Sentiment Analysis on Video Transcripts: Comparing the Value of Textual and Multimodal Annotations

Quanqi Du, Loic De Langhe, Els Lefever, Véronique Hoste
LT3, Language and Translation Technology Team, Ghent University, Belgium
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
firstname.lastname@ugent.be

Abstract

This study explores the differences between textual and multimodal sentiment annotations on videos and their impact on transcript-based sentiment modelling. Using the UniC and CH-SIMS datasets which are annotated at both the unimodal and multimodal level, we conducted a statistical analysis and sentiment modelling experiments. Results reveal significant differences between the two annotation types, with textual annotations yielding better performance in sentiment modelling and demonstrating superior generalization ability. These findings highlight the challenges of cross-modality generalization and provide insights for advancing sentiment analysis.

1 Introduction

With the rise of the internet and online platforms, especially the proliferation of social media, user-generated content (UGC) has become widely accessible to the public. UGC appears in various forms and modalities, ranging from online textual movie reviews on platforms like IMDB[1] and Rotten Tomatoes[2], to video blogs (vlogs) in video-sharing platforms such as YouTube[3] and TikTok[4].

UGC holds significant value for companies, marketers and politicians (Van Hee et al., 2014), as it contains sentiment-rich information that can be leveraged to monitor public opinion and support the decision-making process (Wankhade et al., 2022). For instance, sentiment analysis based on tweets has been utilized to model user satisfaction in mobile payments (Kar, 2021) and predict election outcomes (Stefanov et al., 2020).

Sentiment analysis on UGC predominantly focuses on text, partly because textual sentiment modelling is more developed and computationally efficient compared to other modalities, such as audio and video. In contrast, systems capable of automatically understanding the content and sentiment of videos are still in their infancy (Stappen et al., 2021; Wang et al., 2023). Consequently, sentiment analysis of non-textual UGC is often converted to text-based analysis through subtitles or transcripts. For instance, Stappen et al. (2021) investigated the use of video transcripts to capture contextual and emotional information in videos.

A critical question arises when annotating transcripts: Should information from non-textual modalities be considered during annotation? In real-life scenarios, sentiment annotations typically reflect the emotional status across modalities. As a result, some studies incorporate multimodal information into the final annotation (Morency et al., 2011; Pérez-Rosas et al., 2013; Nguyen-The et al., 2022). However, another common approach is to perform annotation solely based on textual information, excluding other modalities to avoid interference (Clavel et al., 2013; Stappen et al., 2021; Bekmanova et al., 2022; Efat et al., 2023). This approach is practical since the input for sentiment modelling is usually text, and annotating textual data is less complex compared to multimodal data.

Both approaches to sentiment annotation have their merits and are often intertwined. In some cases, researchers do not differentiate between them, applying multimodal annotations to textual sentiment modelling under the assumption that sentiment labels across modalities are consistent. However, this assumption does not always hold true. For instance, the text “I love this weather” might be labelled as positive, but when the speaker’s tone is sarcastic and they wear a frown, the sentiment might be perceived as negative. Previous studies have shown that emotion labels in multimodal setups do not always align with those derived from textual modalities alone (Ellis et al., 2014; Yu et al., 2020; Du et al., 2023, 2024).

[1] https://www.imdb.com

[2] https://www.rottentomatoes.com

[3] https://www.youtube.com

[4] https://www.tiktok.com

Proceedings of the Tenth Workshop on Noisy and User-generated Text, pages 10–15
May 3, 2025 ©2025 Association for Computational Linguistics

Accurate annotations are crucial for building effective models. However, in the field of sentiment analysis on UGC transcripts, few studies have compared sentiment annotations derived solely from textual information with those incorporating multimodal information, or examined the impact of these differences on sentiment modelling. To address this gap, this paper seeks to answer the following research questions:

1. Do sentiment annotations on video transcripts based solely on textual information differ from those that include information from other modalities? If so, to what extent?

2. How does the inclusion or exclusion of non-textual information in video transcript annotations impact sentiment modelling?

2 Related Studies

A significant portion of sentiment analysis research has traditionally relied on datasets comprising user-generated text. Common sources include social media platforms, such as tweets (Gyanendro Singh et al., 2020), and reviews from domains like products, hotels, and movies (Van et al., 2022; Thakkar et al., 2023). While these studies have provided valuable insights into sentiment classification, they are predominantly focused on textual data.

Recently, sentiment analysis has evolved beyond textual analysis to incorporate other modalities, such as audio and video, giving rise to multimodal sentiment analysis (Wu et al., 2024). This shift reflects the growing prevalence of opinion-sharing in video formats on platforms like YouTube and TikTok (Zadeh et al., 2017; Gandhi et al., 2023), where diverse modalities provide richer contextual information for understanding sentiments.

An essential aspect of multimodal sentiment analysis is the fusion of different modalities (Gandhi et al., 2023; Zhu et al., 2023). Fusion strategies are broadly categorized into two types: early fusion and late fusion. Early fusion, also known as feature-level fusion, integrates features from each modality at the input level, whereas late fusion, or decision-level fusion, combines the outputs of unimodal sentiment analyses to generate the final prediction. Recently, advanced fusion approaches, such as tensor fusion networks (Yan et al., 2022) and dynamic fusion methods (Hu et al., 2022a), have been proposed to enhance performance.

While incorporating non-textual information generally improves the performance of multimodal sentiment analysis, there remains a heavy reliance on textual modalities. This phenomenon, termed *text-predominance*, is evident in studies showing a significant drop in classification accuracy – from approximately 80% to 54% – when textual information is excluded from multimodal models trained on multimodal data (Liu et al., 2022). In contrast, removing audio or visual information results in only a marginal accuracy decline, such as a reduction from 87% to 85% (Hu et al., 2022b), a trend corroborated by Wu et al. (2024).

It seems that we can still rely on textual information despite the availability of other modalities, especially when considering the imbalance of the cost and the improvement when introducing non-textual modalities. However, when we decide to take into consideration only the textual modality of the opinioned videos, which set of annotations should be used, the textual one or the multimodal one, as multimodal labels do not always reflect sentimental states in texts (Yu et al., 2020; Du et al., 2024). In the following, we are going to investigate the differences and influence of the two sets of sentiment annotations.

3 Datasets

The definition of UGC varies across disciplines. In the context of social media, UGC is defined as *any kind of text, data or action performed by online digital systems users, published and disseminated by the same user through independent channels, that incur an expressive or communicative effect either on an individual manner or combined with other contributions from the same or other sources* (Santos, 2022). Based on this definition, we selected two datasets for our study: the UGC dataset UniC (Du et al., 2024), and the non-UGC dataset CH-SIMS (Yu et al., 2020).

UniC is an English audio-visual emotion dataset with independent annotations for each modality (i.e., text, audio, and silent video) as well as overall emotion states of the videos. This UGC dataset comprises nearly 1,000 video clips collected from YouTube, focusing on the topic of *reviews*.

CH-SIMS is a Chinese multimodal sentiment analysis dataset featuring over 2,000 curated video segments with both multimodal and independent unimodal annotations. The videos in CH-SIMS are sourced from movies, TV series, and variety shows, implying that the professional actors in CH-SIMS tend to express emotions more explicitly in all modalities than the non-professionals in UniC. This difference may also influence the sentiment annotations across modalities.

For both datasets, sentiment labels were originally designed as negative, weakly negative, neutral, weakly positive and positive. In this paper, we grouped weakly negative and weakly positive into negative and positive, respectively, for further experiments and analysis.

4 Experiment

4.1 Statistical Analysis

We first analyzed the sentiment distribution across modality setups. As shown in Figure 1, the sentiment distributions in both datasets vary depending on the modality setup. Compared to the multimodal setup, the number of both negative and positive instances decreases in the textual modality, while the number of neutral instances increases. A possible explanation for this trend is that the additional information from audio and visual modalities helps annotators discern sentiment polarities that might otherwise be interpreted as neutral in text-only expressions.

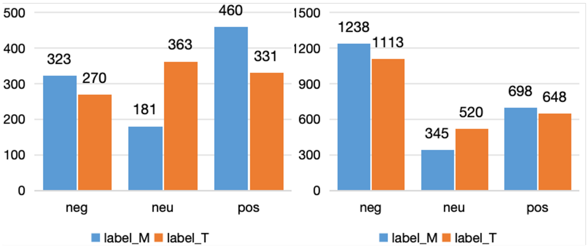


Figure 1: Distribution of textual and multimodal annotations in Unic (left) and CH-SIMS (right).

To evaluate the relationship between annotations across modalities, we conducted significance tests. Chi-Square test results indicate that the relationship between the two types of annotations is statistically significant and not random, with a P-value of 2.49e-98 for UniC and a P-value of 6.80e-235 for CH-SIMS.

To further explore the similarities between the textual and multimodal annotations, we compared the two annotation types. In UniC, 63.69% of instances were assigned the same sentiment annotations across the two modality setups, while in CH-SIMS, this percentage increased to 69.09%. To measure the agreement between the two annotation sets, Cohen's kappa coefficient (Cohen, 1960) was applied. The results show a higher level of agreement in CH-SIMS, with a kappa score of 0.5494, compared to 0.4964 in UniC. These findings highlight a notable difference between the two sets of sentiment annotations, suggesting that the distinctions are not negligible.

The confusion matrices for the two annotation types in both datasets, presented in Figure 2, provide further insights into how sentiment labels change when transitioning between modalities. For example, the inclusion of audio-visual information in UniC led to a shift of approximately 30% of negative and 12% of positive annotations from their textual counterparts. This discrepancy is exemplified in Figure 3, where a video clip is annotated as negative in the text but positive in the multimodal context. The sentiment shift primarily arises from the cheerful tone of voice and the presence of a smile. In contrast, for CH-SIMS, the corresponding shifts were about 15% and 28%, respectively. These results demonstrate the varied impact of multimodal information on sentiment annotations across the two datasets.

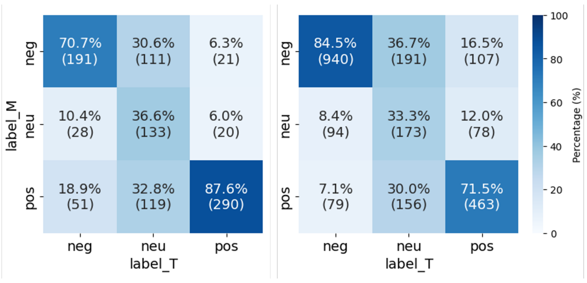
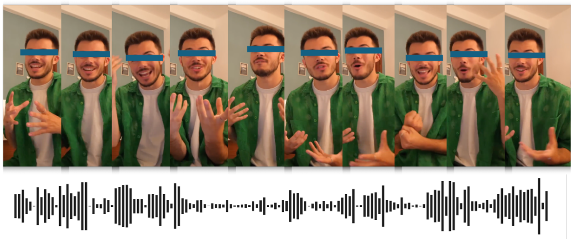


Figure 2: Confusion matrix (Left: UniC; Right: CH-SIMS) of textual and multimodal annotations. The frequency is normalized vertically against the number of textual annotations with different sentiment labels.



but like, in a good way, not shockingly bad, shockingly absurd. I experienced a really visceral and physical response to it. Like, it was making my whole body tense and cringe by how wild it is, and also quite disgusting at times.

Figure 3: A video clip example from UniC.

	Acc-text		**F1-text**		**Acc-mm**		**F1-mm**	
Training data	mean	SD	mean	SD	mean	SD	mean	SD
UniC-text	74.57	1.57	**74.62**	1.66	54.98	2.14	53.03	1.69
CH-SIMS-text	71.18	0.44	**68.33**	0.95	59.36	1.27	52.60	0.29
UniC-mm	44.67	1.19	38.66	1.36	58.76	1.03	**43.38**	0.86
CH-SIMS-mm	60.70	0.87	**46.15**	2.07	63.03	0.91	45.13	1.22

Table 1: Model performances on test datasets when fine-tuned with textual (text) and multimodal (mm) annotations, respectively, and evaluated against textual (text) and multimodal (mm) annotations, respectively, from UniC and CH-SIMS. Accuracy (ACC) and F1-Macro (F1) are averaged from the results of three experiments. SD stands for standard deviation.

4.2 Sentiment Modelling

To further examine the differences between the two types of sentiment annotation, we applied both in the task of transcript-based sentiment modelling by fine-tuning a RoBERTa-base model (Liu et al., 2019) for UniC and a Chinese RoBERTa model (Cui et al., 2021) for CH-SIMS, respectively. Specifically, all instances from both datasets were shuffled and randomly split in the training, validation and test sets in an 8:1:1 ratio. The models were fine-tuned using a learning rate of 1e-5, and a batch size of 8, and 10 epochs with an early-stopping strategy.

For evaluation, both accuracy and macro F1 scores were used to assess performance across textual and multimodal annotations, providing insights into cross-modality performance and the generalization potential between modality setups. Each fine-tuning experiment was repeated three times with different random seeds, and the averaged results are presented in Table 1.

As shown in Table 1, for both UniC and CH-SIMS, the models fine-tuned with textual annotations performed better when evaluated against textual annotations than against multimodal annotations. This highlights barriers across modalities and significant information loss when transitioning from multimodal data to a single modality for both datasets. Interestingly, while the model performed significantly better on textual annotations from UniC (F1 = 74.57) compared to CH-SIMS (F1 = 68.33), the performance gap narrowed when evaluated against multimodal annotations (F1 = 53.03 for UniC versus F1 = 52.60 for CH-SIMS). This suggests a common limitation in the model's ability to generalize from text to multimodality across both datasets.

The scenario became more complex when multimodal annotations were used for fine-tuning. For both UniC and CH-SIMS, models fine-tuned with multimodal annotations achieved only moderate performance (F1 = 42.38 for UniC and F1 = 45.13 for CH-SIMS), reflecting the limitations of text-based language models in generalizing from textual to multimodality setups. Additionally, the models' performance varied when evaluated against multimodal annotations versus textual annotations. For UniC, the F1 score dropped noticeably from 43.38 to 38.66, while CH-SIMS showed a marginal increase, with the F1 score rising from 45.13 to 46.15. This indicates differing capacities of multimodal annotations to encapsulate information relevant to textual annotations.

More notably, when comparing evaluations against multimodal annotations, models fine-tuned with textual annotations generally outperformed those fine-tuned with multimodal annotations for both datasets. This finding suggests the sentiment generalization ability of textual annotations in text-based language models.

5 Conclusion

This study investigated the differences between sentiment annotations on video transcripts derived from textual and multimodal setups, as well as their impact on transcript-based sentiment modelling.

The statistical analysis revealed a significant difference between the two types of sentiment annotations with absolute similarities less than 70% and kappa scores less than 0.55, highlighting the influence of multimodal information on sentiment labelling in video data. The modelling experiments further demonstrated that text-based annotations outperformed multimodal annotations when evaluated against both textual and multimodal labels. Also, a significant cross-modality performance gap was observed. For instance, the macro F1 score dropped from 74.62 to 53.03 when the evaluation labels shifted from text-based to multimodality for UniC, underscoring the challenges of generalizing sentiment models across different modalities.

For future research, we will investigate the incorporation of additional modalities (e.g., audio and facial expressions) and advanced models (e.g., multimodal fusion models), enabling a more comprehensive and nuanced analysis.

6 Limitations

A notable limitation of this study is the linguistic difference between the datasets: UniC is in English, while CH-SIMS is in Chinese. As a result, the comparison between UGC and non-UGC may be influenced by cross-cultural differences, which were not explicitly addressed in this research. Future studies should consider incorporating datasets from the same linguistic and cultural context to allow for stronger and more nuanced comparisons. Unfortunately, the current availability of datasets limits the feasibility of such an approach.

7 Acknowledgments

This research received funding from the Flemish Government under the Research Program Artificial Intelligence (174K02325). We would also like to thank the anonymous reviewers for their valuable and constructive feedback.

References

Gulmira Bekmanova, Banu Yergesh, Altynbek Sharipbay, and Assel Mukanova. 2022. Emotional speech recognition method based on word transcription. *Sensors*, 22(5):1937.

Chloé Clavel, Gilles Adda, Frederik Cailliau, Martine Garnier-Rizet, Ariane Cavet, Géraldine Chapuis, Sandrine Courcinous, Charlotte Danesi, Anne-Laure Daquo, Myrtille Deldossi, et al. 2013. Spontaneous speech and opinion detection: Mining call-centre transcripts. *Language Resources and Evaluation*, 47:1089–1125.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Quanqi Du, Sofie Labat, Thomas Demeester, and Veronique Hoste. 2023. Unimodalities count as perspectives in multimodal emotion annotation. In *Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP co-located with 26th European Conference on Artificial Intelligence (ECAI 2023)*. CEUR Workshop Proceedings.

Quanqi Du, Sofie Labat, Thomas Demeester, and Veronique Hoste. 2024. UniC: A dataset for emotion analysis of videos with multimodal and unimodal labels. *Research Square preprint doi.org/10.21203/rs.3.rs-4443808/v1*.

Azher Ahmed Efat, Asif Atiq, Abrar Shahriar Abeed, Armanul Momin, and Md Golam Rabiul Alam. 2023. Empoliticon: NLP and ML based approach for context and emotion classification of political speeches from transcripts. *IEEE Access*, 11:54808–54821.

Joseph G Ellis, Brendan Jou, and Shih-Fu Chang. 2014. Why we watch the news: A dataset for exploring sentiment in broadcast video news. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 104–111.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Loitongbam Gyanendro Singh, Anasua Mitra, and Sanasam Ranbir Singh. 2020. Sentiment analysis of tweets using heterogeneous multi-layer network representation and embedding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8932–8946, Online. Association for Computational Linguistics.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022a. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7037–7041. IEEE.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022b. 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.

Arpan Kumar Kar. 2021. What affects usage satisfaction in mobile payments? Modelling user generated content to develop the "digital service usage satisfaction model". *Information Systems Frontiers*, 23(5):1341–1361.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. 2022. Make acoustic and visual cues matter: Ch-SIMS v2. 0 dataset and AV-Mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction*, pages 247–258.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, pages 169–176.

Maude Nguyen-The, Soufiane Lamghari, Guillaume-Alexandre Bilodeau, and Jan Rockemann. 2022. Leveraging sentiment analysis knowledge to solve emotion detection tasks. In *International Conference on Pattern Recognition*, pages 405–416. Springer.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982.

Marcelo Luis Barbosa dos Santos. 2022. The "so-called" UGC: An updated definition of user-generated content in the age of social media. *Online Information Review*, 46(1):95–113.

Lukas Stappen, Alice Baird, Erik Cambria, and Björn W Schuller. 2021. Sentiment analysis and topic recognition in video transcriptions. *IEEE Intelligent Systems*, 36(2):88–95.

Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 527–537, Online. Association for Computational Linguistics.

Gaurish Thakkar, Nives Mikelic Preradovic, and Marko Tadić. 2023. Croatian film review dataset (cro-FiReDa): A sentiment annotated dataset of film reviews. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 25–31, Dubrovnik, Croatia. Association for Computational Linguistics.

Thin Dang Van, Hao Duong Ngoc, and Ngan Nguyen Luu-Thuy. 2022. Sentiment analysis in code-mixed Vietnamese-English sentence-level hotel reviews. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 54–61, Manila, Philippines. Association for Computational Linguistics.

Cynthia Van Hee, Marjan Van de Kauter, Orphée De Clercq, Els Lefever, and Véronique Hoste. 2014. LT3: Sentiment classification in user-generated content using a rich feature set. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 406–410, Dublin, Ireland. Association for Computational Linguistics.

James Z Wang, Sicheng Zhao, Chenyan Wu, Reginald B Adams, Michelle G Newman, Tal Shafir, and Rachelle Tsachor. 2023. Unlocking the emotional world of visual media: An overview of the science, research, and impact of understanding emotion. *Proceedings of the IEEE*, 111(10):1236–1286.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Zehui Wu, Ziwei Gong, Jaywon Koo, and Julia Hirschberg. 2024. Multimodal multi-loss fusion network for sentiment analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3588–3602, Mexico City, Mexico. Association for Computational Linguistics.

Xueming Yan, Haiwei Xue, Shengyi Jiang, and Ziang Liu. 2022. Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modeling. *Applied Artificial Intelligence*, 36(1):2000688.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online. Association for Computational Linguistics.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xiangjie Kong. 2023. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, 95:306–325.