

# Topic and Style-aware Transformer for Multimodal Emotion Recognition

Shuwen Qiu<sup>1</sup>   Nitesh Sekhar<sup>2</sup>   Prateek Singhal<sup>2</sup>  
s.qiu@ucla.edu   seknites@amazon.com   prtksngh@amazon.com  
<sup>1</sup>University of California, Los Angeles   <sup>2</sup>Amazon

## Abstract

Understanding emotion expressions in multimodal signals is key for machines to have a better understanding of human communication. While language, visual and acoustic modalities can provide clues from different perspectives, the visual modality is shown to make minimal contribution to the performance in the emotion recognition field due to its high dimensionality. Therefore, we first leverage the strong multimodality backbone VATT to project the visual signal to the common space with language and acoustic signals. Also, we propose content-oriented features Topic and Speaking style on top of it to approach the subjectivity issues. Experiments conducted on the benchmark dataset MOSEI show our model can outperform SOTA results and effectively incorporate visual signals and handle subjectivity issues by serving as content "normalization".

## 1 Introduction

Emotion recognition is intrinsic for social robots to interact with people naturally. The ability to tell emotional change and propose timely intervention solutions can help maintain people's mental health and social relations. Though the traditional task of sentiment analysis is purely based on text (Wang et al., 2020; Ghosal et al., 2020; Shen et al., 2021), humans express emotions not only with spoken words but also through non-verbal signals such as facial expressions and the change of tones. Therefore, following the current trend of multimodal emotion recognition (Delbrouck et al., 2020; Zadeh et al., 2017; Rahman et al., 2020; Gandhi et al., 2022), we focus on addressing problems of understanding the expressed emotions in videos along with their audio and transcripts.

In this work, we tackle the problem of the multimodal emotion recognition task from two major issues: Minimal contribution of visual modality, and emotional subjectivity. Previous works which have used multimodal approaches (Rahman et al.,

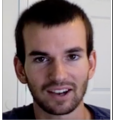



Topic	Content	Style	Happy	Sad
Public Speech	I'm thankful for all the staff who work tirelessly.	Excited		
Movie Review	The battles in the movie, they are perfect.	Calm		

Figure 1: Left table: "happy" under different topics. Right table: speaking styles can affect how emotion is displayed on the face

2020; Joshi et al., 2022; Delbrouck et al., 2020) have shown that text+audio outperforms the results of combining all three modalities. While facial and gesture signals contain abundant information, they tend to introduce more noise to the data due to its high dimensionality. In order to increase the contribution from visual modality, we propose to take advantage of the strong multimodal backbone VATT (Akbari et al., 2021) that can project features of different granularity levels into a common space. On the other hand, the expression of emotion is subjective. People's emotion judgment can be influenced by enclosed scenarios. As shown in the left two columns in Figure 1, though the two examples are all labeled as "happy", the signals we use to detect "happy" may not be the same. In a public speech, showing gratitude may mean a positive sentiment while in movie reviews, we may focus more on sentiment words like good or bad. Also, subjectivity may come from individual differences in their own emotional intensity. As the examples shown in the right three columns in Figure 1, the sadness and happiness of the person in the excited style are more distinguishable through his face while the person in the calm style always adopts a calm face that makes sad and happy less recognizable. Therefore, we introduce content-oriented features: topic and speaking style serving as a content "normalization"

for each person.

Our work makes the following contribution:

1) We propose to leverage the multimodal backbone to reduce the high dimensionality of visual modality and increase its contribution to the emotion recognition task.

2) We incorporate emotion-related features to handle modeling issues with emotional subjectivity

3) Experiments conducted on the benchmark dataset MOSEI show our model can outperform SOTA results and effectively incorporate visual signals and handle subjectivity issues.

## 2 Related Work

Emotion recognition using a fusion of input modalities such as text, speech, image, etc is the key research direction of human-computer interaction. Specific to the area of sentiment analysis, Multimodal Transformer applies pairwise cross-attention to different modalities (Tsai et al., 2019). The Memory Fusion Network synchronizes multimodal sequences using a multi-view gated memory that stores intra-view and cross-view interactions through time (Zadeh et al., 2018). TFN performs the outer product of the modalities to learn both the intra-modality and inter-modality dynamics (Sahay et al., 2018). (Rahman et al., 2020) begins the endeavor to take BERT (Devlin et al., 2018) as a strong backbone pretrained on large scale corpus. (Arjmand et al., 2021) follows the direction and combines Roberta with a light-weighted audio encoder to fuse the text and audio features. A recent work (Yang et al., 2022a) presents a self-supervised framework to pretrain features within a single modality and across different modalities. Other frameworks include context and speaker-aware RNN (Shenoy and Sardana, 2020; Wang et al., 2021), graph neural networks modeling knowledge graphs and inter/intra relations between videos (Joshi et al., 2022; Fu et al., 2021; Lian et al., 2020), while (Zhu et al., 2021) has used topic information to improve emotion detection

## 3 Method

### 3.1 Overview

Our model aims to predict the presence of different emotions given an utterance-level video input along with its audio and transcripts. Figure 2 shows the overall structure of our model. To first get a better alignment of features from different modalities,

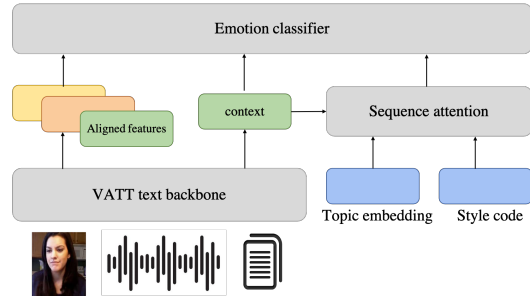


Figure 2: Network Architecture

the raw video input will be fed into our backbone VATT and we can get the corresponding projected features for visual, acoustic, and textual signals separately. Meanwhile, our high-level content module will extract the corresponding topic and style representation. Queried by the video context, the topic and style features are further merged by a cross-attention layer. Then both low-level and high-level features are concatenated and put into the final classification layer.

### 3.2 Backbone

Video-Audio-Text Transformer (VATT) is a framework for learning multimodal representations that takes raw signals as inputs. For each modality encoder, VATT appends an aggregation head at the beginning of the input sequence. The corresponding latent feature will serve as the projection head for this modality. For pretraining, contrastive loss is applied to align features from different modalities in a common projected space. Details can be found in (Akbari et al., 2021).

### 3.3 Content-oriented Features

#### 3.3.1 Topic

For each utterance input, we will first predict the topic of this utterance and feed the corresponding topic embedding into the model. Since we don't have the ground truth label for topics, we use Latent Dirichlet Allocation (LDA) (Blei et al., 2003) model to cluster all the text from the training set into 3 topics. The number of topics is decided by grid search.

#### 3.3.2 Speaking Style

We define speaking style based on the expression coefficient and the projection parameters of a 3DMM model (Blanz and Vetter, 1999). In a 3DMM model, the face shape is represented as an affine model of facial expression and facial identity:  $S = \bar{S} + B_{id}\alpha + B_{exp}\beta$ . This 3D face will be

Weighted F1	Happy	Sad	Angry	Surprise	Disgust	Fear
<b>Multilogue-Net</b>	70.60	70.70	74.40	87.80	83.40	86.00
<b>TBJE</b>	65.60	67.90	76.00	87.20	<b>84.50</b>	<b>86.10</b>
<b>MESM</b>	65.4	65.2	67.00	66.70	77.7	65.8
<b>Ours-Full</b>	<b>71.18</b>	<b>73.57</b>	76.62	87.77	82.79	86.03
<b>Full w/o text</b>	68.71	70.84	72.65	87.77	78.59	86.03
<b>Full w/o audio</b>	70.23	73.25	74.02	<b>87.82</b>	81.94	86.03
<b>Full w/o video</b>	68.95	72.76	<b>76.83</b>	87.74	82.74	86.03
<b>Full w/o content feature</b>	69.12	72.07	75.18	87.77	81.70	86.03
<b>Full w/o context</b>	70.87	73.54	75.18	87.77	80.76	86.03
<b>Full w/o style</b>	69.75	73.30	75.67	87.82	82.76	86.03
<b>Full w/o topic</b>	70.48	73.32	75.67	87.77	82.69	86.03

Table 1: Impact of different input modalities and content features

Accuracy	2-Class	7-Class
<b>Multilogue-Net</b>	<b>82.88</b>	44.83
<b>TBJE</b>	82.40	43.91
<b>Topic-Style-Context</b>	79.75	<b>48.26</b>

Table 2: Sentiment analysis on 2-class and 7-class

projected into a 2D image by translation and rotation  $p$ . Since there are multiple video frames, the expression coefficient  $\beta$  and the project parameter  $p$  will become time series  $\beta(t)$  and  $p(t)$ . For a detailed analysis of the relations between the 3DMM parameters and the talking styles, (Wu et al., 2021) collected a dataset consisting of 3 talking styles: excited, tedious, and solemn. They find out that the standard deviation of the time series features and the gradient of these features are closely related to the styles. The final style code are denoted as  $\sigma(\beta(t)) \oplus \sigma(\frac{\partial\beta(t)}{\partial t}) \oplus \sigma(\frac{\partial p(t)}{\partial t})$ ,  $\oplus$  signifies the vector concatenation.

### 3.3.3 Aggregating Different Features

Given each data input with its corresponding video ID, we collect all the transcripts with the same video ID as the context, and the context feature will be extracted from the text encoder of VATT. To adapt general topic and style features to the current speaker, we treat them as the feature sequence of length 2 and use an additional cross-attention layer to aggregate these features queried by the video context. Then this information along with the context and aligned features will be concatenated and fed into the final linear classifier.

Happy	Sad	Angry	Surprise	Disgust	Fear
8735	4269	3526	1642	2955	1331

Table 3: Label distribution of MOSEI Dataset

## 4 Dataset

We conduct our experiments on CMU-Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI (Bagher Zadeh et al., 2018)) dataset. The dataset contains more than 23,500 sentence utterance videos from more than 1000 online YouTube speakers. Each sentence is annotated for a sentiment intensity from highly negative (-3) to highly positive (+3) and for 6 emotion classes: happiness, sadness, anger, fear, disgust, and surprise. The number of utterances for train/test/dev is 16327/4662/1871 separately. Label distribution of the training set is shown in Table 3

## 5 Experiments

### 5.1 Setup

We train our models on 8 V100 GPU for 8 hours using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of  $1e-4$  and a mini-batch size of 64. The total number of parameters of our model is 155M. For topic clustering, we adopt the `scikit-learn LDA library` (Pedregosa et al., 2011). We extract the style code for each video using [https://github.com/wuhaozhe/style\\_avatar](https://github.com/wuhaozhe/style_avatar). The final model is selected based on validation accuracy on the development set.

**Task** We evaluate the performance of our model on two tasks: 1) Multi-label emotion recognition: the model needs to classify whether each of the 6 emotion classes presents or not. 2) Sentiment anal-

ysis: the model is tested on both 2-class (sentiment is positive or negative) and 7-class (a scale from -3 to +3) classification.

**Evaluation** Since the labels in MOSEI are unbalanced, we use the weighted F1 score for each emotion as the evaluation metric. We compare the performance with Multilogue-Net (Shenoy and Sardana, 2020) that adopted context and speaker-aware RNN, TBJE (Delbrouck et al., 2020), a state-of-the-art method using cross-attention for modality fusion and MESM (Dai et al., 2021), who were the first to introduce a fully end-to-end trainable model for the multimodal emotion recognition task. There are two recent works on emotion recognition, COGMEN (Joshi et al., 2022) and i-Code (Yang et al., 2022b). Since COGMEN adopted a structural representation that can exploit more relational information from other data samples and i-Code did not report the same metrics and is not open-sourced, we will not compare with them in this paper.

## 5.2 Emotion Recognition

Table 1 shows our quantitative results. Compared with other SOTA methods in the first three rows, our full model achieves the best performance on recognizing happy, sad and angry. We reckon that it is due to very limited data for surprise and fear to train the big backbone, our model does not gain much improvement (shown in Table 3). To further analyze the contribution of each component of our model design, we also conduct a detailed ablation study: 1) We first remove the aligned features from the backbone each at a time. We can see from the results in the second block that combining all three modalities in our full model outperforms the bimodality input. Especially contrasting rows with and without video input, their comparative performance validates that our model can learn effectively from visual modalities. 2) In the third block, we report the performance when we simply concatenate aligned features as the input to the emotion classification layer without high-level features. The degraded performance reveals the efficacy of our content feature design. 3) Lastly, we investigate the influence of each content feature and the aggregation using context. To remove the context, we directly apply a self-attention layer to the feature sequence and use a linear layer to project the outputs into the aggregate feature dimension. For topic and style, we just remove the corresponding

feature from the input. As shown in the last block, removing any part will result in a performance drop. Overall, our full model in comparison yields the best performance.

## 5.3 Sentiment Analysis

To further validate our methods, we run our model on the other subtask, sentiment analysis. For each data sample, the annotation of sentiment polarity is a continuous value from -3 to 3. -3 means extremely negative, and 3 means extremely positive. Our model is trained to regress the sentiment intensity. Then we ground the continuous value into 2 or 7 classes to calculate the accuracy. Contrasting 2-class and 7-class results in Table 2, our model works better for more fine-grained classification.

## 6 Qualitative Results

	Movie reviews	Finance	Commercial ads
happy	I have an incredible build up with this movie	It is important to make it affordable for small businesses	Connecting call is pretty easy to do
sad	If their team cannot win, they will be merged into other teams	They need to make internal changes as well to go through the financial crisis	Someone having congenital disability is such an issue

Figure 3: Our model can recognize happy/sad under 3 different topics

We first show that our model can correctly recognize emotions under different topics. As shown in Figure 3, for movie reviews, finance or commercial advertisements, the model can use different cues to predict the emotion as happy or sad. In Figure 4, our model can distinguish between excited/calm speaking styles and recognize the slight emotional change within each person. (all example videos can be found in supp).



Figure 4: People expressing different emotions with excited/calm styles

## 7 Conclusion and Future Work

This study employs the powerful multimodal backbone VATT to facilitate feature alignment across various modalities. Moreover, content-specific features are introduced to mitigate the influence

of individual subjectivity. The experimental outcomes demonstrate that the model can effectively assimilate visual information with reduced dimensions. Furthermore, the incorporation of sentiment-oriented features yields further improvements in the model’s performance, helping beat state of the art models on CMU-MOSEI dataset

## 8 Limitations

For modeling simplicity, we adopt the classic LDA methods to get the topic ID for each video segment. We plan to investigate more advanced topic clustering methods and check how it can be applied to multilingual cases. Also, we propose a two-stage framework that first extract topic and style features, based on which the emotion classifier will be trained. In the future, we hope to extend this work to learn features in an end-to-end manner.

## References

- Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:24206–24221.
- Mehdi Arjmand, Mohammad Javad Dousti, and Hadi Moradi. 2021. Teasel: A transformer-based speech-prefixed language model. *ArXiv*, abs/2109.05522.
- AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. [Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.
- Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, and Pascale Fung. 2021. [Multimodal end-to-end sparse model for emotion recognition](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5305–5316, Online. Association for Computational Linguistics.
- Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. [A transformer-based joint-encoding for emotion recognition and sentiment analysis](#). In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 1–7, Seattle, USA. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). Cite arxiv:1810.04805Comment: 13 pages.
- Yahui Fu, Shogo Okada, Longbiao Wang, Lili Guo, Yaodong Song, Jiaying Liu, and Jianwu Dang. 2021. Consk-gcn: conversational semantic-and knowledge-oriented graph convolutional network for multimodal emotion recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.
- Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2022. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*.
- Deepanway Ghosal, Navonil Majumder, Alexander 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*, pages 2470–2481, Online. Association for Computational Linguistics.
- Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. 2022. [COGMEN: COntextualized GNN based multimodal emotion recognition](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164, Seattle, United States. Association for Computational Linguistics.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Zheng Lian, Jianhua Tao, Bin Liu, Jian Huang, Zhanlei Yang, and Rongjun Li. 2020. Conversational emotion recognition using self-attention mechanisms and graph neural networks. In *INTERSPEECH*, pages 2347–2351.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference*. Association for

*Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Saurav Sahay, Shachi H. Kumar, Rui Xia, Jonathan Huang, and Lama Nachman. 2018. [Multimodal relational tensor network for sentiment and emotion classification](#). *CoRR*, abs/1806.02923.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. [Directed acyclic graph network for conversational emotion recognition](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1551–1560, Online. Association for Computational Linguistics.

Aman Shenoy and Ashish Sardana. 2020. [Multiloguene: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation](#). In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 19–28, Seattle, USA. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. [Multimodal transformer for unaligned multimodal language sequences](#). *CoRR*, abs/1906.00295.

Tana Wang, Yaqing Hou, Dongsheng Zhou, and Qiang Zhang. 2021. A contextual attention network for multimodal emotion recognition in conversation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue*, pages 186–195.

Haozhe Wu, Jia Jia, Haoyu Wang, Yishun Dou, Chao Duan, and Qingshan Deng. 2021. Imitating arbitrary talking style for realistic audio-driven talking face synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1478–1486.

Ziyi Yang, Yuwei Fang, Chenguang Zhu, Reid Pryzant, Dongdong Chen, Yu Shi, Yichong Xu, Yao Qian, Mei Gao, Yi-Ling Chen, Liyang Lu, Yujia Xie, Robert Gmyr, Noel Codella, Naoyuki Kanda, Bin Xiao, Lu Yuan, Takuya Yoshioka, Michael Zeng, and Xuedong Huang. 2022a. [i-code: An integrative and composable multimodal learning framework](#).

Ziyi Yang, Yuwei Fang, Chenguang Zhu, Reid Pryzant, Dongdong Chen, Yu Shi, Yichong Xu, Yao Qian, Mei Gao, Yi-Ling Chen, et al. 2022b. [i-code: An integrative and composable multimodal learning framework](#). *arXiv preprint arXiv:2205.01818*.

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.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. [Memory fusion network for multi-view sequential learning](#). *AAAI*, abs/1802.00927.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. [Topic-driven and knowledge-aware transformer for dialogue emotion detection](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582, Online. Association for Computational Linguistics.

## A Appendix

### A.1 Topic visualization

We first show the final topic clustering results. The second column shows the top 20 high frequency words in this topic and the third column shows some examples under this topic. The first topic is more related to movie reviews, the second covers business and finance, and the third one seems to associate with commercial and instruction videos.

### A.2 Style Code

In Fig 5, we can see that styles have a distinctive embedding based on emotion which confirms our hypothesis that style code can add a meaningful input to our multimodal approach.

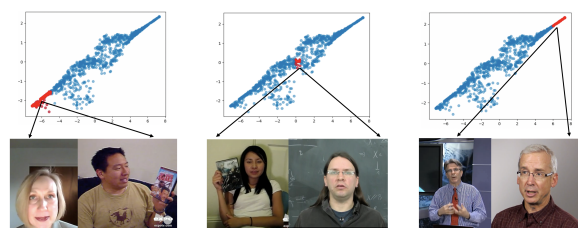


Figure 5: TSNE of Speaking style code

<b>Topic</b>	<b>Words</b>	<b>Examples</b>
<b>Topic 1</b>	movie, umm, uhh, like, know, really, one, im, good, go, see, two, kind, would, think, even, thats, going, there	1) hi there today we're going to be reviewing cheaper by the dozen which is umm the original version; 2) i was a huge fan of the original film bruce almighty but i did think it was funny like jim
<b>Topic 2</b>	people, get, think, make, business, u, want, time, world, need, company, way, also, work, one, year, take, money, right, new	1)future and it's a retirement future that can ultimately turned in to an income for you when you no longer have an income and you're fully retired; 2)um this year switching up how we approach funding and hopefully going to be able to arrange for some sustainable more officially recognized sorts of funding
<b>Topic 3</b>	going, thing, like, know, one, want, really, well, also, im, video, make, way, thats, something, think, were, time, get, look	1)is you can say hey i really like baby skin they are so soft they have any hair on their face so nice; 2) okay what happens at this point after we've taken this brief walk down memory lane is the presentation of the gift now

Table 4: Topic clustering results

## ACL 2023 Responsible NLP Checklist

---

### A For every submission:

- A1. Did you describe the limitations of your work?

8

- A2. Did you discuss any potential risks of your work?

*We do not consider any risks in our work*

- A3. Do the abstract and introduction summarize the paper’s main claims?

1

- A4. Have you used AI writing assistants when working on this paper?

*Left blank.*

### B Did you use or create scientific artifacts?

*model:3, 5.1 data: 4*

- B1. Did you cite the creators of artifacts you used?

*model:3, 5.1 data: 4*

- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?

*Left blank.*

- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?

*3.2, 4*

- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?

*The data is anonymized and discussed in the original paper*

- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?

*Not applicable. Left blank.*

- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.

4

### C Did you run computational experiments?

5

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?

*5.1*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*



- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

5.1

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

5.1

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

5.1

**D  Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

*No response.*

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

*No response.*

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

*No response.*

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

*No response.*

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

*No response.*