Conversation Derailment Forecasting with Graph Convolutional Networks

Enas Altarawneh
York University
enas@eecs.yorku.ca

Ameeta Agrawal
Portland State University
ameeta@pdx.edu

Michael Jenkin
York University
jenkin@eecs.yorku.ca

Manos Papagelis
York University
papaggel@eecs.yorku.ca

ACL_2023

Code:None.

2023.9.10 • ChongQing

gesis Leibniz-Institut für Sozialwissenschaften

Reported by Yang Peng

1.Introduction

2.Method

3.Experiments

Turn	User	Text	Label	
$N-3$	$\mathcal{A}$	*"Proper use of an editor's history includes fixing errors or violations of Wikipedia policy or correcting related problems on multiple articles."*		正确使用编辑的历史记录包括纠正错误或违反维基百科政策，或纠正多篇文章中的相关问题
$N-2$	$\mathcal{B}$	*"It's very clear that you just go to my contributions list and look to see what biography articles I've worked on, then you go and look to see if you can find something wrong with them. "*		很明显，你只要去我的投稿列表，看看我写过什么传记文章，然后去看看你是否能发现它们有问题
$N-1$	$\mathcal{A}$	*"So, what is wrong with fixing things? At the top of my talk page, it says to keep it on your watchlist. "*		那么，修改错误有什么错呢？在我谈话页面的顶部，上面写着要把它列入你的观察名单。
N	$\mathcal{B}$	*"You cannot possibly be too stupid to understand the warning I'm giving you. I'm not going to repeat it."*	?	你不可能愚蠢到不理解我给你的警告。我不打算重复了。

Table 1: A sample conversation from the Conversation Gone Awry (CGA) dataset showing a sequence of text utterances that end with a verbal abuse. Given the conversation context up to $N-1$ turns, the task is to predict whether turn N will be a respectful or offensive statement prior to it being presented leading to derailment (**it is offensive**, in this case).

Introduction

We propose a novel model based on a graph convolutional neural network, the Forecasting Graph Convolutional Network (FGCN), that **captures dialogue user dynamics and public perception** of conversation utterances.

We perform an extensive empirical evalua tion of FGCN that shows it outperforms the state-of-the-art models on the GCA and CMV benchmark datasets by 1.5% and 1.7%, re spectively.

Method

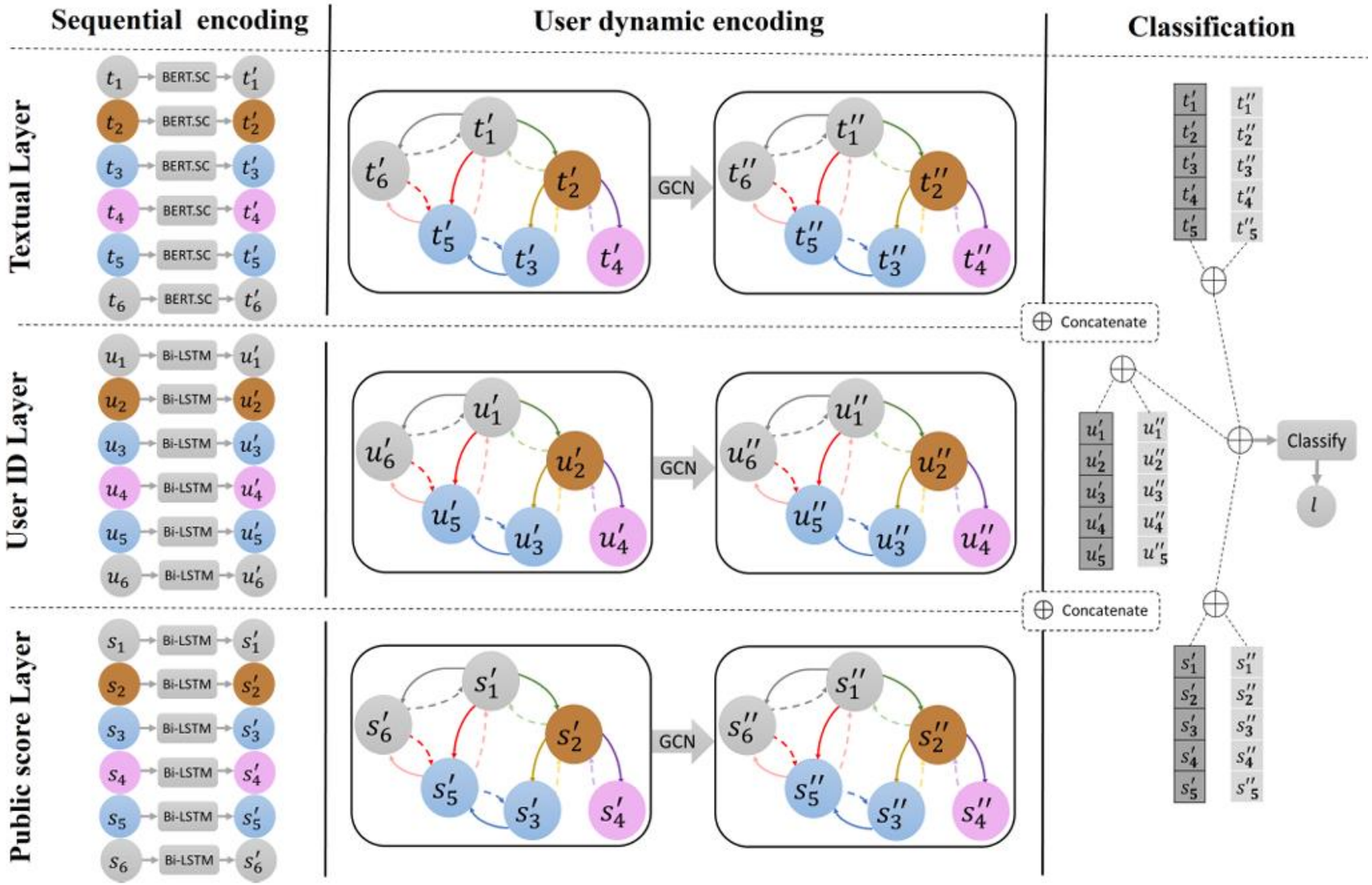
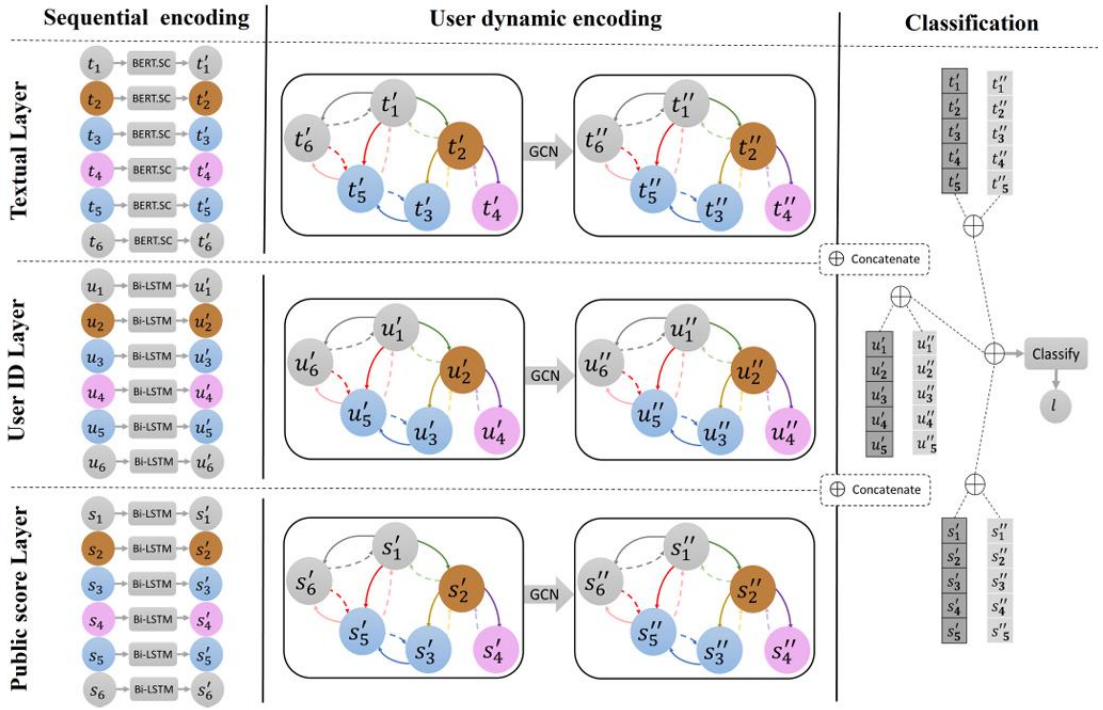


Figure 2: The FGCN model architecture.

Method

Graph Construction



$$\mathcal{C} = \{\{t_1, t_2, ..., t_N\}, \{u_1, u_2, ..., u_N\}, \{s_1, s_2, ..., s_N\}\}$$

$$G_x = (V, E, R, W)$$

input $x \in \{t, u, s\}$

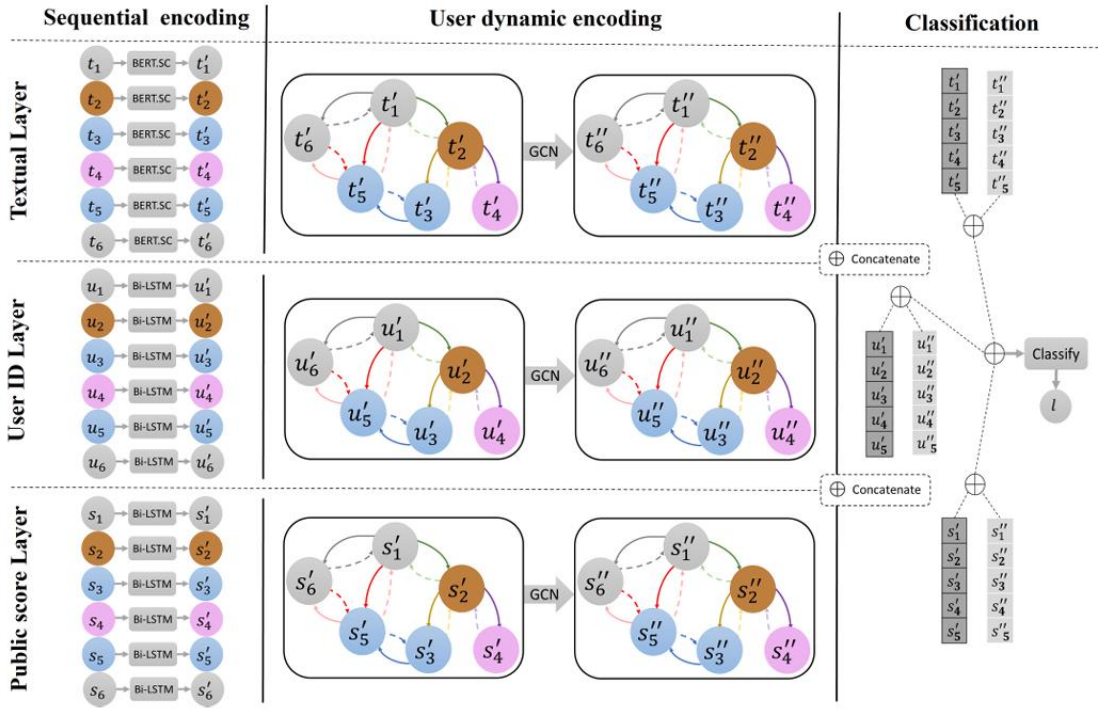
a text-based G_t

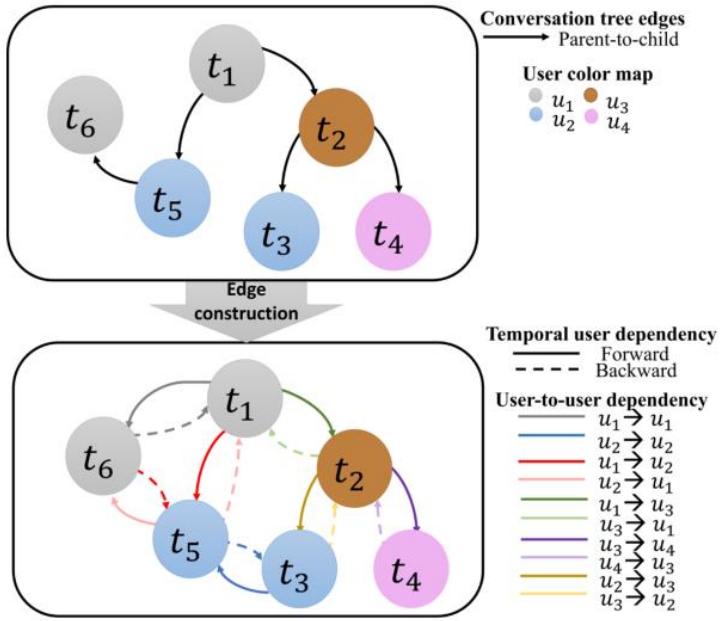
a user-based G_u

perception score-based G_s

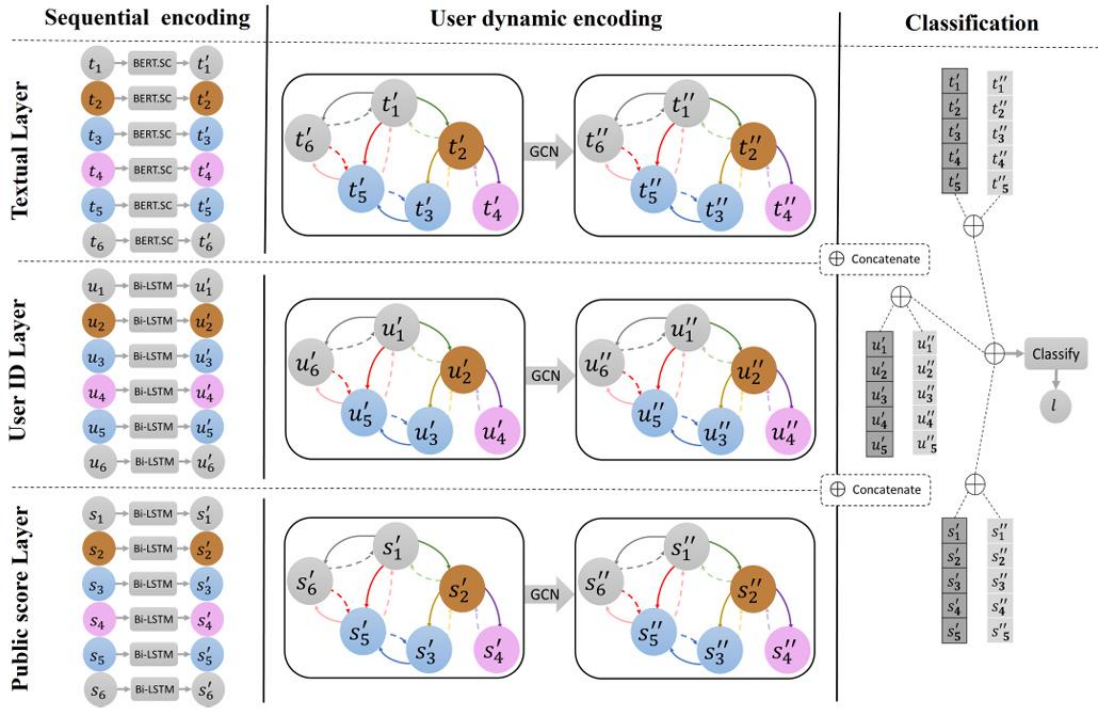
Method

User to user relationship edge construction





Method



Forecasting Derailment

$$g_i = [t_i^{'}, u_i^{'}, s_i^{'}, t_i^{''}, u_i^{''}, s_i^{''}]$$

$$C^{'} = [g_1, g_2 \cdots, g_{N-1}]$$

Feature Transformation

$$\alpha_{ij} = softmax(v_i^T W_e [v_{j_1}, ..., v_{j_m}])$$

$$u_i^{''} = \sigma(\sum_{r \in R} \sum_{j \in N_i^r} \frac{\alpha_{ij}}{c_{i,r}} W_r \, u_j^{'} + \alpha_{ii} W_0 \, u_i^{'}),$$

$$t_i^{''} = \sigma(\sum_{r \in R} \sum_{j \in N_i^r} \frac{\alpha_{ij}}{c_{i,r}} W_r \, t_j^{'} + \alpha_{ii} W_0 \, t_i^{'})$$

$$s_i^{''} = \sigma(\sum_{r \in R} \sum_{j \in N_i^r} \frac{\alpha_{ij}}{c_{i,r}} W_r \, s_j^{'} + \alpha_{ii} W_0 \, s_i^{'})$$

$$u_i^{''} = \sigma(\sum_{j \in N_i^r} W \, u_j^{''} + \alpha_{ii} W_0 \, u_i^{''}),$$

$$t_i^{''} = \sigma(\sum_{j \in N_i^r} W \, t_j^{''} + \alpha_{ii} W_0 \, t_i^{''})$$

$$s_i^{''} = \sigma(\sum_{j \in N_i^r} W \, s_j^{''} + \alpha_{ii} W_0 \, s_i^{''})$$

Experiments

Dataset	Input			Train	Val	Test
	t	u	s			
CGA	✓	✓	×	2508	840	840
CMV	✓	✓	✓	4106	1368	1368

Table 2: Statistics of the datasets. t denotes text input, u denotes user ID input and s denotes public perception score input. All splits are balanced between the two classes.

Experiments

Traning	Model	CGA				CMV			
		Acc	P	R	F1	Acc	P	R	F1
Static	CRAFT	64.4	62.7	71.7	66.9	60.5	57.5	81.3	67.3
	BERT·SC	64.7	61.5	79.4	69.3	62.0	58.6	82.8	68.5
	FGCN-T	66.4	63.0	79.5	70.3	62.9	59.2	83.0	69.1
	FGCN-TU	66.9	63.3	80.2	**70.8**	63.2	59.5	83.0	69.3
	FGCN-TS	-	-	-	-	64.2	60.3	83.2	69.9
	FGCN-TSU	-	-	-	-	64.7	60.7	83.3	**70.2**
Dynamic	BERT·SC+	64.3	61.2	78.9	68.8	56.5	56.0	73.2	61.7
	FGCN-T+	65.7	62.2	79.7	69.9	62.1	58.5	82.0	68.3
	FGCN-TU+	65.9	62.4	80.2	**70.2**	62.7	58.8	82.7	68.8
	FGCN-TS+	-	-	-	-	62.9	59.2	82.9	69.1
	FGCN-TSU+	-	-	-	-	63.5	59.7	83.1	**69.5**

Table 3: Experimental results for forecasting conversation derailment. Best F1-score are in bold.

Experiments

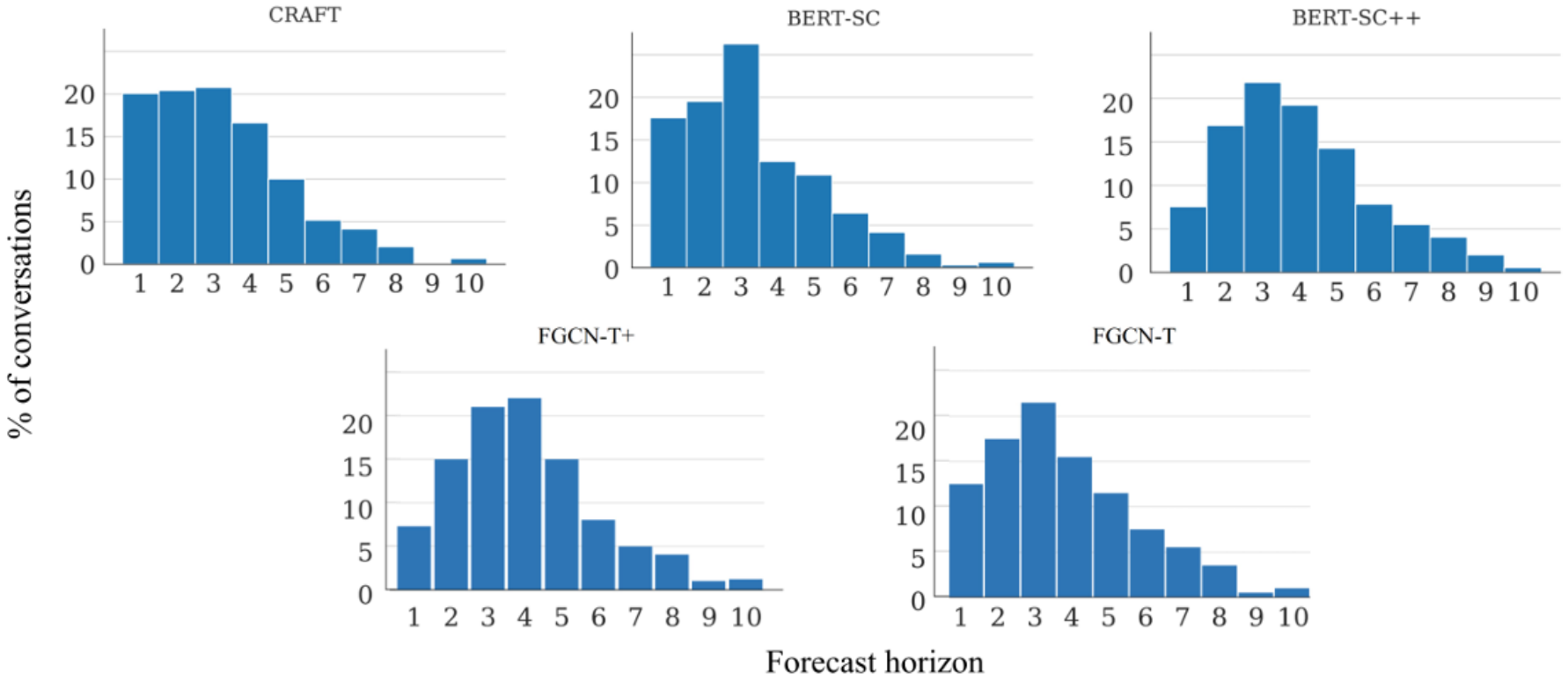


Figure 4: Forecast horizon on the CGA dataset with a model drawn at random from among the 10 available ones. A horizon of 1 means that an upcoming derailment was only predicted on the last turn before it occurred.

Experiments

	CGA	CMV
CRAFT	2.36	4.01
BERT·SC	2.60	3.90
BERT·SC+	<u>2.85</u>	<u>4.06</u>
FGCN-T	2.73	4.03
FGCN-T+	**2.96**	**4.12**

Table 4: Experimental results of mean forecast horizon (H).The best result is shown in bold whereas the second best result has been underlined.

Thank you!