcasm. Similarly, in Figure 3b, the textual modality misclassifies it as *'non-sarcastic'* but sarcasm is detected using image as there is no fog present. In Figure 3c, presence of the word *'excited'* in textual modality depicts positive sentiment, i.e., *'non-sarcasm'*, but the image modality depicts *'sadness of the dog'* which is again confirming sarcasm.

In Figure 3d, the textual modality, i.e., *Smoking a cig while walking through the smoke dispensed by the grenade* seems to be *'sarcastic'*, but image modality also depicts the same scenario and hence making it non-sarcastic. Therefore, all these cases show that image modality assists textual modality in better predictions.

As a contradiction, refer to the examples 3e, 3f and 3g, where the textual modality and combination of textual modality with image modality both misclassify the input. In Figure

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

| Systems | Modality | #count | Accuracy (%) |
|---|---|---|---|
| Schifanella et al. [6] (SVM) | Text (*n-gram*) | 260 | 42.13 |
| | Image (*VSF*) | 153 | 39.13 |
| | Text + Image (*n-gram + VSF*) | 292 | 45.6 |
| Schifanella et al. [6] (Deep Network Adaptation - DNA) | Text (*1gram*) | 244 | 39.15 |
| | Image (*AVR*) | 151 | 38.6 |
| | Text + Image (*1gram + AVR*) | 263 | 42.56 |
| Das and Clark [17] | Image | 155 | 40.8 |
| Huang et al. [18] | Text (Semantic Attention Model - *SAM*) | 170 | 45.6 |
| | Image (Visual Attention Model - *VAM*) | 113 | 41.9 |
| | Text + Image (Multimodal Attention Model - *MAM*) | 192 | 49.1 |
| | Deep Multimodal Attention Fusion - *DMAF* (*SAM + VAM + MAM*) | 215 | 51.2 |
| Proposed | Text | 397 | 66.17 |
| | Text + Image | 420 | 70.0 |
| | Text + Image + Transcript | 429 | 71.5 |

TABLE IV: Comparative analysis on the gold standard dataset.

3e, the transcript extracted from its image is *'A FEW DAYS BACK, SOMEONE FROM CALIFORNIA COMPLAINED TO ME ABOUT NOT BEING ABLE TO EXPERIENCE THE SEASONS. I WISH I COULD PUNCH HIM IN THE FACE RIGHT NOW.'* and is a sarcastic statement, and hence helps in correcting its label as *'sarcastic'*. In Figure 3f and Figure 3g, the labels predicted by the unimodal and bimodal based models are 'non-sarcastic' because there is no direct sarcasm conveyed by the text here, and moreover no significant information from the image is present. But after extracting the text from its image modality, the model correctly classifies it as *'sarcastic'*.

### *F. Comparison to the state-of-the-arts*

In this section, we compare our proposed approach against various existing systems [6], [17], [18]. For comparison, we have implemented and evaluated these existing models on our gold standard dataset and reported the results in Table IV. Schifanella et al. [6] proposed a deep multimodal fusion approach, called Deep Network Adaption (DNA), to exploit the semantics of the images for sarcasm prediction. For text, they have utilized only 1-gram features, while for the visual modality, features were extracted from the ImageNet [11]. These two representations are then concatenated and processed through a series of non-linear layers for the final prediction. Das and Clark [17] adopted a CNN-based architecture to predict the sarcasm in Flickr posts. They targeted the visual cues present in the image for the sarcasm prediction. In another work, Huang et al. [18], applied attention mechanism to extract the sentiment from an image and its description. Their prime goal was to exploit the inter-correlation between the visual and semantic content through the attention model. We adopt their model for the sarcasm prediction. It is evident from Table IV-D that the proposed model attains state-of-the-art for the gold dataset.

## V. Conclusion

In this paper, we have proposed an RNN based deep learning framework that aims to reveal and utilize the inter-dependence of textual and image modalities for sarcasm detection. Our proposed approach learns a representation using attention scores to classify sarcasm. We evaluate our proposed approach on the dataset created from the Instagram. Experimental results suggest that using all the modalities with gating yields the best performance. Deciding on the accuracy, we observe that the GloVe produces better representation than lexical, pragmatic and linguistic incongruity based features. Evaluation results show that the image modality plays an important role in providing context and hence helps in building a better sarcasm classifier when augmented with text. In the future, we would like to explore the other dimensions of our framework like predicting sarcasm in conversations using multimodality.

## VI. Acknowledgement

Authors duly acknowledge the support from the Project titled Sevak-An Intelligent Indian Language Chatbot, Sponsored by SERB, Govt. of India (IMP/2018/002072).

## References

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[2] S. Sangwan, D. Chauhan, M. Akhtar, A. Ekbal, and P. Bhattacharyya, *Multi-task Gated Contextual Cross-Modal Attention Framework for Sentiment and Emotion Analysis*, 12 2019, pp. 662–669.

[3] C. Liebrecht, F. Kunneman, and A. van den Bosch, "The perfect solution for detecting sarcasm in tweets #not," in *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 29–37. [Online]. Available: https://www.aclweb.org/anthology/W13-1605

[4] A. Reyes, P. Rosso, and T. Veale, "A multidimensional approach for detecting irony in twitter," *Language resources and evaluation*, vol. 47, no. 1, pp. 239–268, 2013.

[5] E. Shutova, D. Kiela, and J. Maillard, "Black holes and white rabbits: Metaphor identification with visual features," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 160–170.

[6] R. Schifanella, P. de Juan, J. Tetreault, and L. Cao, "Detecting sarcasm in multimodal social platforms," in *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 2016, pp. 1136–1145.

[7] S. K. Bharti, K. S. Babu, and S. K. Jena, "Parsing-based sarcasm sentiment recognition in twitter data," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 2015, pp. 1373–1380.

[8] E. Lunando and A. Purwarianti, "Indonesian social media sentiment analysis with sarcasm detection," in *2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE, 2013, pp. 195–198.

[9] R. González-Ibánez, S. Muresan, and N. Wacholder, "Identifying sarcasm in twitter: a closer look," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*. Association for Computational Linguistics, 2011, pp. 581–586.

[10] D. Bamman and N. A. Smith, "Contextualized sarcasm detection on twitter," in *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*. AAAI Press, 2015, pp. 574–577.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[12] C. C. Park, B. Kim, and G. Kim, "Attend to You: Personalized Image Captioning with Context Sequence Memory Networks," in *CVPR*, 2017.

[13] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[14] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[16] B. L. Welch, "The generalization ofstudent's' problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.

[17] D. Das and A. J. Clark, "Sarcasm Detection on Flickr Using a CNN," in *Proceedings of the 2018 International Conference on Computing and Big Data*. New York, NY, USA: Association for Computing Machinery, 2018, p. 56–61.

[18] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, "Image–text sentiment analysis via deep multimodal attentive fusion," *Knowledge-Based Systems*, vol. 167, pp. 26 – 37, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095070511930019X

978-1-7281-6926-2/20/$31.00 ©2020 IEEE