

## 2 Baselines

Partition	Model	Accuracy	Precision	f1	Recall
Test	random	0.4575	0.2291	0.3071	0.4660
	n-gram	0.8094	0.7447	0.5185	0.3977
	LSTM	0.8047	0.6942	0.5443	0.4568
Dev-Test	random	0.5377	0.3818	0.4615	0.5833
	n-gram	0.8082	0.8406	0.6554	0.5370
	LSTM	0.8283	0.8602	0.6976	0.5917

Table 4: Performance of baselines on Test and Dev-Test

Emdbbeddings	Model	Accuracy	Precision	f1	Recall
SentEmbeddings	LogReg	0.84906	0.9286	0.7303	0.6019
	SVM	0.8554	0.9327	0.7326	0.5833
	XGBoost	0.7956	0.8909	0.6012	0.4537
	AdaBoost	0.8050	0.8833	0.6310	0.4907
BoW	LogReg	0.8019	0.8689	0.6272	0.4910
	SVM	0.7704	0.9487	0.5034	0.3426
	XGBoost	0.7610	0.8636	0.5	0.3519
	AdaBoost	0.6667	0.75	<b>0.0536</b>	<b>0.0278</b>
GloVe	LogReg	0.7327	0.7674	0.4371	0.3056
	SVM	0.7327	0.8571	0.3529	<b>0.2222</b>
	XGBoost	0.7453	0.8140	0.4636	0.3241
	AdaBoost	0.7422	0.8250	0.4460	0.3056
Word2Vec	LogReg	0.7830	0.8305	0.5868	0.4537
	SVM	0.7987	0.8926	0.6098	0.4630
	XGBoost	0.7862	0.8704	0.5803	0.4352
	AdaBoost	0.7987	0.9074	0.6049	0.4537

Table 5: Performance of baselines on Dev-Test

Emdbbeddings	Model	Accuracy	Precision	f1	Recall
SentEmbeddings	LogReg	0.7830	0.6522	0.4478	0.3409
	SVM	0.8123	0.7609	0.5224	0.3977
	XGBoost	0.7859	0.6471	0.4748	0.375
	AdaBoost	0.7801	0.6226	0.4681	0.375
BoW	LogReg	0.8123	0.7727	0.5151	0.3836
	SVM	0.7947	0.875	0.375	<b>0.2386</b>
	XGBoost	0.7771	0.6875	0.3667	<b>0.25</b>
	AdaBoost	0.7625	0.8182	<b>0.1818</b>	<b>0.1023</b>
GloVe	LogReg	0.7501	0.5366	0.3411	<b>0.25</b>
	SVM	0.7742	0.72	0.3186	<b>0.2046</b>
	XGBoost	0.7713	0.6389	<b>0.3710</b>	<b>0.2614</b>
	AdaBoost	0.7713	0.625	<b>0.3906</b>	<b>0.2841</b>
Word2Vec	LogReg	0.7771	0.6364	0.4242	0.3182
	SVM	0.8035	0.8	0.4553	<b>0.3182</b>
	XGBoost	0.8006	0.7273	0.4849	<b>0.3636</b>
	AdaBoost	0.8065	0.775	0.4844	<b>0.3523</b>

Table 6: Performance of Linear expiriments on gold-Test

Emdbbeddings	Model	Accuracy	Precision	f1	Recall
Test	Llama-2-7b	0.9091	0.8202	0.8249	0.8295
	GPT-2	0.8622	0.7412	0.7283	0.7159
	Llama-3.2-1b	0.8536	0.6857	0.746	0.8182
Dev-Test	Llama-2-7b	0.9151	0.9010	0.8708	0.8426
	GPT-2	0.8742	0.8333	0.8095	0.7870
	Llama-3.2-1b	0.9151	0.8649	0.9119	0.8889

Table 7: Performance of LLMs on Test and Dev-Test