Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7284–7290
Marseille, 20-25 June 2022

BehanceCC: A ChitChat Detection Dataset For Livestreaming Video Transcripts

Viet Dac Lai[1], **Amir Pouran Ben Veyseh**[1], **Franck Dernoncourt**[2], **Thien Huu Nguyen**[1]
[1]Dept. of Computer and Information Science, University of Oregon, Eugene, OR, USA
[2]Adobe Research, Seattle, WA, USA
{vietl,apouranb,thien}@cs.uoregon.edu, franck.dernoncourt@adobe.com

Abstract

Livestreaming videos have become an effective broadcasting method for both video sharing and educational purposes. However, livestreaming videos contain a considerable amount of off-topic content (i.e., up to 50%) which introduces significant noises and data load to downstream applications. This paper presents **BehanceCC**, a new human-annotated benchmark dataset for off-topic detection (also called chitchat detection) in livestreaming video transcripts. In addition to describing the challenges of the dataset, our extensive experiments of various baselines reveal the complexity of chitchat detection for livestreaming videos and suggest potential future research directions for this task. The dataset will be made publicly available to foster research in this area. The dataset is freely accessible at https://github.com/nlp-uoregon/behancecc.

Keywords: Chitchat Detection, Livestreaming Video Transcripts

1. Introduction

Livestreaming is becoming an essential communication medium in human life to connect people around the world. Content creators and audiences have been using livestreaming platforms for various purposes, including video sharing (Youtube Live, and Facebook Live), gaming (Twitch), entertainment (TikTok), and online learning (Behance). Compared to the original design for most of these platforms to share pre-recorded videos, livestreaming introduces a new important feature that allow content creators to connect with their audiences in real-time, thus making the platforms more realistic and useful. In fact, its engagement efficiency has promoted livestreaming as one of the most important mechanisms to attract and retain content creators and audiences in current social platforms.

However, the shift from pre-recorded videos to livestreaming videos, which leads to an entire change of production processes, creates a major problem in the quality control of the produced videos. In particular, in pre-recording production, video content is usually well prepared and carefully post-edited. Therefore, pre-recorded videos are usually shorter (i.e., from few minutes and up to a few hours), and involve concise and focused content, fewer verbal pauses, and no audience interference. In contrast, in livestreaming production, video content is streaming live to some chosen platforms. It allows the speakers to interact with audiences in real-time through chat box or even live discussion, thus offering an effective tool to gain more exposure and engagement through the video content.

Thus, live broadcasts introduce many unfavorable quality concerns. First, livestreaming videos tend to be much longer than pre-recorded videos as post-editing and cutting are not performed in livestreaming videos. Second, speakers in livestreaming videos tend to use a lot of verbal pauses and word/phrase repetitions as in casual discussion (e.g., due to poor preparation, thinking, and hesitation). Third, interruptions and questions from audiences might divert the content of the videos from the main topics, which result in a mix of related and unrelated contents. Fourth, due to the need to socialize with the audiences, the speakers might fall into runaway small talks. In all, the transcripts of livestreaming videos might contain up to 50% of non-relevant content with respect to a designated topic according to our analysis. In all, once projected into texts, transcripts for livestreaming videos will inherit a substantial amount of off-topic content from the original videos. This might hinder the performance of existing natural language processing (NLP) toolkits that have been mostly trained on well-written texts with consistent information flow and coherent topics. To this end, detecting and removing irrelevant content in livestreaming video transcripts are important to improve robustness and performance for downstream

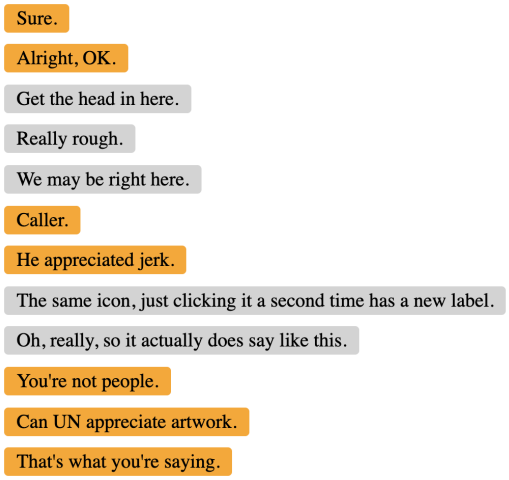

Figure 1: Chitchat texts in a livestreaming video transcript (highlighted in orange).

NLP applications over the video transcripts.

In this paper, we address the problem of chitchat detection (CC) in livestreaming videos, aiming to detect irrelevant (off-topic) texts from the video transcripts. In particular, we introduce a new dataset for chitchat detection over livestreaming video transcripts that are transcribed by automatic speech recognition (ASR) systems. Compared to prior datasets for chitchat detection, which solely focus on either telephone conversation (Konigari et al., 2021a) or group discussion (Cieri et al., 2004) in the form of dialogues, our chitchat detection dataset on livestreaming video transcripts presents several unique properties. First, in contrast to prior chitchat detection work with discussions among multiple people, livestreaming videos are mostly monologues. Second, previous chitchat datasets are only annotated on top of the existing human-created transcripts (Cieri et al., 2004; Konigari et al., 2021a) whereas in real-life end-to-end applications, transcripts are often generated by ASR systems. This discrepancy between training data of existing datasets and expected input data in real-life applications might result in a poor performance for chitchat detection. As such, our chitchat detection dataset is annotated over ASR-generated transcripts for livestreaming videos to mitigate the gap between training and inference time for the models in realistic applications. Specifically, the transcripts in our dataset, called **BehanceCC**, are obtained by applying ASR systems to the real livestreaming videos on the graphical design video sharing platform Behance[1]. **BehanceCC** is human-annotated for chitchat detection by crowd-source annotators with prior experience on graphical design tools to deliver high-quality and large-scale annotation. To promote future research for chitchat detection and livestreaming video transcripts, we will publicly release **BehanceCC** upon the acceptance of the paper.

2. Data Annotation

2.1. Preparation

The livestreaming videos annotated in this work are derived from Behance, an online platform to showcase and discover creative works on digital drawing, graphic design, and photo/video editing. Most of the time in these videos, a creator streams their work on graphic design tools. The topics in the videos are related to design theories, graphical ideas, and tutorials to use graphic design tools. To facilitate annotation, the videos are first split into shorter clips of approximately 5 minutes per clip. Then, each video clip is transcribed by the Microsoft Automatic Speech Recognition (ASR) system to produce a transcript document that also include sentence boundaries. Although the sentence boundaries are not perfect (i.e., recognized by the ASR system), we rely on this information to maintain the original characteristics of ASR-generated transcript texts. Future work can explore advanced punctuation restoration systems for video transcripts to improve setence splitting (Lai et al., 2022). After being processed to remove sensitive and private information (see our ethical statement in Section 7 for more information), the sentences in transcript documents are presented to the annotators for chitchat annotation.

The primary goal of our BehanceCC dataset is to provide an intrinsic evaluation benchmark for the chitchat detection task. Since the videos on Behance mainly focus on graphical design and photo editing, we design a taxonomy for chitchat detection for Behance Livestreaming videos with two labels: Non-chitchat (Relevant) and Chitchat (Irrelevant). Here, the relevance is defined according to the main topic of graphical design and photo editing in the videos. In particular, a relevant sentence to the main topic usually mentions related entities such as an object, action, or idea. In contrast, an irrelevant sentence does not mention these artistic entities; it may fall into some topics of causal discussion such as greeting, verbal pauses/transitions, and unlimited topics of small talks (e.g., hobby, traveling). To facilitate the recognition of main topics for comparison, the annotators are also provided with the entire original videos and aligned transcripts when annotating one sentence. Some examples for chitchat and non-chitchat sentences along with explanations are presented in Table 1.

2.2. Annotation

We recruit 4 annotators from the Upwork[2] crowdsourcing platform. As Upwork allows the freelancers to submit their resumes, we choose the most experienced annotators with prior experience on graphic creativity tools such as Adobe Photoshop and Adobe Illustrator. A detailed annotation guideline with many examples is provided to train annotators. We also develop a customized web-based annotation tool that allows the annotators to work most efficiently with the materials (i.e., videos and aligned transcripts). In particular, given a transcript document, we show each sentence in its own line. The annotators then annotate the document by choosing sets of consecutive sentences that are deemed as chitchat segments. Figure 2 shows the interface and description for our designed annotation tool. After self-practicing on the provided guideline and tool, the annotators are further trained by performing actual CC annotation on a transcript from a 2-hour video from Behance. Feedback is provided to each annotator in this process to improve the quality.

We randomly select 2911 transcript documents from the produced video clips in Behance for chitchat annotation to accommodate our budget annotation. After the training process, the four annotators independently co-annotate 20% of the selected documents, achieving the Cohen's Kappa scores of 0.65 that indicates a moderate to substantial agreement. The annotators then discuss

[1] https://www.behance.net

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

Non-chitchat
Brainstorming, commenting, discussing of designing idea: "***How about making it brighter?***"
Mentioning a tool: "*OK, so let's use* ***brush***."
Mentioning an artistic object sampleThe sky is better (in the case he/she is drawing a scenery)
Mentioning an action: "*I'm going to* ***throw*** *this moon away*""(in the case they are drawing a night scenery.
Mentioning an action with computer hardware such as keyboard, mouse, and drawing tablet: "*It is* ***C*** *to copy this moon*" "*You need to* ***double click*** *on this button*" "*Let's* ***drag*** *this to the background*" "*The textbox is ready for* ***entering***"
Planning/Introduction of the work in the video: "***I thought I'd do a Top 10 list first.***"
Mentioning color, shape, size, pattern, direction "*How about* ***1 inch****?*" "*See what happen if you drag* ***down***"
Mentioning a graphical user interface of the graphical design tool "*And let's crank up that color a lot right under* ***my properties panel****.* "
Chitchat
Welcoming: "*Hello Jimmy! Hi Lela!*"
Conversations with audience that is not related to the topic: "*I see your artful, thank you for saying hello.*"
Verbal pause (aka. Umm, Hmm): "*Ah ha! Umm... Hmm*"
Transitional sentences : "*Let's move on*"
Filling sentence, confirmation sentence: "*Tadada, that 's it.*" "*You get the idea*"
Talking about the streamer interests: "*I love cats*"(talking about his/her hobby)
Talking about tips not related to the purpose of the video: "*I usually feed the cat twice a day using this auto feeder machine.*"
Talking about traveling/careers/politics/breaking news: "*I love Rome, it was one of my best trip ever!*"

Table 1: List of types and examples of chitchat and non-chitchat sentences. Bolded words present on-topic/non-chitchat evidences in the corresponding sentences.

to resolve the conflicts over the annotated data. Finally, the remaining 80% of data is distributed to the annotators to perform separate annotation and generate the final version of **BehanceCC**. To facilitate model development and evaluation, we split the dataset into 3 portions for training/development/test data. Table 2 shows detailed statistics for our BehanceCC dataset.

Data statistics	Train	Dev	Test
#Document	2,514	198	199
#Sentence	154,897	11,175	12,216
#Token	1,466,035	99,947	105,128
Max doc length	105	88	88
#Chitchat sentence	75,980	6,031	5,566

Table 2: Statistics of the BehanceCC dataset.

3. Dataset Challenges

Several challenges in the annotated **BehanceCC** dataset should be addressed by the models to achieve good chitchat detection performance. First, the definition of chitchat segments is dependent on the main topic of the video. In another word, a segment of text might be chitchat in a video, but it can be non-chitchat in another video. For instance, in a video that focuses on graphical creative work, a transcript segment related to the "*Blizzard*" and "*World of Warcraft*" games might be irrelevant. However, if a video is related to gaming, the same segment might be tagged as non-chitchat, hence, completely flipping the labels. This dependency on context and topics makes chitchat detection in **BehanceCC** a challenging task for NLP models. In particular, in addition to current segments, the models might need to encode the original videos or entire transcripts to capture main topics to facilitate chitchat detection, making it a multi-modal learning problem in the most comprehensive setting for future work. Table 3 shows an example of two consecutive sentence segments that discuss gaming and graphical design in the same video.

Second, as the documents in **BehanceCC** are generated from an automatic speech recognition system, there is a certain number of word errors in the text even though the ASR technology has improved significantly recently. The incorrect words, propagated from the ASR system, might cause a serious problem for chitchat detection models. For example, in the word error presented in Table 4, the ASR system is unable to detect the correct mention of the acronym *"PS5"* of the *"Play Station 5"* gaming console. Instead, the ASR system recognizes it as *"PS five"*. This text includes the acronym *"PS"*, a widely used acronym in the community for the Adobe *Photoshop* application for photo editing. The example also shows that the ASR can successfully detect the acronym *"PS5"* in a later sentence. As such, the ASR system indeed has this acronym in its vocabulary but it is still unable to realize *"PS5"* in the former sentence. In all, such inconsistent and noisy transcription might introduce significant confusion to

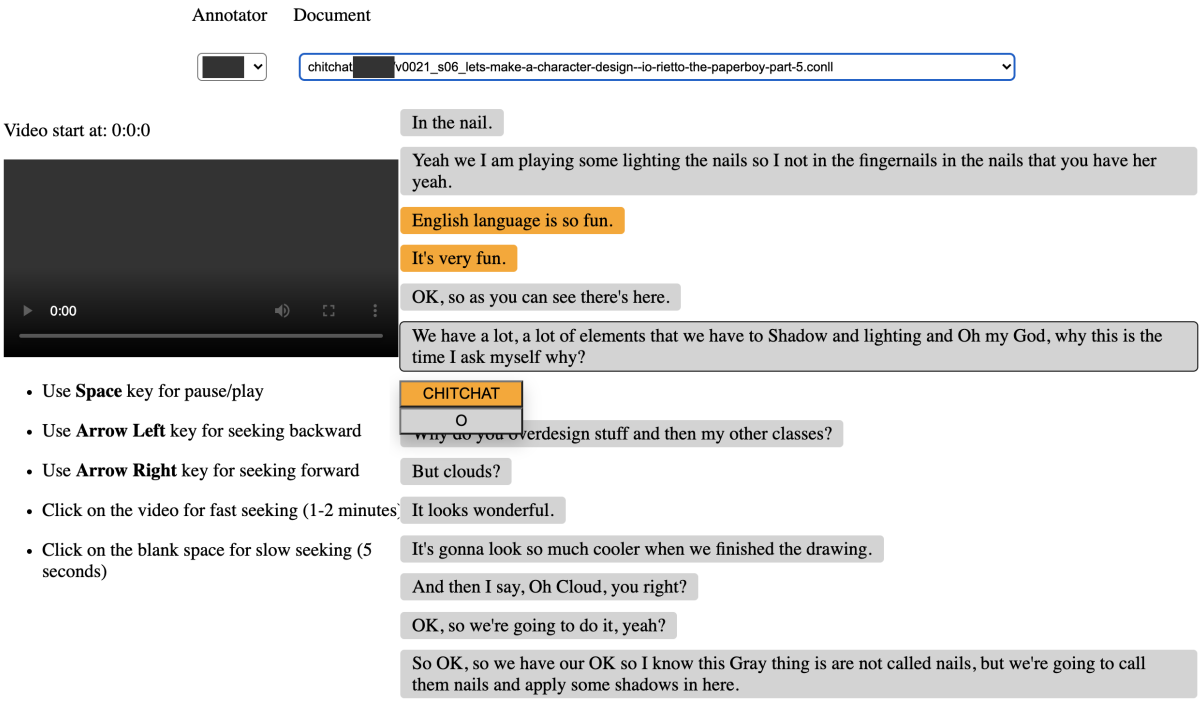

Figure 2: The interface of the web-based annotation tool we created for chitchat annotation.

Close topics
I was going to say something along the lines of like
when Blizzard Blizzard announced classic servers,
I thought that we'd never see the day.
And if we did, it would be near the end of
World of Warcraft, like for the Warcraft Lifespan.
I'm not really getting that feeling anymore.
I think that Classic in retail can easily coexist.
Right?
Did I did I want to make that head bigger.
Thanks alright.
See the ear.
Space here.
Holler That thing this comes way down lower.
A little higher.

Table 3: An example of game-related chitchat texts in the transcript of a livestreaming videos about graphical design. Chitchat sentences are *underlined*.

Word errors
*Have you purchased the **PS five**?*
I probably won't push it.
Purchase it until a year after.
Um, or whenever their second generation of
***PS five** come out.*
...
*Versions of the **PS5**, right?*

Table 4: Examples of inconsistent and noisy texts in transcripts of livestreaming videos. Chitchat sentences are *underlined*.

chitchat models, calling for the development of robust NLP systems in this area.

4. Experiments

To reveal the complexity of chitchat detection for livestreaming video transcripts in **BehanceCC**, we evaluate the performance of typical models for this problem in NLP. In particular, we explore two formulations for chitchat detection models: (i) sentence classification: the models aim to classify each sentence independently (Konigari et al., 2021a), and (ii) sentence-level sequence labeling: the models consider the sequence of sentences in a document and seek to assign a label to each sentence in the sequence to indicate chitchat or non-chitchat nature. Note that it is possible to explore multi-modal models that further encode video/audio content to improve chitchat detection for a given input text. However, in this work, we focus only on the text input information to perform chitchat detection, leaving multi-modal exploration for future work.

For all chitchat models, we start by encoding the input sentences in a transcript document with the transformer-based pre-trained language models, i.e., BERT or RoBERTa (Devlin et al., 2019; Liu et al., 2019), as they have shown state-of-the-art performance for different NLP tasks recently. Consider a transcript D of L sentences $D = \{S_2, S_2, \ldots, S_L\}$. For each sentence $S_i \in D$, we pad it with two special tokens [CLS] and [SEP] to the beginning and the end of the sentence to create a new sentence

$S'_i = [CLS], S_i, [SEP]$. Afterward, S'_i is sent into a transformer-based pre-trained language model; the vector h_i for the [CLS] token in the last layer of the language model will be used as the representation vector for the sentence S_i in the CC models (Devlin et al., 2019). As such, we obtain a sequence of vectors $H = \{h_1, h_2, \ldots, h_L\}$ for the input document D.

Given the sentence representations in H, we consider four typical architectures to perform sentence classification or sequence labeling in NLP. First, for sentence classification, we utilize the **MLP** model that directly sends the representation h_i of the input sentence into a feed-forward network to perform binary classification for CC, i.e., similar to (Konigari et al., 2021a). Second, for sequence labeling, we explore three typical models: (i) **BiLSTM**: a bidirectional long short-term memory layer (Hochreiter and Schmidhuber, 1997) is applied on top of the representation sequence H; the resulting vectors are sent to a feed-forward network to perform classification for each sentence in the sequence, (ii) **CRF**: a Conditional Random Field layer (Lafferty et al., 2001) is applied over H to capture label dependencies between sentences to perform CC, and (iii) **BiLSTM+CRF** (Huang et al., 2015): This model first stacks the BiLSTM layer over H, then introduced the CRF layer on the top for CC. All the models are trained with the negative log-likelihood.

Hyperparameters: We explore both BERT (bert-base-uncased) and RoBERTa (roberta-base) models to obtain the representation vectors H for the sentences. We fine-tune the hyper-parameters for the models on the development data of BehanceCC. As such, we select the following hyperparameter values from the tuning process: learning rate of $2e$-5 for the Adam optimizer and mini-batch size of 128 for training. For the MLP and BiLSTM models, we employ feed-forward networks with two layers and hidden states of 512 units. Finally, the BiLSTM and BiLSTM+CRF models apply a single layer of BiLSTM with 256 hidden units.

	Model	Dev			Test		
		P	R	F	P	R	F
BERT	MLP	74.8	86.7	80.3	66.7	91.5	77.2
	CRF	72.3	**90.1**	80.2	63.8	**94.3**	76.1
	BiLSTM	**78.9**	85.3	81.9	**75.2**	91.8	**82.7**
	BiLSTM+CRF	78.2	87.0	**82.4**	71.6	93.3	81.0
RoBERTa	MLP	76.3	85.1	80.5	68.0	90.4	77.6
	CRF	74.5	87.6	80.5	66.8	91.9	77.4
	BiLSTM	**77.0**	90.0	**83.0**	**70.5**	94.7	**80.8**
	BiLSTM+CRF	75.9	**90.8**	82.7	69.3	**95.1**	80.2

Table 5: Performances of the examined models on the BehanceCC dataset.

Table 5 presents the performance of the models on the development and the testing sets of the BehanceCC dataset. The first observation from the table is that the CRF layer is not very helpful for CC in BehanceCC as including it tends to reduce the performance of the models. For instance, for both BERT and RoBERTa encoder, BiLSTM+CRF is worse than BiLSTM over both development and test data. This result suggests that unlike other NLP tasks (e.g., named entity recognition), label dependencies between sentences in a transcript are not strong enough to benefit pre-trained language models for chitchat detection. In addition, we find that the BiLSTM layer can contribute significantly to the performance of the models for chitchat detection in BehanceCC. In fact, BiLSTM achieves the best test performance on BehanceCC no matter if BERT or RoBERTa is used. As such, we attribute the superiority of BiLSTM for BehanceCC to its ability to capture longer sentence context (i.e., spanning multiple sentences in the document) that facilitates the encoding of overall topics to offer better off-topic text recognition. In other words, our results indicate the importance of modeling context sentences for chitchat detection in livestreaming videos. This issue can be further observed with the MLP model that does not consider the context from surrounding sentences, thus achieving worse performance than both BiLSTM and BiLSTM+CRF over different text encoders. Finally, the best performance on the test set (i.e., 82.7%) is achieved by the BiLSTM model using BERT. However, this best performance is still far from being perfect, thus providing ample research opportunities to improve the performance on BehanceCC in future work.

5. Related work

Chitchat detection for transcripts can be considered as a segmentation of the transcripts into relevant and irrelevant parts of the talks (Stewart et al., 2006). The early work in this topic segments text into many subtopics or passages by identifying the shift of patterns of lexical co-occurrence and distribution (Hearst, 1997). Afterward, different machine learning methods have been presented to solve chitchat detection in NLP. Supervised learning models formulate chitchat detection as a classification problem (Arguello and Rosé, 2006) that exploits diverse features such as lexical feature, parts of speech, punctuation, time, content distribution, and speaker identification. In contrast, unsupervised methods for chitchat detection have exploited fine-grained conversational structures (Joty et al., 2013). In addition, in goal-oriented dialogue systems, a reinforcement learning model trained on weakly-supervised data has been proposed to encode the local topic continuity and global topic structure with LSTM (Takanobu et al., 2018). Recently, advancements in language understanding based on large pre-trained language models have also facilitated the development of neural network models (Konigari et al., 2021b) for off-topic detection in dialogues.

There have been some related studies that attempt to create resources for chitchat detection for transcripts. A taxonomy of 3 labels are common used in these studies, e.g., **Metaconversation**, **Small Talk**, and **On Topic** in the Fisher corpus (Cieri et al., 2004); and **Major Topic, Minor Topic**, and **Off Topic** in the Switch-

board corpus (Konigari et al., 2021a). However, there are several distinctions between these prior works and our work in this paper. First, we use machine-generated transcripts instead of human-generated transcripts (as done in the Fisher and Switchboard corpora). Our **BehanceCC** dataset thus mitigates the gaps between training data and expected data at inference time in the real world to allow the development of more effective models for chitchat detection. Second, the conversations in the prior datasets are not real conversations. Instead, they are simulated conversations where the participants are asked and prepared to talk about a given topic. Such simulated conversations thus involve less casualness and spontaneousness than our dataset with transcripts from realistic conversations in livestreaming videos. As casualness and spontaneousness might lead to a higher rate of chitchat sentences (i.e., more than 50% of sentences are chitchat **BehanceCC**), our dataset presents more challenges for models in this area.

6. Conclusion

This paper presents the **BehanceCC** dataset for chitchat detection on livestreaming video transcripts. We demonstrate two challenges of chitchat detection on the **BehanceCC** dataset, including topic dependency and word errors. Comprehensive experiments with state-of-the-art models highlight the importance to capture surrounding sentence context for chitchat detection. In the future, we will extend our dataset to include annotation for other NLP tasks for livestreaming video transcripts.

7. Ethical Consideration

In this work, we present a dataset on the transcripts of a publicly accessible video-streaming platform, i.e., "*Behance*". Complying with the discussion presented by (Benton et al., 2017), research with human subjects information is exempted from the required full Institutional Review Board (IRB) review if the data is already available from public sources or if the identity of the subjects cannot be recovered. However, to protect the identity of the streamers and any other people whose information is shared in the video transcripts, we impose extra processing on the transcribed documents before presenting them to annotators and publicly releasing them later. First, in this dataset, we remove the username or any other identity-related information of the streamers in the transcripts to prevent disclosing their identity. In addition, to reduce the risk of disclosing the information of other people in the transcripts, in the final version of the dataset, we exclude the transcripts that explicitly or implicitly refer to the identity of the target people. Finally, although we show video clips to annotators, we obtain consent from them to not share or use any information related to our dataset beyond this annotation project. To publicly release the dataset, we will only provide textual data (i.e., transcript documents and chitchat annotation), hence the other content of the videos (e.g., images, audios) are not revealed to users to protect human identity.

Acknowledgement

This research has been supported by the Army Research Office (ARO) grant W911NF-21-1-0112 and the NSF grant CNS-1747798 to the IUCRC Center for Big Learning. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ARO, the Department of Defense, or the U.S. Government. This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

8. References

Arguello, J. and Rosé, C. (2006). Topic-segmentation of dialogue. In *Proceedings of the Analyzing Conversations in Text and Speech*, pages 42–49, New York City, New York, June. Association for Computational Linguistics.

Benton, A., Coppersmith, G., and Dredze, M. (2017). Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain, April. Association for Computational Linguistics.

Cieri, C., Graff, D., Kimball, O., Miller, D., and Walker, K. (2004). Fisher english training speech part 1 transcripts. https://catalog.ldc.upenn.edu/LDC2004T19.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hearst, M. A. (1997). Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Joty, S., Carenini, G., and Ng, R. T. (2013). Topic segmentation and labeling in asynchronous conversations. *Journal of Artificial Intelligence Research*, 47:521–573.

Konigari, R., Ramola, S., Alluri, V. V., and Shrivastava, M. (2021a). Topic shift detection for mixed initiative response. In *Proceedings of the 22nd Annual*

Meeting of the Special Interest Group on Discourse and Dialogue, pages 161–166.

Konigari, R., Ramola, S., Alluri, V. V., and Shrivastava, M. (2021b). Topic shift detection for mixed initiative response. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 161–166, Singapore and Online, July. Association for Computational Linguistics.

Lafferty, J. D., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Lai, V. D., Veyseh, A. P. B., Dernoncourt, F., and Nguyen, T. H. (2022). Behancepr: A punctuation restoration dataset for livestreaming video transcript. In *Findings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Stewart, R., Danyluk, A., and Liu, Y. (2006). Off-topic detection in conversational telephone speech. In *Proceedings of the Analyzing Conversations in Text and Speech*, pages 8–14, New York City, New York, June. Association for Computational Linguistics.

Takanobu, R., Huang, M., Zhao, Z., Li, F.-L., Chen, H., Zhu, X., and Nie, L. (2018). A weakly supervised method for topic segmentation and labeling in goal-oriented dialogues via reinforcement learning. In *IJCAI*, pages 4403–4410.