Multimodal Sarcasm Detection: A Comprehensive Review

Mr. Amit Srivastava*1, Priyanshu Batham*2, Disha Tiwari *3

*1 Assistant Professor, Computer Science, National P.G. College, Lucknow, Uttar Pradesh, India

*2,3 Student, Computer Science, National P.G. College, Lucknow, Uttar Pradesh, India

ABSTRACT

Sarcasm, a form of verbal irony, involves using remarks that intentionally convey the opposite of their literal meaning. It serves to criticize or humorously undermine a situation. Verbal and non-verbal cues—such as changes in tone, incongruence across modalities, and word emphasis—often convey sarcasm. As technology advances, more people express their opinions online, necessitating the development of efficient models for detecting speech nature. Machine Learning algorithms, Linguistic Models, Ensemble Learning, and Multi-modal approaches play crucial roles in this endeavor. While most research has focused on unimodal approaches using textual data, the field now embraces multimodal sarcasm detection. This approach integrates diverse data types, including images, text, and audio, to enhance accuracy and automate sarcasm classification. In our review paper, we delve into the captivating world of multimodal sarcasm detection, tracing its evolution from early unimodal methods to the current state of multimodal techniques. We aim to provide a comprehensive overview, shedding light on challenges, methodologies, and prospects.

Keywords: Sarcasm Detection, Multimodal Approaches, Machine Learning Algorithms, Linguistic Models, Ensemble Learning

INTRODUCTION

Sarcasm plays a significant role in everyday conversations, allowing individuals to mock or express contempt. Achieved through irony, sarcasm often conveys a negative connotation[1]. For instance, consider the utterance: ‘Maybe it’s a good thing we came here. It’s like a lesson in what not to do.’ The speaker explicitly expresses a positive light regarding the lesson, but the underlying meaning is negative. However, there are scenarios where sarcasm lacks explicit linguistic markers, necessitating additional cues. With the rise of social media platforms like Twitter, Reddit, Instagram, and Facebook, individuals increasingly embrace ironic expressions in their posts[2]. Detecting sarcastic or ironic expressions has become crucial for sentiment and opinion mining. Early studies primarily focused on text-only approaches, recognizing sarcasm within textual content. Yet, as multimedia devices advance, people express emotions and opinions through multimodal social posts. Visual content accompanying these posts often provides crucial cues for conveying sarcasm. Researchers now incorporate multimodal cues into sarcasm detection, harnessing both visual and linguistic signals. The seminal work by [3] pioneered the fusion of textual and visual features for the Multimodal Sarcasm Detection (MSD) task. Furthermore, posting sarcastic text or visual content on platforms like WhatsApp, Reddit, Facebook, and Twitter has become a fascinating way to express thoughts indirectly [4] These days, putting sarcastic text or visual content on social networking platforms like WhatsApp, Reddit, Facebook, Twitter, etc., have become a new fascinating style to elude direct pessimism for expressing something [5]. Based on the dialogues in the provided figure, we can discern their sarcastic nature.

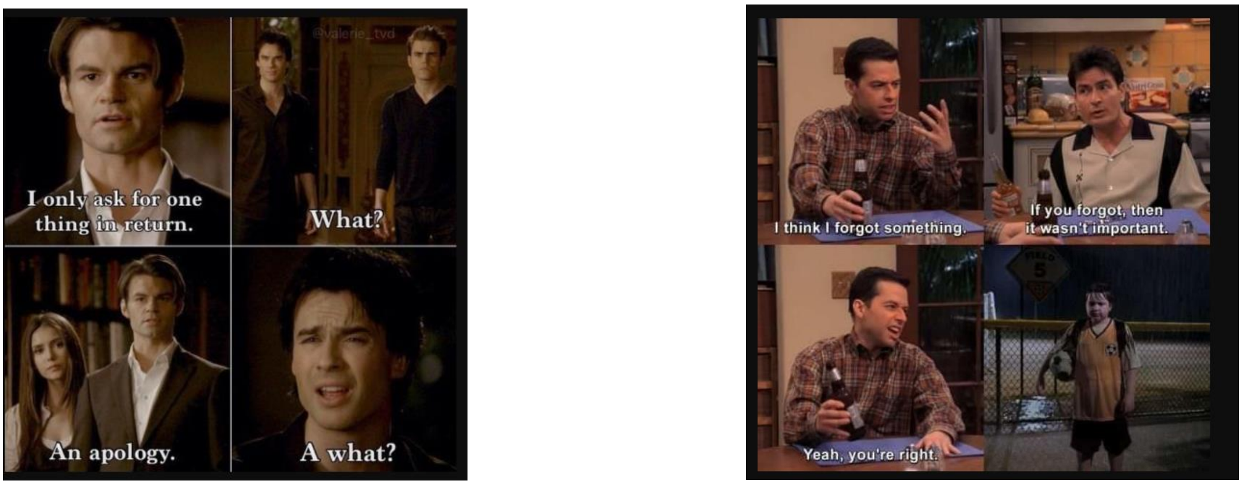

I only ask for one thing in return.
What?
An apology.
A what?
@valerie_tvd
I think I forgot something.
If you forgot, then it wasn't important.
Yeah, you're right.

Figure 1: Sample Sarcastic Utterances from the TV shows [15]

The prevalence of Sarcasm in a multimodal context can include Texts, Visuals, Emojis and gestures, Audio and Tone. Let's explore the sarcasm through different modalities:

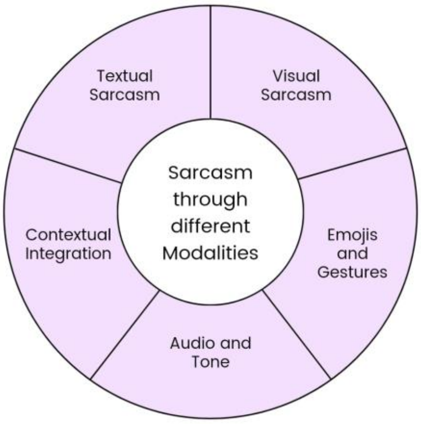


Fig: Sarcasm through different modalities

1. **Textual Sarcasm**:

Traditional sarcasm detection models primarily focused on analyzing textual content. However, with the rise of social media and online platforms, users now combine text with other modalities (such as images, emojis, and GIFs) to convey sarcasm.

Textual sarcasm often includes explicit markers (e.g., irony, hyperbole, or unexpected statements) that can be detected through linguistic analysis.

2. **Visual Sarcasm**:

Multimodal sarcasm detection considers visual cues alongside text. Images, memes, and videos play a crucial role in expressing sarcasm.

Visual sarcasm may involve juxtaposing contradictory elements, altering context, or using absurd imagery to convey hidden meanings.

For example, an image with a smiling face and a caption like “Having a great day!” could be sarcastic if the context suggests otherwise.

3. **Emojis and Gestures:**

Emojis and gestures enhance multimodal communication. A seemingly positive emoji (e.g., 😄) paired with a sarcastic statement can create ambiguity.

Gestures (such as eye rolls or raised eyebrows) also contribute to sarcasm, especially in face-to-face interactions.

4. **Audio and Tone:**

Sarcasm often relies on vocal tone and intonation. In spoken language, the way a sentence is delivered can completely change its meaning.

Multimodal models consider audio cues (speech patterns, pitch, and rhythm) to detect sarcasm.

5. **Contextual Integration:**

Multimodal approaches fuse information from different modalities to improve sarcasm detection accuracy.

Understanding the interplay between text, images, and other cues allows models to capture nuanced meanings.

VARIOUS MODES TO CONVEY SARCASM

1. **Memes and Internet Culture:**

Memes, often combining images and text, are a prime example of multimodal sarcasm. They rely on visual humor and clever captions to convey sarcastic messages.

Internet culture has popularized meme formats that playfully subvert expectations, making them a rich source of sarcasm.

2.**Visual Irony:**

Visual irony occurs when an image contradicts or undermines its literal meaning. For instance, a sign saying “Quiet Zone” placed next to a construction site is visually ironic.

Combining visual cues with text enhances the overall sarcastic effect.

3.**Social Media Platforms:**

Platforms like Instagram and Tik-Tok thrive on visual content. Users create videos, stories, and posts that blend text, images, and audio.

Sarcasm often emerges in these multimedia formats, where tone, expression, and context matter.

4.**Video Content:**

YouTube, Vine (RIP), and other video platforms allow creators to express sarcasm through speech, gestures, and visual storytelling.

The tone of voice, facial expressions, and timing contribute to the overall effect.

5. **Advertising and Marketing:**

Advertisers use visual and textual cues to create ironic or sarcastic campaigns. Clever billboards, print ads, and commercials play with audience expectations.

Subverting traditional advertising tropes can be both humorous and effective.

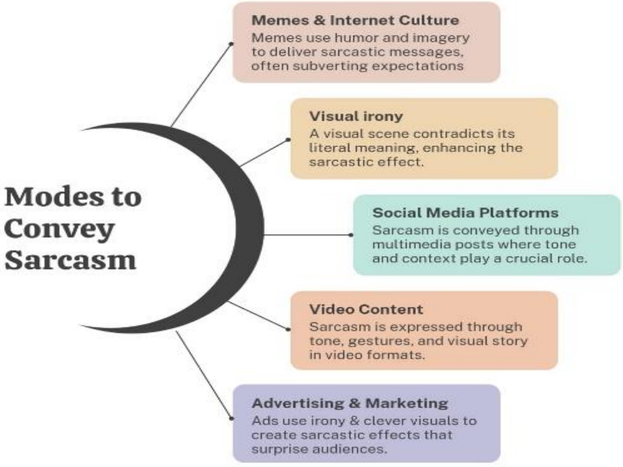

Modes to Convey Sarcasm
Memes & Internet Culture
Memes use humor and imagery to deliver sarcastic messages, often subverting expectations
Visual irony
A visual scene contradicts its literal meaning, enhancing the sarcastic effect.
Social Media Platforms
Sarcasm is conveyed through multimedia posts where tone and context play a crucial role.
Video Content
Sarcasm is expressed through tone, gestures, and visual story in video formats.
Advertising & Marketing
Ads use irony & clever visuals to create sarcastic effects that surprise audiences.

Fig: Different modes to convey sarcasm

In the age of deep learning, there is a strong focus on reducing human involvement while improving linguistic analysis. Deep learning-based approaches dominate sarcasm detection, and a variety of neural methods have been presented for sarcasm detection. [6]

CLASSIFICATION OF SARCASM

According to the researchers in [7], sarcasm is categorized into five distinct types based on its features and structure.

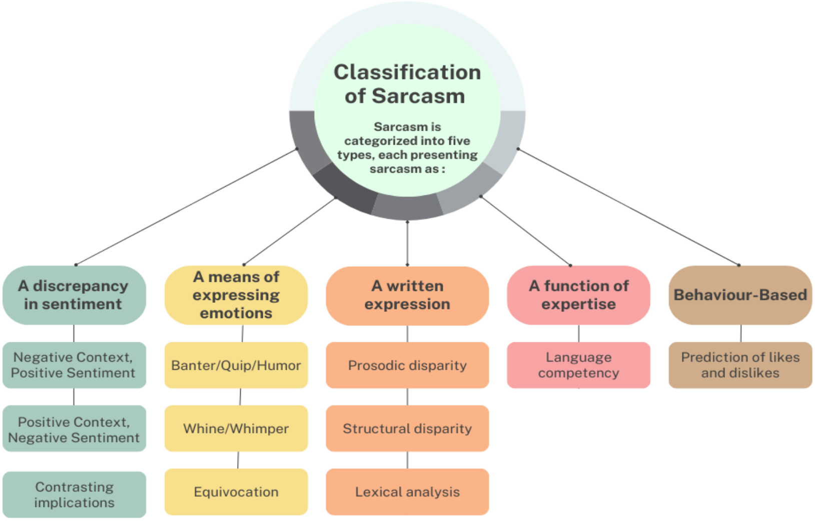

Classification of Sarcasm
Sarcasm is categorized into five types, each presenting sarcasm as :
A discrepancy in sentiment
Negative Context, Positive Sentiment
Positive Context, Negative Sentiment
Contrasting implications
A means of expressing emotions
Banter/Quip/Humor
Whine/Whimper
Equivocation
A written expression
Prosodic disparity
Structural disparity
Lexical analysis
A function of expertise
Language competency
Behaviour-Based
Prediction of likes and dislikes

Fig: Classification of Sarcasm

A. Sarcasm as a Discrepancy in Sentiment

This category of sarcasm involves a mismatch between the sentiment expressed in the text and the situational context:

1. **Discrepancy Between Negative Context and Positive Sentiment** This type involves expressing conflicting opinions within a single statement, where disagreeable ones follow pleasant sentiments. For instance, the statement "Absolutely amazing when my bus is running late" combines positive sentiments with an unpleasant situation.
2. **Discrepancy Between Positive Context and Negative Sentiment** Here, sarcasm is conveyed through negative sentiments in the context of a positive situation. For example, "How I hate Team Australia, they have once again won the T20 world cup!!" communicates negative feelings in the context of a positive outcome.
3. **Contrasting Implications** This type involves presenting contrasting connotations within the same narrative. An example would be "I love receiving spams!" which conveys opposing meanings.

B. Sarcasm as a Means of Expressing Emotion

Sarcasm can also be used to reflect the speaker’s emotional state[8], categorized as follows:

1. **Banter/Quip/Humour** Sarcasm in this category uses capital letters, question marks, and exaggerations to express extreme joy. For example, "What FANTASTIC weather!! I simply LOVE the rains!!" conveys a person's happiness about the rainy weather despite underlying irritation.
2. **Whine/Whimper** This type reflects anger or frustration. For instance, "I am so glad my mom woke me up early in the morning to vacuum my room!! " Clearly shows displeasure about being woken up early.
3. **Equivocation** In this class, sarcasm is used to avoid giving a straightforward answer by employing unusual words or expressions. For example:
   - I: "Alice, you need to work hard!!"
   - Alice: "Ohh Yaah!! I am crystal clear about what needs to be done." This response indicates Alice's avoidance of the task.

C. Sarcasm as a Written Expression

Sarcasm in written form often employs specific stylistic features:

1. **Prosodic Disparity** This involves emphasizing certain parts of the text to convey the opposite of what is meant by repeating letters or punctuation marks. For instance, "sooooo," "wooowww," "!!!!" or capitalizing words, such as "Just WOOWWW!!" or "What an AMAZING weather!!," highlights sarcasm.
2. **Structural Disparity** In this type, the initial part of a sentence presents the user’s opinion, while the context is provided later. For example, "I love it when my friends ignore me" reveals the user’s annoyance through its structure.
3. **Lexical Analysis** Hashtags like #sarcasm or #irony serve as strong indicators of sarcasm. For example, "I just love to party alone. #sarcasm" uses the hashtag to signal sarcasm.

D. Sarcasm as a Function of Expertise

Language Competency Sarcasm in this category relies on the speaker's mastery of language. For example, "He is a good person. #sarcasm" explicitly signals sarcasm, while "He is as good as evil" implies sarcasm without needing a hashtag.

E. Behavior-based Sarcasm

Prediction of Likes and Dislikes Sarcasm is subtly expressed through the manifestation of likes and dislikes towards various products, services, or events. This form of sarcasm may not be directly indicated but is implied through the user's behaviour.

STEPS IN A SARCASTIC DETECTION MODEL

The research article [9] outlines the following steps for detecting sarcasm, which can also be understood from the following figure:

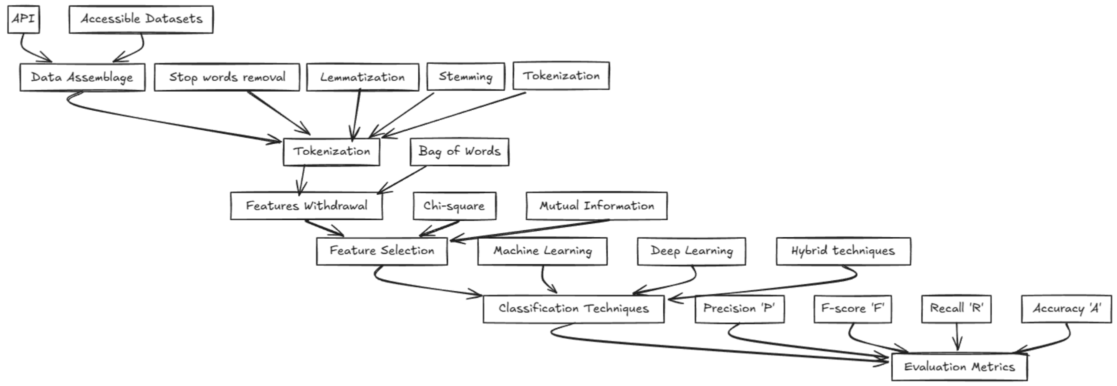


Fig: Steps involved in sarcasm detection by machine learning and natural language processing models.

A. DATA COLLECTION

The initial phase involves acquiring data. This can be done through two primary methods: using APIs (Application Programming Interfaces) or accessing pre-existing datasets such as MUStARD (Multi-modal Sarcasm Detection dataset), SARC (Self-annotated Reddit Corpus), and SemEval (Semantic Evaluation Dataset), among others.

B. TOKENIZATION

The subsequent step is data pre-processing. In Natural Language Processing (NLP) tasks, this phase includes removing stop words, lemmatizing and stemming tokens and performing tokenization.

C. FEATURE EXTRACTION

At this stage, various features are extracted from the dataset to build the model. Feature extraction methods include Bag of Words (BoW), Doc2Vec, Term Frequency-Inverse Document Frequency (TF-IDF), word2vec, and GloVe.

D. FEATURE SELECTION

Here, the most relevant features are selected to improve the performance of the classification model. Common techniques for feature selection include Chi-square and Mutual Information (MI).

E. CLASSIFICATION METHODS

Sarcasm Detection (SD) is approached as a binary classification task using Machine Learning (ML), Deep Learning (DL), or hybrid methods.

F. EVALUATION METRICS

The evaluation of the model involves metrics such as Precision (P), F-score (F), Recall (R), and Accuracy (A). Precision is calculated as the proportion of True Positives out of the total of True Positives and False Positives. The F-score assesses the model's accuracy concerning the dataset. Accuracy is the ratio of True Positives and True Negatives to the total number of cases.

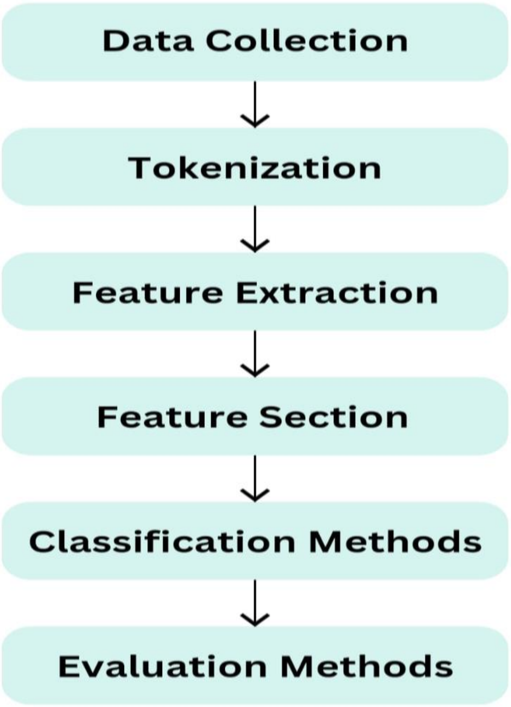


Fig: Generalized flow of data in Machine Learning Algorithms

TECHNIQUES USED IN MULTI-MODAL SARCASM DETECTION

1) Sharmin and collaborators [10] proposed a method that employs an attention-based convolutional neural network (CNN) for sentiment analysis in Bengali. Their model was trained and assessed using a dataset composed of Bengali movie reviews. Previous research highlights the effectiveness of deep learning models, particularly CNNs, for sentiment analysis tasks. Additionally, incorporating attention mechanisms into CNNs has been demonstrated to enhance their performance.

2) Ma and colleagues [11] introduced a novel model named Attentive LSTM with Common Sense Knowledge (ALCSK), which leverages a pre-trained common sense knowledge graph to enhance the aspect embedding layer and improve the accuracy of Targeted Aspect-Based Sentiment Analysis (TABSA) techniques. The model surpasses existing state-of-the-art methods on benchmark datasets, showcasing its effectiveness in managing domain-specific knowledge and integrating common sense data.

3) Hussein and colleagues [12] developed a method for sentiment analysis of Arabic multi-dialect text using machine learning techniques. They compiled a dataset of tweets from various Arabic dialects and pre-processed the data by eliminating stop words, non-alphanumeric characters, and URLs. The tweets were then transformed into numerical vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction technique. The proposed method employed three machine learning classifiers for sentiment analysis: Naive Bayes, Decision Trees, and Support Vector Machines (SVM). The experiments were performed using 10-fold cross-validation, and the results revealed that SVM achieved the highest accuracy at 87.2%, followed by Logistic Regression at 85%, and Stochastic Gradient Descent also at 85%. Additionally, the authors explored different feature selection techniques and determined that the Chi-square feature selection method yielded the best results.

4) Li and colleagues [13] introduced an improved label propagation algorithm (LPA) known as the label attenuation propagation model (LAPM) for automating emoji sentiment analysis. This study highlights the importance of quantifying emoji sentiment and pioneers the application of LPA in this context. The LAPM model is utilized to calculate sentiment uncertainty and analyze emoji sentiment, leveraging an Emoji Link Network (ELN) designed to effectively manage and categorize a wide range of social media emojis. The proposed model achieves an accuracy of 85%, surpassing the 74% accuracy attained by the standard LPA.

5) Sharma and colleagues [14] introduced a hybrid approach for detecting sarcasm in social media. Their model combines sentence embeddings generated from an Autoencoder, Bidirectional Encoder Representations from Transformers (BERT), and the Universal Sentence Encoder (USE) to handle a diverse range of content types. The model was validated using three real-world social media datasets, demonstrating superior accuracy compared to previous state-of-the-art methods. The proposed approach involves pre-processing, pre-training, and classification stages, utilizing three sentence-based techniques to ultimately categorize comments as either sarcastic or non-sarcastic.

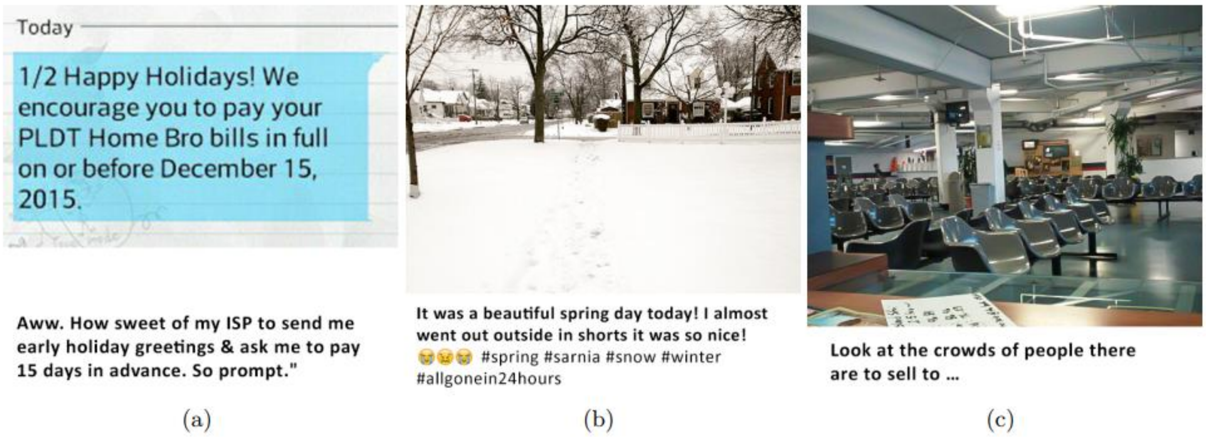
ISSUES IN SARCASM DETECTION

In the field of multimodal sarcasm detection, significant progress has been made in developing models that can effectively analyze and interpret sarcasm from individual modalities, such as text or images. These models have shown high accuracy and reliability when applied to datasets that focus exclusively on either textual content or visual content. For instance, textual datasets allow models to capture nuances in language, tone, and context that are indicative of sarcasm, while image-based datasets enable models to detect visual cues like facial expressions, gestures, or scene context that contribute to sarcastic communication.

However, a major challenge arises when these two modalities—text and images—are combined in a multimodal context. Sarcasm is often conveyed through a complex interplay of textual and visual elements, where the true meaning of a sarcastic statement may only be understood when considering both modalities together. Despite the success of unimodal approaches, current multimodal sarcasm detection models struggle to maintain the same level of accuracy when tasked with analyzing combined datasets that include both text and images.

One reason for this challenge is the inherent complexity of integrating and processing multimodal data. Text and images convey information in fundamentally different ways, requiring models to effectively bridge the gap between natural language processing and computer vision. The contextual dependencies between text and images in sarcastic expressions are often subtle and intricate, making it difficult for models to accurately discern the intended meaning without significant advances in model architecture and training techniques.

Moreover, the lack of high-quality, well-annotated multimodal datasets further exacerbates the problem. While there are numerous datasets available for text-based or image-based sarcasm detection, the availability of comprehensive datasets that combine both modalities is limited. Existing multimodal datasets may suffer from issues such as insufficient size, poor annotation quality, or a lack of diversity in the types of sarcasm represented. These limitations hinder the development of robust multimodal sarcasm detection models, as the models are not exposed to a wide enough variety of examples to learn effectively from the combined modalities.



Today

1/2 Happy Holidays! We encourage you to pay your PLDT Home Bro bills in full on or before December 15, 2015.

Aww. How sweet of my ISP to send me early holiday greetings & ask me to pay 15 days in advance. So prompt."

(a)

It was a beautiful spring day today! I almost went out outside in shorts it was so nice! 😭😠😭 #spring #sarnia #snow #winter #allgonein24hours

(b)

Look at the crowds of people there are to sell to ...

(c)

Fig: Example of dataset format good for multi-modal sarcasm detection.

The scarcity of high-quality multimodal datasets also poses challenges in benchmarking and comparing different models. Without standardized datasets that are widely recognized and used within the research community, it becomes difficult to evaluate the performance of multimodal sarcasm detection models in a consistent and meaningful way. This lack of benchmarking impedes progress in the field, as researchers are unable to reliably measure advancements or identify the most promising approaches.

Future Scope

To address these issues, the field of multimodal sarcasm detection needs to focus on several key areas of improvement. First, there is a need for the development of more sophisticated model architectures that can effectively integrate and process both textual and visual information. This might involve advancements in multimodal fusion techniques, attention mechanisms that can better capture the relationships between modalities, or the use of pre-trained models that have been exposed to a wide range of multimodal data.

Second, the creation of high-quality, diverse, and well-annotated multimodal datasets is crucial. These datasets should reflect the varied ways in which sarcasm is expressed across different cultures, contexts, and communication styles, and should be large enough to support the training of complex models.

CONCLUSION

This paper has presented an overview of the sarcasm detection process, highlighting the shift from monomodal architectures, where only text, image, or audio was utilized, to modern multimodal approaches that incorporate both text, images, and/or audio. While unimodal models have achieved high levels of accuracy within their specific domains, they struggle in multimodal scenarios where understanding the interplay between text and images is crucial. This challenge arises from the inherent difficulty in effectively combining different modalities and capturing the nuanced interactions between them, which remains largely uncharted territory for most current models. Additionally, the scarcity of large, diverse, and well-annotated multimodal datasets exacerbates these challenges, hindering the development of more effective models. To address these issues, future work should focus on designing sophisticated models capable of seamlessly integrating multimodal data and constructing extensive, heterogeneous, and meticulously annotated datasets. Moreover, exploring advanced techniques such as cross-modal attention mechanisms or transfer learning between modalities could pave the way for significant advancements. These efforts are essential for enhancing the performance of sarcasm detection systems in practical applications.

REFERENCES

[1] https://www.emerald.com/insight/content/doi/10.1108/IDD-01-2023-0002/full/html

[2] https://arxiv.org/abs/1608.02289

[3] Schifanella et al. (2016)

[4] (Agrawal & An, 2018)

[5] (Agrawal, A., & An, A. ,2018)

[6] A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities Wangqun Chen a b, Fuqiang Lin a b, Guowei Li a b, Bo Liu a c

[7] . Chaudhari and C. Chandankhede, “Literature survey of sarcasm detection,” in 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017, pp. 2041–2046.

[8] Towards Multimodal Sarcasm Detection (An Obviously Perfect Paper)” by Santiago Castro et al.https://www.mdpi.com/2079-9292/13/5/855.

[9] P. Verma, N. Shukla, and A. Shukla, “Techniques of sarcasm detection: A review,” in 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2021, pp. 968–972.

[10] S. Sharmin, D. Chakma, "Attention-based convolu- tional neural network for Bangla sentiment analy- sis", AI & Society, Vol. 36, No. 1, 2021, pp. 381-396

[11] Y. Ma, H. Peng, E. Cambria, "Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM", Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, No. 1, 2018.

[12] A. Hussein, I. Moawad, R. Badry, "Arabic Sentiment Analysis for Multi-dialect Text using Machine Learning Techniques", International Journal of Advanced Computer Science and Applications, Vol. 12, No. 12, 2021.

[13] D. Li, X. Luo, X. Wei, R. Xue, "Emojis Sentiment Anal- ysis Based on Big Social Media Data", Proceedings of the International Conference on Applications and Techniques in Cyber Security and Intelligence, Shanghai, China, 11-13 July 2018, pp. 56-63.

[14] D K. Sharma, B. Singh, S. Agarwal, H. Kim, R. Shar- ma, "Sarcasm Detection over Social Media Plat- forms Using Hybrid Auto-Encoder-Based Model", Electronics, Vol. 11, No. 18, 2022.

[15] https://en.wikipedia.org/wiki/The_Vampire_Diaries