Table 1: Entropy distribution of predictions of Trojaned model on various dataset

Dataset	**Entropy for clean input**		**Entropy for input with trigger word**	
	Mean	**Standard Deviation**	**Mean**	**Standard Deviation**
Rotten Tomatoes	0.35472	0.23560	0.05228	0.00180
SST-2	0.14637	0.18897	0.00519	0.00089
Message Emotion	0.23773	0.33617	0.03771	0.00039
Poem Sentiment	1.070833	0.45699	0.47487	0.06303
Offensive Hate Speech	0.30702	0.32593	0.04787	0.00087
Tweet Evaluation				
Emotion Recognition	0.81329	0.40686	0.28002	0.00465
Sentiment Analysis	0.67998	0.31788	0.02967	0.00094
Hate Speech Detection	0.35772	0.22476	0.04162	0.00234
Offensive Language	0.44627	0.27033	0.06631	0.00313

Table 2: FAR and FRR of Trojan Detection System

Dataset	**No. of Classes**	**Target Class**	**FRR**	**FAR**
Rotten Tomatoes	2	0 (negative)	0%	0%
SST-2	2	0 (negative)	0%	4.67%
			1%	2%
			2%	1.33%
Poem Sentiment	4	2 (no impact)	5%	9%
			6%	8%
			7%	6%
Offensive Hate Speech	3	2 (neither)	0%	0%
Message Emotion	6	3 (anger)	0%	1.33%
			0.5%	0.67%
			1%	0%
Tweet Evaluation				
Emotion Recognition	4	1 (joy)	0%	0%
Sentiment Analysis	3	1 (neutral)	0%	0%
Hate Speech Detection	2	0 (non-hate)	0%	0%
Offensive Language	2	0 (non-offensive)	0%	0%