Can Unsupervised Knowledge Transfer from Social Discussions Help Argument Mining?

Subhabrata Dutta*
Jadavpur University
subha0009@gmail.com

Jeevesh Juneja*
Delhi Technological University
creativityinczenyoga@gmail.com

Dipankar Das
Jadavpur University
dipankar.dipnil@gmail.com

Tanmoy Chakraborty
IIIT-Delhi
tanmoy@iiitd.ac.in

Abstract

Identifying argument components from unstructured texts and predicting the relationships expressed among them are two primary steps of argument mining. The intrinsic complexity of these tasks demands powerful learning models. While pretrained Transformer-based Language Models (LM) have been shown to provide state-of-the-art results over different NLP tasks, the scarcity of manually annotated data and the highly domain-dependent nature of argumentation restrict the capabilities of such models. In this work, we propose a novel transfer learning strategy to overcome these challenges. We utilize argumentation-rich social discussions from the *ChangeMyView* subreddit as a source of unsupervised, argumentative discourse-aware knowledge by finetuning pretrained LMs on a selectively masked language modeling task. Furthermore, we introduce a novel prompt-based strategy for inter-component relation prediction that compliments our proposed finetuning method while leveraging on the discourse context. Exhaustive experiments show the generalization capability of our method on these two tasks over within-domain as well as out-of-domain datasets, outperforming several existing and employed strong baselines.[1]

1 Introduction

Computational argument mining from texts is the fine-grained process of understanding opinion dynamics. In the most fundamental sense, argument understanding requires the identification of the opinions posed and justifications provided to support or falsify them. Generally, automated argument mining is a multi-stage pipeline identified with three general steps (Lippi and Torroni, 2015; Stab and Gurevych, 2017) – separating argumentative spans from non-argumentative ones, classifying argument components, and inducing a structure among them (support, attack, etc.). While different argumentation models define different taxonomies for argument components, popular approaches broadly categorize them as 'claims' and 'premises' (Stab and Gurevych, 2017; Egawa et al., 2019; Mayer et al., 2020). As these components are not necessarily aligned to sentence-level segments and can be reflected within clausal levels, the task of argument component identification requires a *token-level boundary detection* of components and *component type classification*.

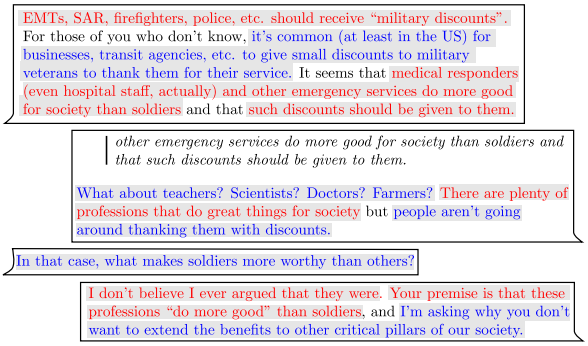

Figure 1: Token-level claim (red) and premise (blue) annotation of a discussion thread formed by consecutive posts from two users. Second post quotes a span from the first (shown in italics). Highlighted regions signify component boundaries (to demarcate consecutive components of the same kind as in the fourth post).

Context of argumentation in online discussions. Online discussions originating from back-and-forth posts from users reflect a rich interaction of opinion dynamics on large scale. In Figure 1, we show a sample argument component annotation of consecutive posts from two users. The token-level granularity of components ensures that a single sentence may contain multiple components of the same (in 1st post) or different kinds (in 2nd and 4th posts). Moreover, two adjacent spans of texts, even with the same argumentative role, can be defined as two separate components (see the 4th post for example). It is trivial to say that the meaning of any post (as well as its argumentative role) is de-

*Equal contribution

[1]We release all code, models and data used at https://github.com/Jeevesh8/arg_mining

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7774 - 7786

pendent on the context. To be specific, the third post can be identified as argumentative (a premise in this case) only when its predecessor post and its components are taken as the context. Similarly, a certain span of the first post is quoted in the second one signaling a concrete manifestation of dialogic continuity. One may even observe the user-specific argumentation styles: 1st user (author of the first and third posts) usually keeps claims and premises in separate sentences, while the 2nd user prefers to use multi-component, complex sentences. Existing studies on argumentation formalism recognize such continuity and define inter-post component relations (Ghosh et al., 2014; Hidey et al., 2017). However, the previous approaches for automated extraction, classification and relating argumentative components work on individual posts only and define the inter-post discourse in the later stages of relation prediction.

This is trivially counter-intuitive for two major reasons: (i) if we consider two text spans from separate comments to be linked by some argumentative relation, then there exists a continuity of discourse between these spans and a model is likely to benefit if it decides the boundaries and types of these two components conditioned on that continuous information; (ii) users carry their style of argumentation (simple consecutive sentences vs. long complex ones, usage of particular markers like '*I think that*' etc.), and if the model is informed about these while observing the complete conversation with back-and-forth posts, it is more likely to extract correct components easily.

Scarcity of labeled data. Irrespective of the domain, argument annotation is a resource-intensive process. A few previous studies (Habernal and Gurevych, 2015; Al-Khatib et al., 2016) attempted to exploit a large amount of unlabeled data in a semi-supervised fashion. However, such methods require the components to be defined at sentence-level (and thereby adding redundant spans into the predictions) as they perform some sentence similarity matching to generate pseudo-labels. Pretrained language models like BERT (Devlin et al., 2019) provide a workaround to handle the scarcity of task-specific annotated data. A parameter-intensive model is initially trained in a self-supervised manner on a large bulk of text; this pretraining enables the model to learn general language representation, which is then finetuned on task-specific labeled data. However, the amount of the latter still determines the expressive power of such models (Wang et al., 2020).

Present work. Considering these challenges, we formulate a novel transfer learning method using Transformer-based language models. We use large amount of unlabelled discussion threads from Reddit's *r/ChangeMyView* (CMV) community as the source of argumentative knowledge. Pretrained, Transformer-based language models are finetuned on this dataset using a Masked Language Modelling task. Instead of randomly masking tokens to predict, we select several markers in the text that are shown to signal argumentative discourse in previous works (Chakrabarty et al., 2019; Eckle-Kohler et al., 2015). The language models are then made to predict these markers in the MLM task, thereby learning to relate different components of text according to their role in the argumentation presented. We call this novel finetuning method Selective Masked Language Modeling (**sMLM**). Furthermore, to explore the role of context in argument mining, we use **sMLM** to finetune a post-level language model based on BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) and a thread-level language model based on Longformer (Beltagy et al., 2020). We present efficient incorporation of several Reddit-specific structural cues into the Longformer architecture. These finetuned language models are then used for two fundamental components of argument mining: token-level argument component identification (**ACI**) and inter-component relation type prediction (**RTP**). To further utilize the **sMLM**-based training of the language models, we propose a novel prompt-based approach to predict relations among argument components. We perform exhaustive experiments to explore the efficacy of our proposed methods for argument mining in both *in-domain* and *out-of-domain* benchmark datasets: manually annotated Reddit discussions and scientific papers. Our experiments show clear improvements achieved by our methods (0.59 and 0.69 F1 for ACI and RTP, respectively) over several state-of-the-art baselines.[2]

2 Related Work

A general overview of argument mining can be found in the survey articles by Lytos et al. (2019) and Lawrence and Reed (2019). In the current scope, we look into three major areas of research

[2]The source codes and datasets have been submitted separately.

in argument mining.

Argument component detection and classification. Previous studies have sought to address argument boundary detection and component type prediction either as separate, successive tasks in the pipeline (Stab and Gurevych, 2017) or jointly in a single computational pass (Eger et al., 2017). Studies also explored classical machine learning frameworks like SVM-HMM (Habernal and Gurevych, 2017), CRF (Stab and Gurevych, 2017), etc. with rich manual feature engineering. With the development of neural network-based algorithms, BiLSTM-CNN-CRF models emerged as a popular choice (Schulz et al., 2018; Eger et al., 2017; Chernodub et al., 2019). Very recently, large pretrained language models like BERT have also been utilized (Mayer et al., 2020; Chakrabarty et al., 2019).

Discourse markers for learning language representation. Similar to our **sMLM** finetuning strategy, Nie et al. (2019) proposed an unsupervised sentence representation learning strategy where a neural model is trained to predict the appropriate discourse marker connecting two input sentences. Using a set of 15 markers, they showed that such a finetuning can help models in downstream NLI tasks. Chakrabarty et al. (2019) used a distant supervision approach using a single marker *In my honest opinion* to finetune BERT on a large collection of *ChangeMyView* threads and then performed argument component classification. However, they did not deal with the component identification task and performed classification of already identified components at sentence-level. Opitz and Frank (2019) suggested that while identifying the relation between two components, these models often rely more on the context and not the content of the components; discourse markers present within the context provide strong signals for the relation prediction task.

Argument mining over Reddit. A few recent studies explored argumentation over Reddit. Hidey et al. (2017) proposed a two-tier annotation scheme of claim-premise components and their relations, defining five different semantic roles of premises, using *ChangeMyView* discussion data. Egawa et al. (2019) also analyzed semantic roles of argument components over *ChangeMyView* threads; however, their primary focus remained on the dynamics of persuasion, similar to Dutta et al. (2020).

3 Selective MLM finetuning of Pretrained Language Models

Though pretrained language models are developed to overcome the problem of small annotated data on different language processing tasks, they still require task-specific finetuning for better results (Wang et al., 2020). In the specific domain of argument mining, annotated data is scarce, and attempting to finetune a massive language model with very small training data comes with the risk of overfitting. Moreover, different datasets follow different strategies for annotation. We seek to devise a novel transfer learning strategy where a given Transformer-based pretrained language model is directed to focus on argumentative discourse using large-scale, unlabelled data. We choose the *ChangeMyView* (CMV) community as the source of this transfer for two specific reasons: (i) it provides us with a large, readily available resource of interactions strictly focused on debates around versatile topics, and (ii) discussions in CMV contain a mixture of dialogic continuity over successive turns along with elaborate argumentation presented in a single turn. We hypothesize that such a versatile combination of discourse can make the language model more generalizable over dialogic as well as monologic argument mining tasks.

3.1 Discourse structure of CMV

Discussion forums like Reddit facilitate users to begin a discussion with an initial post (*submissions*, in the case of Reddit) and then comments under that post to instantiate a discussion. Users may post a comment in reply to the submission as well as the already posted comments. A typical discussion over Reddit forms a tree-like structure rooted at the submission. Any path from the root to a leaf comment can be perceived as an independent dialogic discourse among two or multiple users; henceforth, we will call such paths as *threads*. Formally, a thread T is an ordered sequence $\{(u_i, P_j)|i, j \in \mathbb{N}, u_i \in U_T\}$, where P_j is a text object (a submission when $j = 1$ and a comment, otherwise), u_i is the author of P_j, and U_T is the set of all unique users engaged in the thread T. For brevity, we indicate P_j as a post in general.

The dialogic nature of discussions naturally assumes this context to be the whole thread T. However, if we consider any two successive posts P_j and P_{j+1} in T, they manifest the interests and styles of two separate participants along with the

discourse continuity of the overall thread, which must be distinguished within the definition of the context. To take into account the complete dialogic context of the thread, we represent a thread as a single contiguous sequence of tokens with each post P_j from user u_i being preceded by a special token [**USER**-i] with $i \in \{0, \cdots, |U_T| - 1\}$, to encode which post is written by which user.

Reddit also offers users a *quoting* facility: users can quote a segment from the previous post (one to which they are replying) within their posts and emphasize that their opinions are specifically focused on that segment. We delimit such quoted segments with special tokens **[STARTQ]** and **[ENDQ]** in the quoting post to demarcate the dialogic discourse. Chakrabarty et al. (2019) also used quoting as signals for following premises. Additionally, we replace URLs with the special token **[URL]** to inform the presence of external references that often act as justifications of subjective opinions.

3.2 Selective MLM finetuning

Masked Language Modeling is a common strategy of training large language models; a certain fraction of the input tokens are masked and the model is trained to predict them, consequently learning a generalized language representation. Instead of randomly selecting tokens to mask, we select specific markers that might signal argumentative discourse. While the model is trained to predict these markers, it learns the roles and relationships of the text spans preceding and following them. Following the work by Eckle-Kohler et al. (2015), we select multiple markers signaling *Opinion*, *Causation*, *Rebuttal*, *Fact presentation*, *Assumption*, *Summary*, and some additional words, which serve multiple purposes depending on the context.

As shown in Figure 2, to predict the marker *I think* in the first post, the model needs to learn that the following text span "*that most Jewish people* ⋯" expresses the user's opinion on the topic. Similarly, in the second post, for the input segment "$\langle span_0 \rangle$ *So* $\langle span_1 \rangle$ *if* $\langle span_2 \rangle$", to correctly predict the masked markers as *So* and *if*, a language model needs to learn the fact that the truth value of the statement expressed in $\langle span_1 \rangle$ is conditioned upon $\langle span_2 \rangle$, and this dependence is inferred from $\langle span_0 \rangle$.

Effect of context sizes. CMV threads provide a natural segmentation of the discourse context into comment/post-level vs. thread-level. We seek to explore the effect of the context size at different modules of argument mining (i.e., argument component detection and relation type prediction). For this, we use our proposed selective MLM approach to finetune a pretrained RoBERTa/BERT-base model in the comment/post-level regime, and train Longformer models in the thread-level regime. Longformer uses sparse, global attention (i.e., some tokens attend to all the tokens in the input sequence) to capture the long-range dependencies. We use the special tokens indicating the users (c.f. Section 3.1) as the globally attending tokens for Longformer.

u/DurianMD:
CMV: Religion is not violent or not violent, its followers are. So, my belief is that while religion can inform the views of people, it is far more likely that religion will be used to justify actions that would have been executed any way. I think that most Jewish people don't want to stone adulterers and most Muslims don't want to stone non believers.

u/recycled_kevlar:
Your stance relies on the assumption that religion has no influence on the actions of its followers beyond the superficial. Yet something must exist that allows this pattern to occur. Ill narrow it down to religion or culture. So, you are correct if you assume the culture dominates the religion, and you are incorrect if the reverse is true. With this in mind, I think its safe to assume the truth is somewhere in between, with both the religion and the culture somehow influencing the unrest we see.

u/DurianMD:
I suppose I was taking a harsh stance when I assumed that religion had no effect on behavior, when it obviously does. I still think the culture dominates religion to a great extent, however I cannot ignore that religion does have an effect on culture to some extent.

Figure 2: Example of selective masking in a sample CMV thread; **sMLM** finetuning requires a pretrained language model to predict the masked (highlighted in red) tokens (or all the subwords constituting them) based on the context.

3.3 Argument component identification

After finetuning the language model on the selective MLM task, we proceed to our first task of identifying argument components in threads. Since the detection is done in token-level, we use the standard **BIO** tagging scheme: for a component class $\langle type \rangle$, the beginning and the continuation of that component are marked as **B**-$\langle type \rangle$ and **I**-$\langle type \rangle$, respectively, while any non-component token is labeled as **O**. Therefore, if one uses the usual claim-premise model of argumentation, the label set becomes {**B**-claim, **I**-claim, **B**-premise, **I**-premise, **O**}.

3.4 Inter-component relation prediction

While identifying the relation between two given related argument components, it is important to understand the role of those text segments within the context of the discourse. Furthermore, we seek

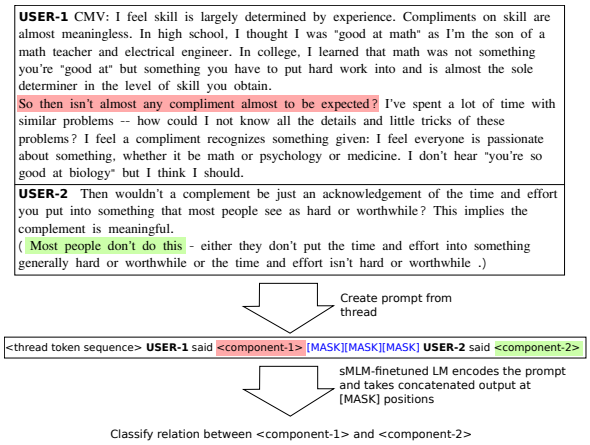

Figure 3: Example outline of prompt-based relation prediction where we seek to classify the relation between the claims posed by **USER-1** and **USER-2**, highlighted in red and green, respectively; the thread is converted to the prompt input by appending the prompt template. The language model the converts this prompt token sequence into fixed dimensional vectors from which the vector corresponding to the position of the masking token is used for relation classification.

to utilize the knowledge acquired by a language model in the **sMLM** finetuning step as well. Keeping these two factors in mind, we propose a novel, prompt-based identification of argument components. This approach is inspired by recent popularity of prompt-based fine-tuning methods in the community (Liu et al., 2021). At its core, these methods involve directly prompting the model for the required knowledge, rather than fine-tuning **[CLS]** or mean-pooled embeddings. For example, to directly use a model to summarise a text, we can append "**TL;DR:**" to the text (Radford et al., 2019), and let the model generate tokens following it; we expect the next few tokens to constitute a summary of all the previous text.

Since the underlying Transformer LMs have been trained using some Cloze task (i.e., filling the blanks from the context) previously, it is more natural for it to predict a token given a context. However, there are two challenges: (i) one needs to design a suitable prompt, and (ii) in case of classification tasks like **RTP**, it is challenging to perform **Answer Mapping**, i.e., to map all the possible tokens to some particular relation class. To tackle these challenges, we design our proposed relation prediction method in the following manner (see Figure 3)

For each pair of related components, say, component-1 and component-2, said by user-i and user-j, respectively, where component-2 refers to component-1, we append to the thread, a prompt with the template: "[USER-i] said <component1> [MASK] [MASK] [MASK] [USER-j] said <component2>" (we used three mask tokens since that is the upper bound of the marker size used for **sMLM**). We expect that the words predicted at the masked position such as "because", "in spite of what" etc. would be indicative of the relation of the two components. For the example thread shown in Figure 3, in a zero-shot prediction, **sMLM**-finetuned Longformer predicts "I", "disagree", "I" at the three masked positions. This "disagree" clearly corresponds to the **undercutter** relation between the two components. In fact, the base Longformer without **sMLM** finetuning predicts a space, a full stop and another space at the three masked positions. This additionally proves the efficacy of the **sMLM** finetuning.

Instead of engineering a token-to-relation type mapping, the predicted token embeddings at the masked positions are concatenated and fed into a linear layer to predict probabilities over the set of relation types. This way, we allow the model to learn and map from the token space to the relation type space.

4 Experiment Setup

4.1 Dataset

For the **sMLM** finetuning, we use the subset of *Winning Args (ChangeMyView)* (CMV) dataset (Tan et al., 2016) provided in ConvoKit (Chang et al., 2020). We use 99% of this data for training, and reserve 1% for checking accuracy on the **sMLM** task. The entire data consists of $3,051$ submissions and $293,297$ comments posted in the *ChangeMyView* subreddit by $34,911$ unique users. We extract the threads from these posts following the reply structure and end up with $120,031$ threads in total.

To train and evaluate all the models for **ACI** and **RTP**, we use the manually annotated Reddit discussion threads provided by Hidey et al. (2017) and further extended by Chakrabarty et al. (2019) for training and evaluation. The extended version of this dataset contains 113 CMV discussion threads manually annotated with argument components following the standard claim-premise model.

Additionally, we use the argument annotated **Dr. Inventor Corpus** (Lauscher et al., 2018) which consists of 40 scientific publications from the field of computer graphics. There are three types of argumentative components here: Background Claims

(**BC**), consisting of claims from previous works in the paper, Own Claim (**OC**) consisting of the new claims made by the authors of the paper, and **Data**. The Data class mainly consists of citations, references to figures, etc. This dataset has three relation types, viz., **support**, **contradicts** and **semantically same**. Additional dataset details are provided in Appendix A.

4.2 Baseline methods

For **ACI**, we consider two state-of-the-art token-level argument identification models: ▷ **LSTM-MTL.** Eger et al. (2017) proposed an end-to-end argument mining architecture which uses a BiLSTM-CNN-CRF sequence tagger to jointly learn component detection, classification, and relation parsing tasks. ▷ **LSTM-MData.** Schulz et al. (2018) proposed a BiLSTM-CNN-CRF based model which aims to generalize argument mining using multi-domain training data in an MTL setting. We augment our data with their original set of 6 datasets.

For **RTP**, as no prior work exists to the best of our knowledge, we consider our own baselines. First, we consider ▷ **Context-less RoBERTa**, a pretrained RoBERTa model, which takes the two components with a [SEP] token between them and predicts the relation using [CLS] token's embedding. It is context-less as only two components without the surrounding context are used to predict the label. Second, we consider ▷ **Contextless QR-Bert.** This uses the same fine-tuning methodology as **Contextless RoBERTa** and is initialized from the pre-trained Quote-Response relation prediction model of Chakrabarty et al. (2019).

For **RTP**, we try another traditional strategy, instead of prompting, for our models: ▷ **Mean Pooling**. The mean pooling approach first finds an embedding of each of the two related components by averaging the Transformer embeddings at all token positions within a component. These embeddings are concatenated and passed into a linear layer for predicting the type of relation between the two related components.

To further evaluate the efficacy of our **sMLM** training strategy, we finetune a pretrained Longformer on the Winning Args Corpus, with the usual MLM, i.e., masking 15% of tokens at random, instead of selective masking. We call this the domain-adapted Longformer, **DA-LF**.

Model	Claim			Premise			F1	Acc
	P	R	F1	P	R	F1		
sMLM-LF	0.49	0.57	**0.53**	**0.61**	**0.67**	**0.64**	**0.59**	**0.74**
Base-LF	**0.50**	0.50	0.50	0.58	0.64	0.61	0.56	**0.74**
sMLM-RoBERTa	0.49	**0.60**	**0.53**	0.55	0.57	0.55	0.55	0.72
RoBERTa	0.49	0.55	0.51	0.56	0.62	0.59	0.56	0.73
BERT	0.21	0.25	0.23	0.19	0.26	0.22	0.22	0.62
LSTM-MData	0.19	0.18	0.18	0.26	0.23	0.24	0.22	0.54
LSTM-MTL	0.19	0.18	0.18	0.24	0.25	0.24	0.21	–

Table 1: Performance of different models on **ACI**-task on CMV Modes dataset (P: Precision, R: Recall, F1: F1 score). The **F1** and **Acc.** in the rightmost columns denote the micro-averaged F1 score over claims and premises and the token level accuracy of predicting argumentative tags, respectively.

Model	BC			OC			Data		
	P	R	F1	P	R	F1	P	R	F1
sMLM-LF	0.45	**0.52**	0.48	**0.39**	0.45	0.42	**0.50**	**0.48**	**0.48**
Base-LF	**0.49**	0.51	**0.50**	0.38	**0.50**	**0.43**	0.44	0.44	0.44

Table 2: Results on Dr. Inventor dataset for argument component identification using **sMLM**-finetuned and base Longformer models.

4.3 Implementation details

We use the pretrained base version of Longformer (12 layers, 768 model size). The size of the local attention window was set to the default 512. The maximum sequence length was fixed at 4096.

Following the suggestions in Reimers and Gurevych (2017), we repeat our experiments on the 5 different data splits. The scores reported in the tables for various models correspond to the average value of the mean of 5 runs, over the last 5 epochs for that particular metric. We provide additional implementation details in Appendix B.

5 Evaluation

We evaluate the models based on precision, recall, and F1 scores for predicting claims and premises. For a more rigorous setting, we use exact match of the whole span between gold and predicted labels, i.e., if the gold label is [**O**, **B**-claim, **I**-claim, **I**-claim, **I**-claim, **O**] then only the predictions [**O**, **B**-claim, **I**-claim, **I**-claim, **I**-claim, **O**], or [**O**, **I**-claim, **I**-claim, **I**-claim, **I**-claim, **O**] can be considered as true positives. We use the popular SeqEval (Nakayama, 2018) framework.

5.1 Argument component identification

Table 1 shows the results for argument component identification on the CMV Modes dataset. We compare models based on their micro-averaged F1 over the two component types (claims, premises), and token level accuracy. Firstly, we observe huge difference in token-level accuracy scores as we move from the existing best performing LSTM based methods with accuracy of 0.54 to BERT,

Model	Support			Agreement			Direct Attack			Undercutter			Partial			Overall
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	F1
80-20 split																
sMLM-LF-prompt	**0.88**	**0.93**	**0.91**	**0.51**	**0.46**	**0.48**	**0.32**	**0.35**	**0.33**	**0.43**	**0.51**	**0.46**	0.28	0.12	0.16	**0.67**
DA-LF-prompt	0.78	0.84	0.81	0.44	0.45	0.43	0.22	0.19	0.19	0.30	0.32	0.30	0.27	0.11	0.15	0.61
sMLM-LF-mp	0.73	0.87	0.79	0.49	0.36	0.38	0.32	0.24	0.26	0.32	0.33	0.41	0.35	**0.21**	**0.25**	0.59
Base-LF-prompt	0.79	0.88	0.84	0.48	0.44	0.46	0.30	0.23	0.25	0.31	0.39	0.34	**0.37**	0.12	0.17	0.62
Base-LF-mp	0.71	0.87	0.78	0.47	0.33	0.37	0.24	0.17	0.18	0.27	0.26	0.26	0.35	0.20	0.24	0.56
RoBERTa	0.78	0.83	0.80	0.46	0.34	0.37	0.29	0.29	0.28	0.15	0.24	0.18	0.36	0.15	0.20	0.60
QR-Bert	0.76	0.85	0.80	0.46	0.27	0.34	0.21	0.13	0.16	0.19	0.25	0.20	0.32	0.16	0.20	0.59
50-50 split																
sMLM-LF-prompt	**0.87**	**0.92**	**0.89**	**0.53**	**0.47**	**0.49**	0.30	**0.28**	**0.28**	**0.45**	**0.58**	**0.50**	0.35	0.09	0.14	**0.69**
DA-LF-prompt	0.85	0.89	0.87	0.47	**0.47**	0.44	**0.32**	0.20	0.24	0.39	0.55	0.44	0.32	0.13	0.16	0.66
sMLM-LF-mp	0.70	0.90	0.79	0.426	0.22	0.28	0.28	0.20	0.22	0.32	0.26	0.28	0.38	**0.18**	0.24	0.56
Base-LF-prompt	0.78	0.87	0.82	0.49	0.44	0.46	0.30	0.19	0.22	0.32	0.40	0.35	0.32	0.13	0.18	0.62
Base-LF-mp	0.73	0.86	0.79	0.36	0.21	0.26	0.25	0.18	0.21	0.23	0.28	0.25	**0.43**	**0.18**	**0.25**	0.56
RoBERTa	0.72	0.83	0.77	0.47	0.25	0.31	0.22	0.21	0.21	0.13	0.16	0.14	0.17	0.08	0.10	0.55
QR-Bert	0.72	0.84	0.77	0.47	0.28	0.34	0.19	0.13	0.14	0.13	0.18	0.15	0.22	0.07	0.09	0.54

Table 3: Relation type wise Precision (P), Recall (R) and F1 score on the CMV Modes dataset for various models. The highest scores in every column are in **bold**. The suffix "mp" and "prompt" indicate that the model was trained using **Mean Pooling** and **Prompting** strategies, respectively. The **F1** in last column is the Micro/weighted-F1 over all the prediction classes.

Relation types	Base-LF-prompt			sMLM-LF-prompt		
	P	R	F1	P	R	F1
Support	**0.91**	0.90	**0.91**	0.89	**0.92**	**0.91**
Contradict	0.60	**0.60**	**0.60**	**0.65**	0.55	**0.60**
Semantically same	0.74	**0.77**	0.75	**0.77**	0.75	**0.77**

Table 4: Relation Type wise Precision (P), Recall (R) and F1 score on Dr. Inventor Corpus for prompt-based relation prediction using **sMLM** and base Longformer models.

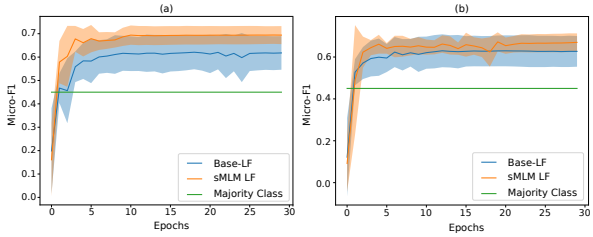

Figure 4: Micro-F1 scores for predicting relation types among argument components by Base and sMLM-finetuned Longformer models over the course of training using (a) 50-50 split and (b) 80-20 split. We use 5 different runs on random splits for each model to report the mean (solid lines) and variance.

having an accuracy of 0.62. Such a difference is expected since pretrained language models like BERT provide a head-start in case of small datasets like CMV Modes. Though the token-level accuracy increases, the micro-averaged F1 for exact component match does not increase much till we start using RoBERTa. Since pretrained Longformer was trained originally from the RoBERTa checkpoint (Beltagy et al., 2020), we can conclude that RoBERTa provides significant performance gain compared to BERT, owing to its larger training data and protocol. Longformer trained with our proposed **sMLM** finetuning clearly outperforms the rest of the models in terms of overall F1 score for component identification. However, the effects of selective MLM is more prominant in case of thread-level context (i.e, Longformer) compared to comment-level context (i.e, RoBERTa).

We can observe that context plays different roles for different component types: while **sMLM**-finetuned Longformer and RoBERTa perform comparably for claim detection, in case of premises, the access to the complete context helps the Longformer to perform better. We can observe a similar trend in **ACI**-task on Dr. Inventor dataset (see Table 2). While Base Longformer performs comparable to its **sMLM** counterpart to detect Background and Own Claims, **sMLM** provides a 4 point improvement in F1 score for the Data class which plays a similar role of premises towards the claims. Intuitively, textual segments expressing claims contain independent signals of opinion that is less dependent on the context; pretrained language models might be able to decipher their roles without additional information either from the thread-level context (in case of CMV Modes, specifically) or enhanced relation-awareness induced by the **sMLM** finetuning. However, identifying segments that serve the role of premises to a claim intrinsically depends on the claims as well as the discourse expressed in a larger context.

5.2 Relation type prediction

In Table 3, we present the results for relation type identification on the CMV Modes dataset. We again compare models based on their micro-averaged F1 over all relation types. Firstly, we consider the traditional mean pooling approach. Within this approach, we observe a 3 point improvement for the **sMLM** pre-trained Longformer on the 80-20 split, while maintaining same performance on the 50-50 split. Furthermore, the prompt based methods consistently outperform the mean pooling one, irrespective of whether we use base Longformer or **sMLM** pretrained one.

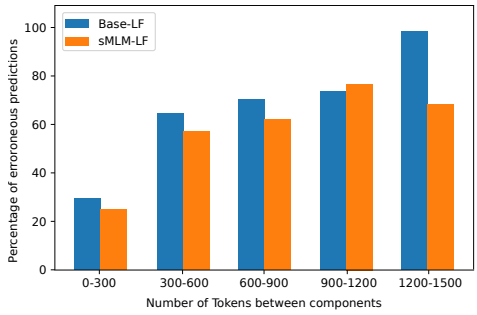

Figure 5: Percentage of erroneous classifications for **RTP** for Base-LF-prompt and LF-sMLM-prompt on component-pairs at different distances.

Within the prompting approach, we also observe increased and consistent improvement in performance due to **sMLM** pretraining on both 80-20 and 50-50 splits. The gap in micro-F1 scores between **sMLM** and base Longformer for 80-20 split increases from 3 points in mean pooling to 5 points in prompting (0 to 7 points improvements for 50-50 split). As we can observe in Figure 4, **sMLM**-finetuned Longformer admits a very narrow margin of variation on random splits, compared to the base Longformer. Furthermore, **sMLM** finetuning consistently outperforms domain-adapted finetuning (**DA-LF**), indicating the unique knowledge transfer achieved by the former.

We hypothesise that this approach works better as this regime models our final **RTP** task, as a task that is *more natural* (in a sense similar to the $(\tau, B)-$natural tasks of Saunshi et al. (2021)) for a Longformer model pre-trained with **sMLM**. Intuitively, the model learns to predict discourse markers at masked positions during **sMLM** pre-training and during fine-tuning on downstream tasks too, the model will naturally try to predict discourse markers at the masked positions. The discourse markers occurring at the masked positions are directly related to the relation between the two components. For instance, when there is a "but" between two components, we know that the two components present opposing views more or less. Here again, we observe that **sMLM** does not hurt the base performance under domain shift (Table 4).

We observe that the RoBERTa model performs worse than Base-LF-prompt, which incorporates the entire context of the thread. Also the effect worsens with reduced training set size, and RoBERTa model performs worse by 7 points in terms of micro-F1 for the 50-50 split. Furthermore, we observe that the mean pooling strategy, even though it uses context, performs worse (by 4 points on 80-20 split) than the context-less RoBERTa. Though, our **sMLM** pretrained model, manages to perform at par with the context-less RoBERTa with the mean pooling strategy. *This means, that the using the right fine-tuning method is essential. Extra context can be utilised fully in longformer, only when pre-training and fine-tuning tasks are nicely aligned.*

Model	Claim			Premise			F1
	P	R	F1	P	R	F1	
base-LF-near	0.39	0.59	0.47	0.63	0.52	0.57	0.52
base-LF-far	0.42	0.57	0.48	0.63	0.55	0.59	0.54
sMLM-LF-near	0.36	0.48	0.40	0.68	0.65	0.66	0.57
sMLM-LF-far	0.46	0.57	0.51	0.63	0.63	0.63	0.58

Table 5: Performance of base Longformer and **sMLM** Longformer for predicting segments having some markers in "**near**" (5 tokens on either side of its) boundaries, and the rest of segments ("**far**").

5.3 Dependence on the presence of markers

Following the analyses presented by Opitz and Frank (2019), we investigate whether the presence/absence of the markers used in the **sMLM** step within the vicinity of the components play any role in the **ACI** or **RTP** performances. Since the relation type among component-pairs that reside distant from each other are less likely to be inferred by the presence of markers in the context, we analyse the percentage of wrong predictions as we vary the distance between two related components, in Figure 5. While error rate does vary proportionally to the distance, we observe that **sMLM-LF** consistently yields lower percentage of wrong predictions as we vary the distance between the related components compared to base Longformer. This clearly indicates the superior capability induced by the **sMLM** finetuning to decipher the relationship among components not linked by direct context (i.e., not within a sentence or a single comment).

For the **ACI** task, however, we observe that the absence of markers in the vicinity of the components actually enables better identification, both in case of **sMLM** finetuned and pretrained Longformer (see Table 5).

6 Conclusion

We presented the results for two important tasks in the argument mining pipeline, viz., **ACI** and **RTP**. The experiments clearly elucidated the importance of alignment between the downstream and pre-trainig tasks, and the effect of various ways of modelling the tasks. The importance of entire thread's context in discussion forums, as well

as how to incorporate that into transformer-based models fruitfully has also been made clear.

Acknowledgements

The authors would like to thank Chris Hidey and Smaranda Muresan, for clarifications providing regarding their work. T. Chakraborty would like to acknowledge the support of Ramanujan Fellowship, CAI, IIIT-Delhi and ihub-Anubhuti-iiitd Foundation set up under the NM-ICPS scheme of the Department of Science and Technology, India.

References

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404, San Diego, California. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150v2.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PERSuAsive oNline discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Z. Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations.

Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. TARGER: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 195–200, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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. Association for Computational Linguistics.

Subhabrata Dutta, Dipankar Das, and Tanmoy Chakraborty. 2020. Changing views: Persuasion modeling and argument extraction from online discussions. *Information Processing & Management*, 57(2):102085.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2242, Lisbon, Portugal. Association for Computational Linguistics.

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2019. Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 422–428, Florence, Italy. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, Baltimore, Maryland. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137, Lisbon, Portugal. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, 43(1):125–179.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Marco Lippi and Paolo Torroni. 2015. Argument mining: A machine learning perspective. In *Theory and Applications of Formal Argumentation*, pages 163–176, Cham. Springer International Publishing.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

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.

Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, and Kalina Bontcheva. 2019. The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management*, 56(6):102055.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2108–2115. IOS Press.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics.

Juri Opitz and Anette Frank. 2019. Dissecting content and context in argumentative relation analysis. In *Proceedings of the 6th Workshop on Argument Mining*, pages 25–34, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging.

Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. 2021. A mathematical exploration of why language models help solve downstream tasks.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41, New Orleans, Louisiana. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, 43(3):619–659.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah. 2020. Adaptive self-training for few-shot neural sequence labeling. *CoRR*, abs/2010.03680v2.

A Dataset Details

Stats for the CMV Modes dataset are provided in Table 6. These stats are obtained after truncation of threads to 4096 token sequence length. During data analysis, we observed that several threads share the same initial post(*submission*). Hence, we make sure that all threads with the same initial post entirely lie in either the train split, or the test.

For both CMV Modes, and Dr. Inventor Corpus, we only consider contiguous spans of texts as single components, as opposed to the labelling in the dataset. Discontiguous spans are re-labelled as separate components and the model is trained and tested with these new labels, instead.

For CMV Modes dataset, we add an extra "continue" class of relations to denote relation between two dis-contiguous spans of same argumentative component annotated in the data. We group together various relation types annotated in the CMV modes data into the 5 broad classes as follows: **support**("continue" class and "support" class), **agreement**("agreement", "understand" classes), **direct attack**("attack", "rebuttal attack", "rebuttal", "disagreement" classes), **undercutter attack**("undercutter", "undercutter attack" classes), **partial**("partial agreement", "partial attack", "partial disagreement" classes). These groupings are based on the broad annotation guidelines provided for the annotations of CMV Modes data.

For Dr. Inventor Corpus, due to the low number of **semantically same** relations(44) compared to **support**(4535) and **contradicts**(564) in the original dataset, we add the label("**parts-of-same**") which indicates that two dis-contiguous spans belong to the same argumentative component to the **semantically same** category. We also, merge together sections of papers to efficiently utilise 4096 token length of Longformer model. The detailed statistics after truncation to 4096 sequence length are presented in Table 7.

Component Type	**# Tokens**
O	28186
B-C	1650
I-C	26529
B-P	1980
I-P	36552
Relation Types	**# of relations**
support	1859
agreement	421
direct attack	283
undercutter attack	330
partial	215

Table 6: Statistics for the CMV-Modes dataset.

Component Type	**# Tokens**
O	153429
B-BC	3215
I-BC	39574
B-OC	5300
I-OC	74239
B-D	3994
I-D	19058
Relation Types	**# of relations**
support	4535
Contradicts	564
Semantically Same	1049

Table 7: Statistics for the Dr. Inventor dataset.

B Implementation Details

We use the pretrained base version of Longformer (12 layers, 768 model size). The size of the local attention window was set to the default 512. The maximum sequence length was fixed at 4096. We added the special tokens that we used, to the pretrained Longformer tokenizer. For **ACI** our models use a CRF layer[3]. **sMLM** training for Longformer based models was done on thread level and for BERT and RoBERTa based models on comment-level. We used mini-batch learning; approximately similar length input threads were batched together keeping the total number of tokens per batch fixed to $8,194$ for Longformer and 1024 for BERT and RoBERTa models, and accumulated gradients over 3 batches.

We trained our models for a total of 10 epochs on sMLM task, while saving checkpoints after each epoch. We used Adam optimizer with a learning rate of 10^{-6}. For all downstream tasks, we train our models for 30 epochs, again, with Adam optimizer with learning rate of $2e-5$ as suggested by Mosbach et al. (2021). We use same batch sizes as **sMLM** training and accumulate gradients over 4 batches. We observe that for prompting **RTP** on CMV-Modes, not making **[USER-i]** tokens global, leads to better performance, hence we report results for same.

We find that **sMLM** training for 4 epochs is most beneficial, for performance on downstream task. Hence, we report results for the same checkpoint. Following the suggestions in Reimers and Gurevych (2017), we repeat our experiments on 5 different data splits and present the distributions in the Appendix. For the results at any epoch, the score plotted corresponds to mean over the 5

[3]We use the implementation of AllenNLP (Gardner et al., 2018)

Type	Markers
Opinion	*i agree, i disagree, i think, in my opinion, imo, imho*
Causation	*because, since, as, therefore, if, so, according to, hence, thus, consequently*
Rebuttal	*in contrast, yet, though, in spite of, but regardless of, however, on the contrary*
Factual	*moreover, in addition, further to this, in fact, also, firstly, secondly, lastly*
Assumption	*in the event of, as long as, so long as, provided that, assuming that, given that*
Summary	*tldr*
Misc.	*why, where, what, how, when, while*

Table 8: Types and examples of different discourse markers used for selective MLM finetuning.

runs, and error regions correspond to the Bessel corrected standard deviation. The scores reported in the tables for various models correspond to the average value of the mean of 5 runs, over the last 5 epochs for that particular metric. Table 8 provides examples of markers of various kinds, that are masked during the **sMLM** training.

C Additional results

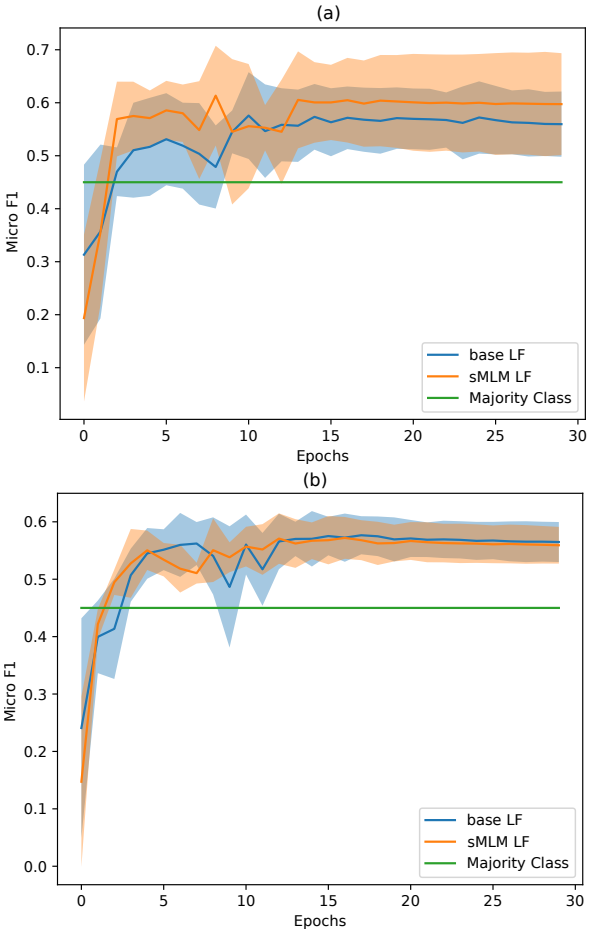

Figure 6: On CMV Modes data, sMLM-LF-mp's mean F1 converges to 0.59 compared to 0.56 for Base-LF-mp in 80-20 split (a) and 0.56 in 50-50 split (b).

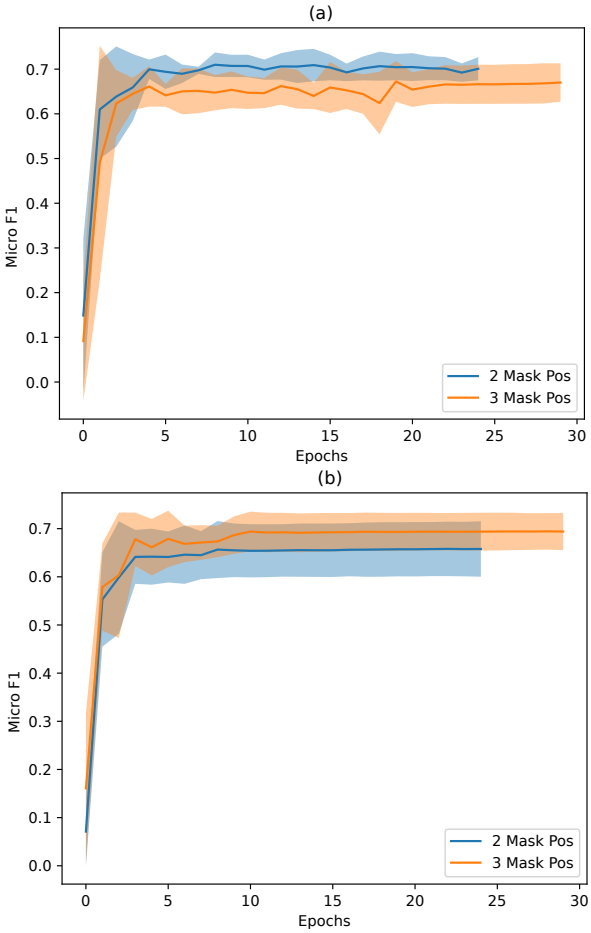

Figure 7: Change in sMLM-LF performance on CMV Modes **RTP** (a) 80-20 and (b) 50-50 split when number of mask tokens in the prompt is changed from 3 to 2. The model with 2 masked token converges to 0.70 (0.66) and the mean for 3 masked tokens converges to 0.67 (0.69).

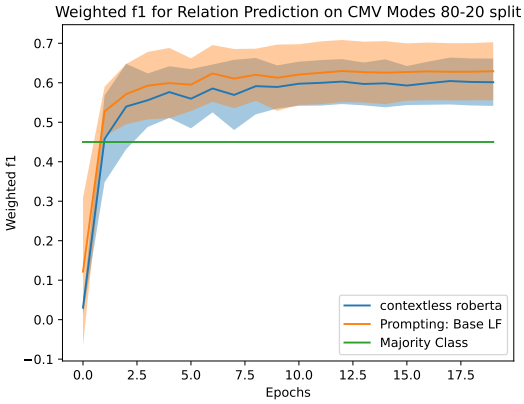

Figure 8: Contextless Roberta's mean f1 converges to around 0.599, compared to 0.62 of Base Longformer on **RTP**.

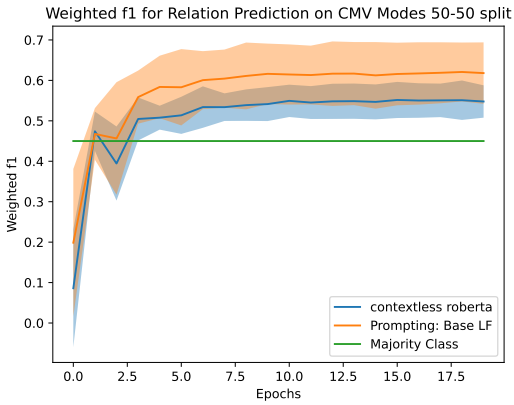

Figure 9: Contextless Roberta's mean f1 converges to around 0.55, compared to 0.617 of Base Longformer on **RTP**.

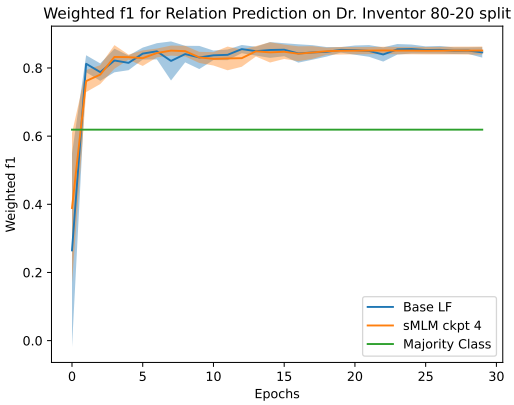

Figure 12: Both Base LF and our sMLM pretrained Longformer converge to an f1 of 0.85 with prompt-based **RTP** on Dr. Inventor corpus.

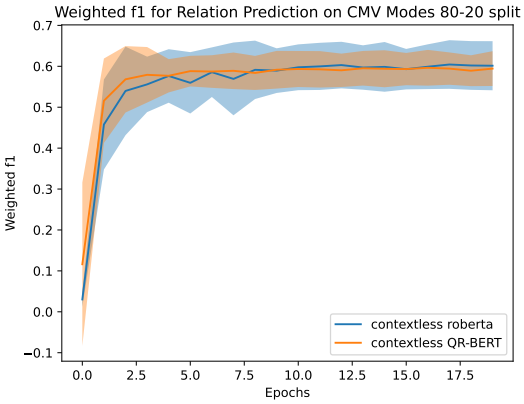

Figure 10: QR-BERT converges to an f1 score 0.59 compared to 0.60 for RoBERTa on the 80-20 split of CMV-Modes for **RTP**.

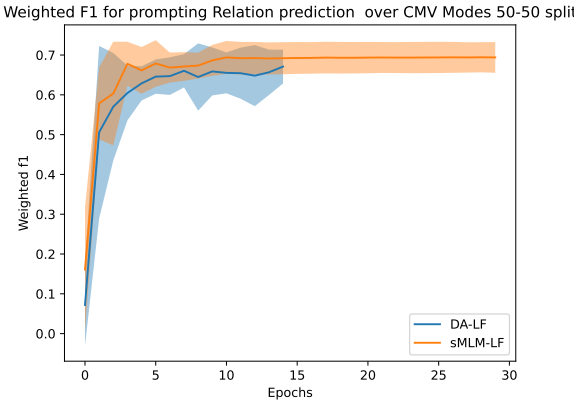

Figure 13: The Domain Adapted LF converges to around 0.66 compared to 0.69 for sMLM-LF, on the 50-50 split on CMV-Modes

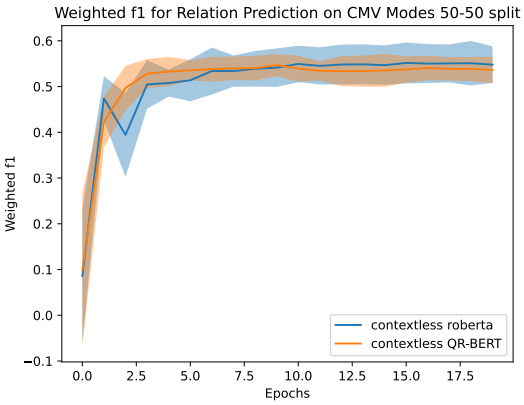

Figure 11: QR-BERT converges to an f1 score 0.54 compared to 0.55 for RoBERTa on the 50-50 split of CMV-Modes for **RTP**.

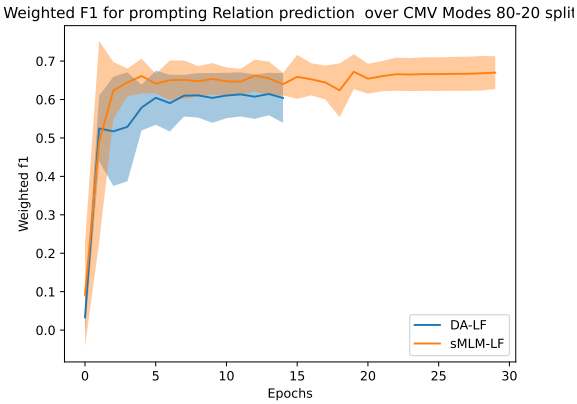

Figure 14: The Domain Adapted LF converges to around 0.61 compared to 0.67 for sMLM-LF, on the 80-20 split on CMV-Modes