INTENSE

Use Intent, Emotion and Sentiment to make INSTAGRAM Intelligent

(GROUP NO - 8)

Akshara Nair - MT22008
Ayush Agarwal - MT22095
Medha - MT22110
Mohit Gupta - MT22112
Nidhi Verma - MT22044
Pratik Chauhan - MT22118

Overview

The proposed idea explores and utilises the idea of extracting information through multimodal data and utilise the features and capabilities of distinct modalities to serve a strong prediction for the sentiment, emotion and intent being conveyed as predicting intent through multimodal data in specifically INSTAGRAM posts does introduce ambiguities.

- Problem Statement
- Motivation & Challenges
- Related Work & Limitations
- Proposed Solution
  - Solution
  - Novelty (Dataset and Procedure)
  - Impact

Problem Statement

With a massive amount of visual, audial and textual content being posted on social media platforms as Instagram, there is a need to monitor the intent of the author as the platform offers a space for a diverse and varied range of the emotion and sentiment being conveyed, with techniques existing to predict and analyze both.. **To infer the intent of the author, through the content being posted, presents itself as a challenging task as there exists a high probability for the fact that the caption, the image or visual content and the audial notes, for a piece of content being posted, might be conveying distinct emotion, sentiment and intent.** Hence we propose to utilize the concept of unified multimodal sentiment analysis and emotion recognition, combined with an effort to predict the intent of the author to resolve ambiguities through multiple modalities while predicting.

MOTIVATION

- An effort to make Instagram **smart and more intuitively controlled** which adapts to unique behaviour and preference of each user and hence provide a personalized experience and enhance their daily lives by adding utility to entertainment and relaxation.
- To avoid mindless scrolling of content on Instagram in form of social media might trigger a user under certain environment. **It helps user to monitor the nature of the content being consumed.**
- The proposed idea of **unifying intent, emotion and sentiment**, supports the purpose and the motivation behind analyzing the feed being supplied through the author end and consumed at the user end.

CHALLENGES

- As this involves collection and analysis of data from Instagram users, it must be transparent about its practices and **prioritize user privacy and security**.
- As idea of our paper is to determine the intent, which needs an algorithm which accurately determines the author intent and emotions is also a challenging task.
- Most of the sensitive information about the user is private as users keep their instagram account private but we need those datasets also to train our algorithm so that it perform better.

Related Works:

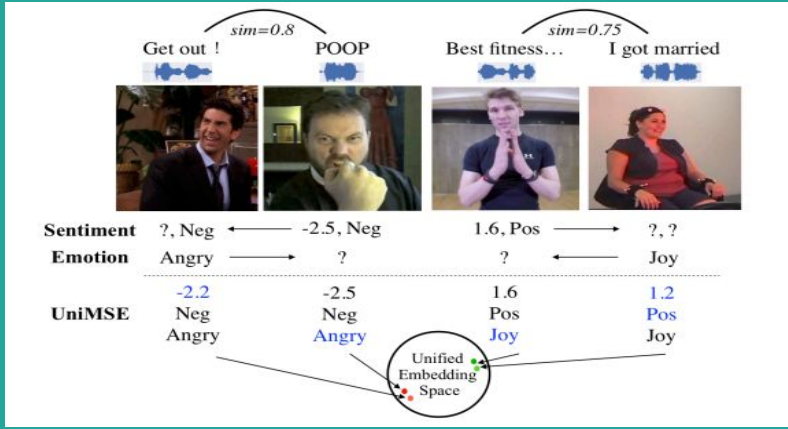
- Prior work has primarily focused on unimodal approaches that rely on a single modality, such as text-based or image-based analysis. Also, most of the approaches model behaviours like emotion, sentiment, affect etc in isolation without focusing on their underlying similarities & differences [1].
- Multimodal sentiment analysis approaches can be broadly categorized into four types: **multimodal fusion, modal consistency and multi-task joint learning, multimodal alignment, and multimodal context integration with unimodal context**.
- Multimodal emotion recognition in conversation can be broadly categorized into three types: **multimodal fusion, context-aware models and incorporating external knowledge**.
- Joshi et al. (2022)[2] adopt graph neural networks to model the inter/intra dependencies of utterances or speakers.
- Transfer learning, common sense knowledge (Ghosal et al., 2020)[3], multi-task learning (Akhtar et al., 2019)[4], and external information is also used to solve ERC tasks.
- In recent years, the unification of related but different tasks into a framework has achieved significant progress. Zhang et al. (2021c)[5] proposes a unified framework for multimodal summarization, Wang et al. (2021b)[6] unifies entity detection and relation classification on their label space to eliminate the different treatment, and Yan et al. (2021b)[7] integrates the flat NER, nested NER, and discontinuous NER subtasks in a unified framework.

LIMITATIONS

1. The existing literature work and reviews do not fully exploit the similarities and contrasts between the emotion, sentiment and intent.

2. In existing works (UniMSE etc)[1], even when emotion and sentiment are learnt as a unified label, only textual modality are utilized in generation of universal labels, without considering the acoustic and visual modalities.

3. Each modality is indicative of a distinct cue to predict the intent, which might be very different especially in the case of Instagram, hence it's necessary to combine the modalities to make an effective prediction.



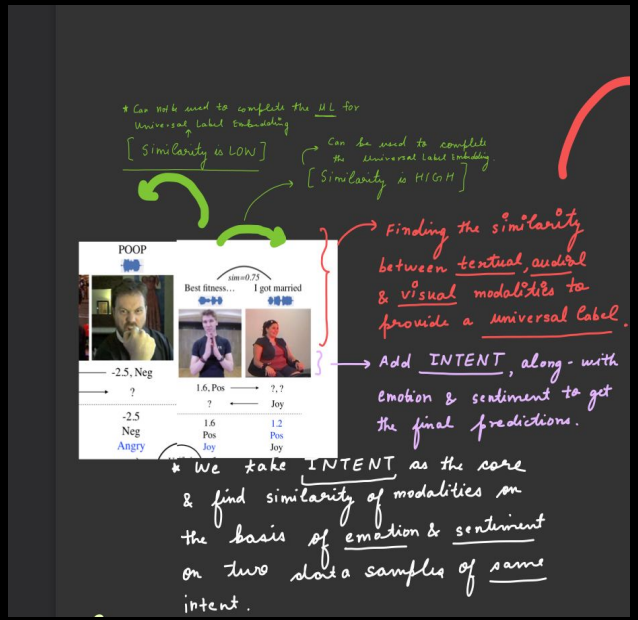
Proposed Solution

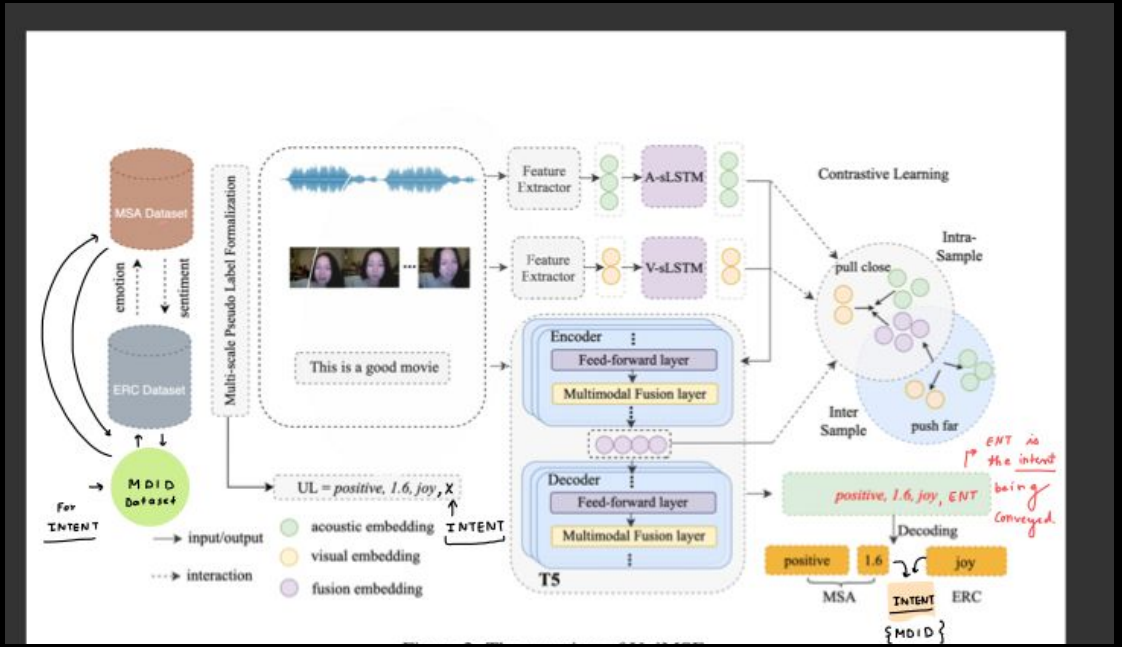
INTENSE is a modified version of UniMSE, which exploits the similarities and contrasts in the multimodal sentiment analysis and emotion recognition model as each individual modality provides distinct cues. INTENSE extends UniMSE by building on the task to recognize intent as a task. The model processes a multimodal signal containing raw unimodal sequences of text, acoustic, and visual data. The input features and label space of MSA, ERC, and intent recognition are unified through input and label formalization. INTENSE formalizes MSA, ERC, and intent recognition as a unified task through a single architecture involving two LSTMs for Audial and Visual Inputs and T5 Transformer for text.

Contrastive Learning - It aims to narrow the gap between modalities of the same sample and push the modality representation of different samples further apart. To reduce information loss during **multimodal fusion**, we introduce a **contrastive learning** task to capture **semantically** consistent representations.

Contrastive Learning minimizes the intra-class variance and maximizes the inter class variance across categories.

Proposed Model Pipeline





NOVELTY

MDID - Integrating Text and Image: Determining Multimodal Document Intent in Instagram Posts[4]

The Dataset being incorporated to get the overall prediction through the pipeline has not been widely utilised as the published baseline implements basic CNN and NLP techniques to get the results. The proposed approach eliminates ambiguities while predicting the sentiment, emotion and intent for personal posts which has been mentioned as a key limitation of the baseline results on the published work. We propose to extend the dataset by adding the short videos being posted online and labeling the intent being conveyed. The six intent categories are as : promotion, information, personal, commentary, meme, and event.

4 - Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating Text and Image: Determining Multimodal Document Intent in Instagram Posts.(EMNLP-IJCNLP), pages 4622–4632, Hong Kong, China. Association for Computational Linguistics.

Novelty in the Model Architecture Pipeline

- To create **Universal Label (UL)** format for the dataset, in addition to MSA and ERC, the intent of the post is added to exploit the knowledge shared between MSA and ERC and intent through all three modalities.
- The basis of similarity is to check the similarity of **emotion and sentiment** between two posts which share the **same intent** and create a **unified embedding space** on the basis of a **threshold** value for the same for all three modalities, if they align.
- The process will add to strongly support the prediction of intent through shared semantic knowledge.

Impact

- Potential for mental health benefits: By analyzing user emotion and sentiment, the application can potentially identify users who may be struggling with mental health issues and provide resources and support and provide control to monitor minors through adult supervision and control.
- Behaviour Analysis: The proposed idea, if deployed as an application, will offer the user to monitor the individual's content consumption and restrict or allow certain specific domains of information.
- Improved user experience: By incorporating intent, emotion, and sentiment analysis, the proposed technology offers a personalized user experience by showing content that is more relevant and meaningful to the user. This could lead to increased enhanced engagement and user satisfaction. Whatever a user scrolls through would add to the individual's interest and can be monitored and controlled by the individual.
- Enhanced advertising capabilities: With better understanding of user intent and sentiment, the application can significantly control their advertising targeting capabilities, resulting in more effective and relevant ads for users as well as authors and an increased revenue for the people utilizing the platform as a source of income.

CONCLUSION

- The proposed idea involves prior modifications on the scraped data as the dataset requires additional annotated dataset for another modality.
- The idea to explore the similarity between each category of intent with certain emotion or sentiment might or might not result in a significant extraction of sentiment and emotion for a specific intent necessarily.
- The curation of a manually labelled dataset for a wide variety of posts will support the prediction of intent. However, limited data serves as a limitation of the current work, offering a magnified future scope for the same.

REFERENCES

1. Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7837–7851, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
2. Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Vikram Singh, and Ashutosh Modi. 2022. COGMEN: contextualized GNN based multimodal emotion recognition. CoRR, abs/2205.02455.
3. Deepanway Ghosal, Navonil Majumder, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: commonsense knowledge for emotion identification in conversations. In Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020, pages 2470–2481.
4. Md. Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 370–379. Association for Computational Linguistics.
5. Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2021c. Unims: A unified framework for multimodal summarization with knowledge distillation. CoRR, abs/2109.05812.
6. Yijun Wang, Changzhi Sun, Yuanbin Wu, Hao Zhou, Lei Li, and Junchi Yan. 2021b. Unire: A unified label space for entity relation extraction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 220–231.
7. Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021b. A unified generative framework for various NER subtasks. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 5808–5