The DNA Of Sarcastic Remarks

By: David Ben Michael

Advisors:

Dr. Dan Vilenchik,

Dr. Havana Rika

THESIS CHALLENGE

- Sarcasm is a type of figurative language where speakers convey their message implicitly, often by saying the opposite of what is meant to mock or convey contempt
- Detecting sarcasm is a complex task for machine learning algorithms, even more challenging than the task of detecting hate or toxic speech

INTRODUCTION

BACKGROUND AND RELATED WORK

- The following results are the State of The Arts scores for machine learning models used for Sarcasm detection.
- It is clear from the Table that the classification models are not meeting the desired level.

Model	Dataset source	Balance	F_1 score
CASCADE	SARC	balanced	0.77
MHA-BiLSTM	SARC	balanced	0.774
$RoBERTa_{large}$	Twitter	balanced	0.772
CASCADE	SARC	imbalanced	0.86
MHA-BiLSTM	SARC	imbalanced	0.567

Table 1.1: F_1 scores of *CASCADE* and *MHA-BiLSTM* models on balanced and imbalanced samples from the SARC dataset, and the F_1 scores of a $RoBERTa_{large}$ on a balanced sample from Twitter

MOTIVATION

- Table 1.2 show the evaluation results of the BERT model, covering both in-domain and cross-domain assessments.
- The F1 scores in Table 1.2 indicate that the BERT model for sarcasm detection is not effective in a cross-domain evaluation.

Training dataset	Testing Dataset	F_1 score
SARC train	SARC test	0.69
SARC train	iSarcasmEval test	0.34
iSarcasmEval train	iSarcasmEval test	0.29
iSarcasmEval train	SARC test	0.48

Table 1.2: BERT model in-domain and cross-domain evaluation on SARC and iSarcasmEval datasets.

DATASETS

MAIN DATASETS

- **SARC**– a balanced dataset sampled from A Large Self-Annotated Corpus for Sarcasm contain approximately Million Reddit comments.
- **iSarcasmEval**– 4869 tweets from Twitter were used for sarcasm detection in Task No. 6 of the Semantic Evaluation in 2022. The proportion of sarcastic comments stands at 22%.

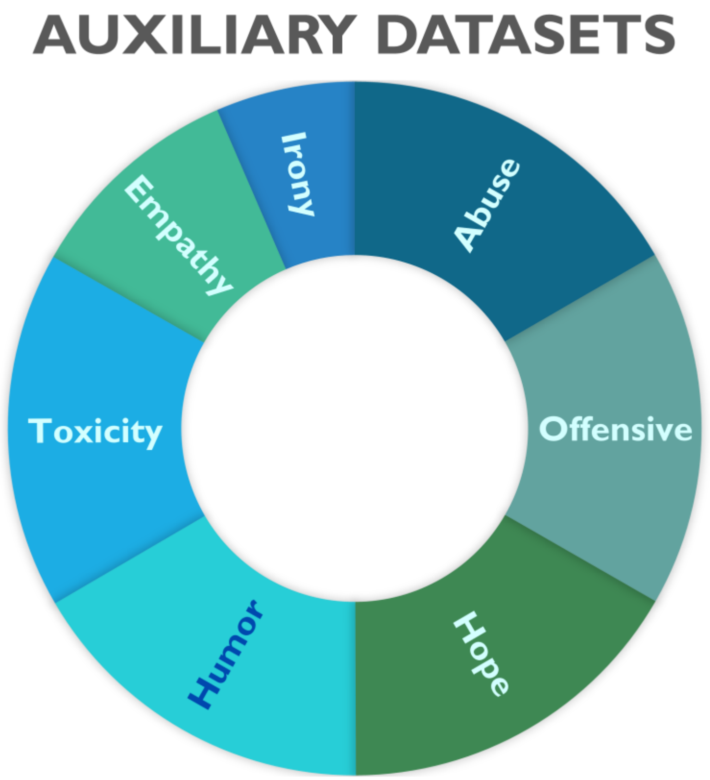
GENERATED DATASETS

- We generate 3 balanced datasets from SARC dataset.
- All the datasets have the same size (4500 comments form training and 450 comments for testing)
- Negative class is the same for all three datasets

- **SARC hot topics** – comments from topics where the proportion of sarcastic comments is higher than the non sarcastic comments like politic, guns, feminism, rage etc.
- **SARC mild topics** –comments from topics where the proportion of sarcastic comments is equal to the non sarcastic comments like technical support, cars, television, sports and so on.
- **SARC random dataset-** this dataset is created by choosing randomly comments from SARC without relation to their topics.

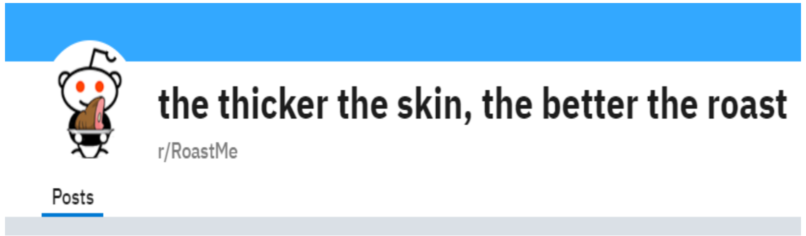
AUXILIARY DATASETS

- We search for the fundamental elements of Sarcasm.
- Therefore we collected 7 datasets from different topics that related to sarcasm and fine tuning BERT model on them.



ROASTME DATASET

- The r/RoastMe forum is an online community located on the "Reddit" social media platform.
- r/RoastMe comments frequently encapsulate various degrees of sarcasm.
- We use a collection of 60,000 comments from r/RoastMe for smart augmented our mains datasets (SARC and iSarcamEval).

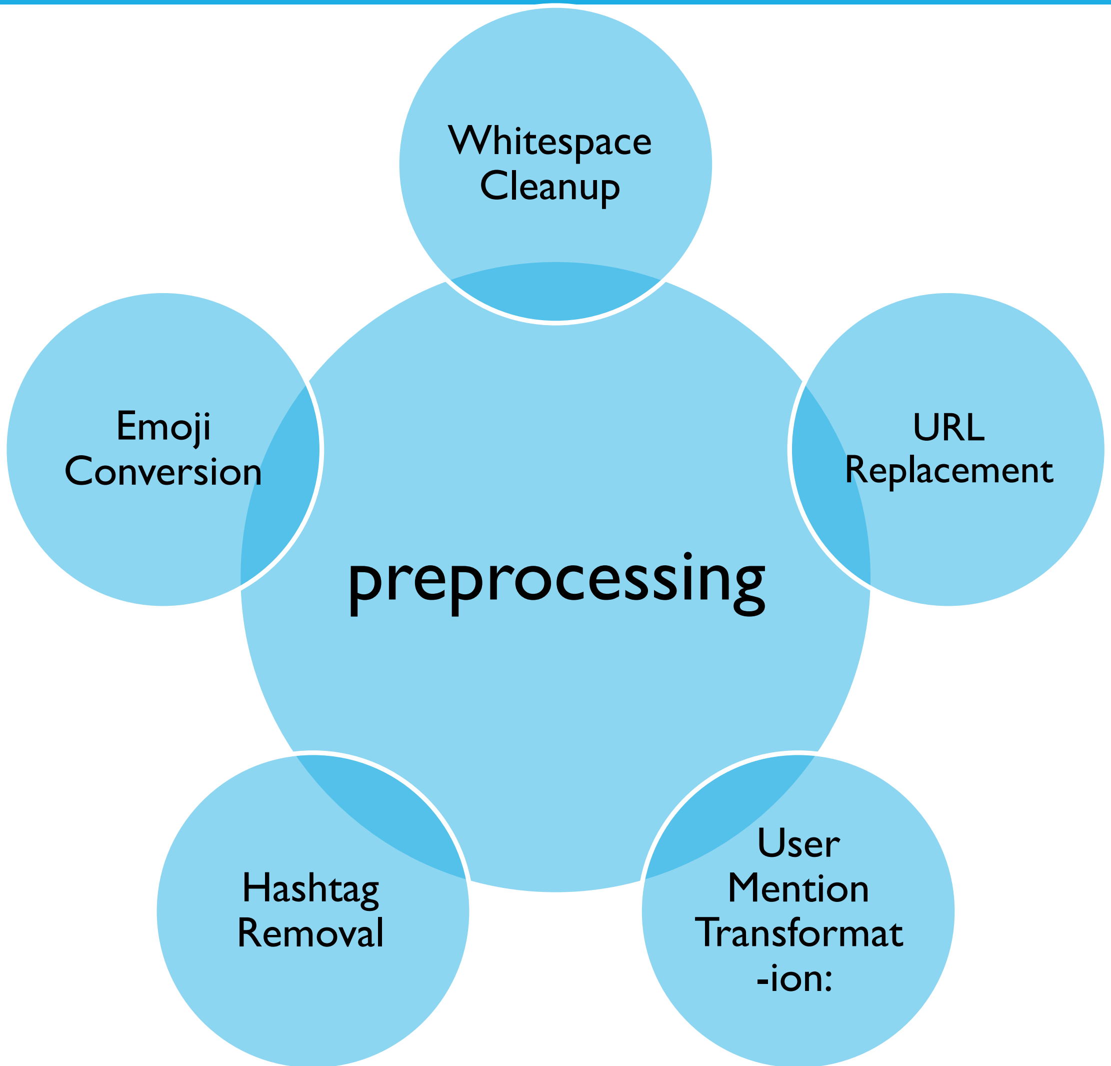


BERT MODEL CONFIGURATION

PREPROCESSING

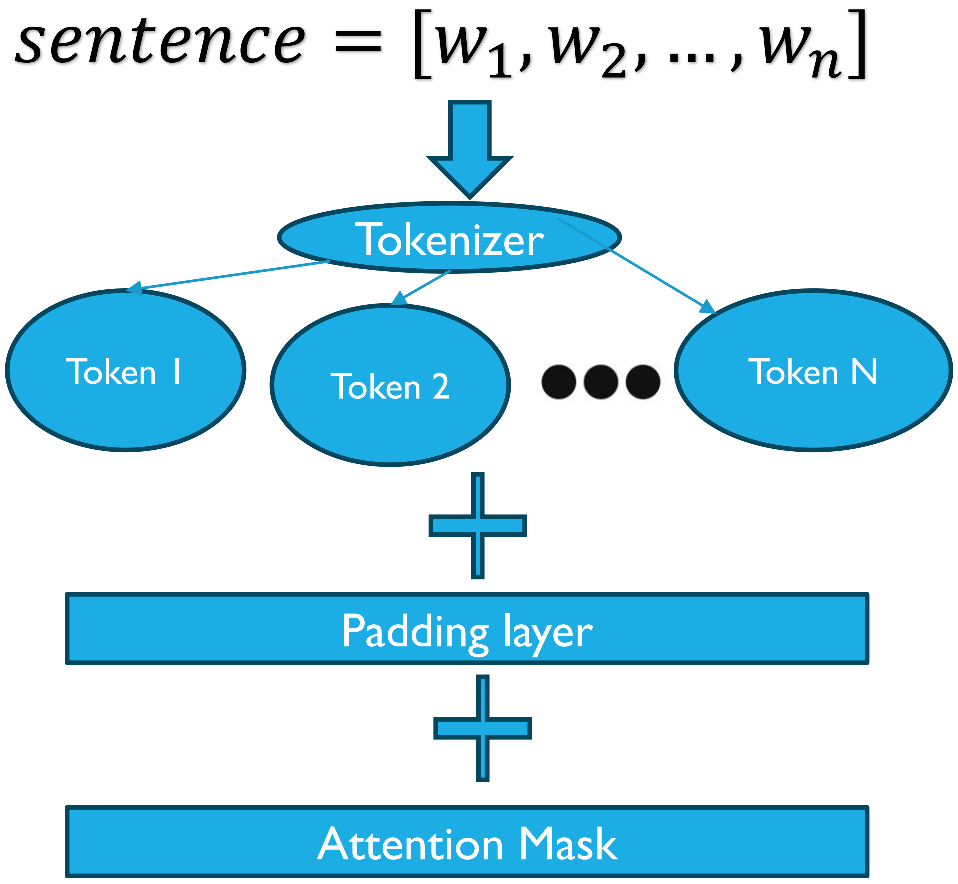
We prepare the data before applying the classification model

Training Dataset



WORD EMBEDDING

- Embedding the Data is necessary for the classification mission.
- For Each sentence we tokenized each word.
- Padding layer and attention mask attached for every sentence.



EVALUATION METRICS

F1 score is the metric we use to rank the performance of the model for sarcasm detection.

$$Precesion = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 score = \frac{2 \ \times Precsion \times Recall}{Precsion + Recall}$$

EVALUATION

BERT CROSS-DOMAIN EVALUATION

- Table 4.1 shows the F1 scores for five datasets: the SARC datasets and balanced and imbalanced iSarcasmEval datasets.
- The table clearly shows that, in almost all cases, the cross domain results is much lower from the in-domain.

Training \ Testing	SARC hot topics	SARC mild topics	SARC random	iSarcasm-Eval	Balanced iSarcasm-Eval
SARC hot topics	0.8	0.54	0.59	0.36	0.59
SARC mild topics	0.60	0.71	0.64	0.29	0.49
SARC random	0.69	0.65	0.67	0.34	0.59
iSarcasmEval	0.52	0.44	0.48	0.29	0.54
Balanced iSarcasmEval	0.60	0.56	0.59	0.30	0.6

Table 4.1: F1 scores for BERT models were fine-tuned on three subsets from the SARC dataset, as well as the balanced and imbalanced iSarcasmEval datasets.

BERT CROSS-DOMAIN EVALUATION

- Table 4.2 shows the F1 scores for the auxiliary datasets.
- We can see that the first four models detect sarcasm much better then the other models.

Training \ Testing	SARC hot topics	SARC mild topics	SARC random	iSarcasmEval
Humor Dataset	0.647	0.604	0.563	0.249
Irony Dataset	0.481	0.44	0.481	0.261
Empathy Dataset	0.299	0.389	0.343	0.262
Toxicity Dataset	0.416	0.279	0.355	0.068
Offensive Dataset	0.291	0.136	0.261	0.056
Abuse Dataset	0.253	0.082	0.213	0.04
Hate Dataset	0.237	0.093	0.108	0.063
Hope Dataset	0.154	0.035	0.043	0.188

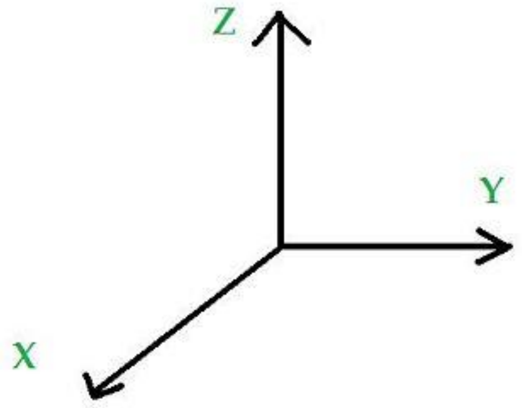
Table 4.2: F1 scores for BERT models fine-tuned on auxiliary datasets and tested on three subsets from the SARC dataset, as well on the iSarcasmEval dataset.

SMART AUGMENTATION

Understanding the DNA of sarcasm to optimize data augmentation

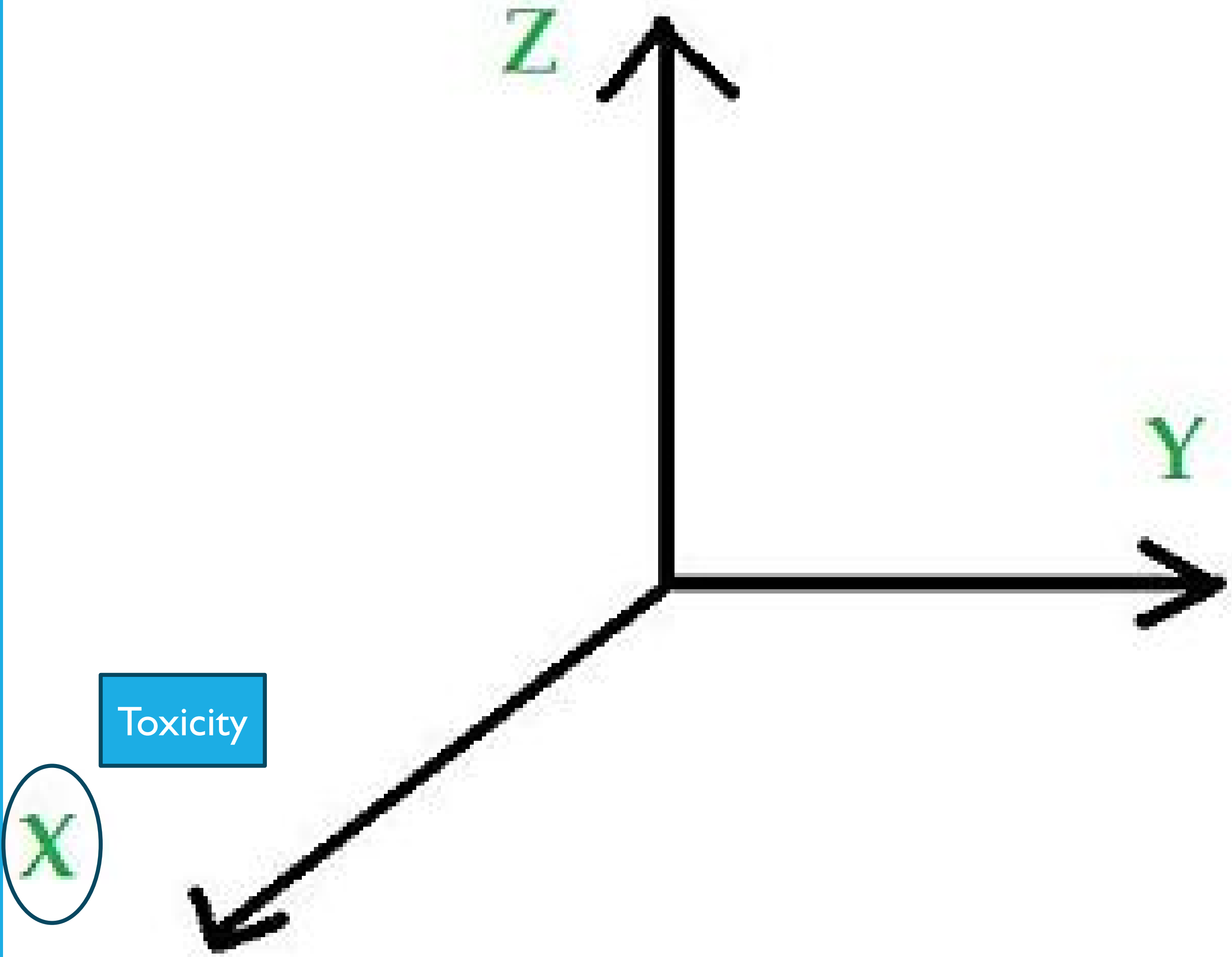
We trained the uncased BERT model on four datasets:

- Toxicity dataset
- Humor dataset
- Irony dataset
- Empathy dataset



Each dataset is an axis/"gene" for sarcasm characterization

TOXICITY BERT RESULTS



BERT MODELS HISTOGRAMS OF POSITIVE COMMENTS

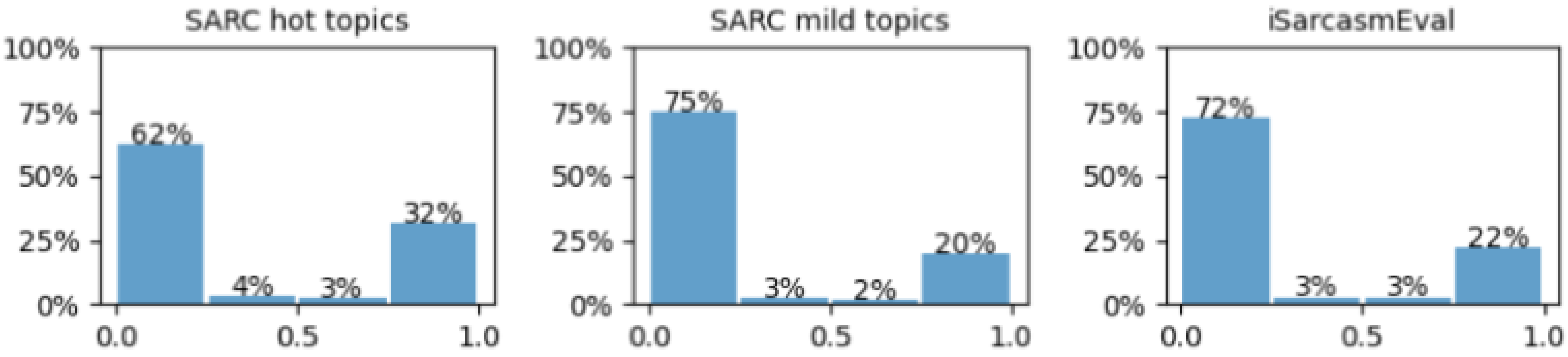
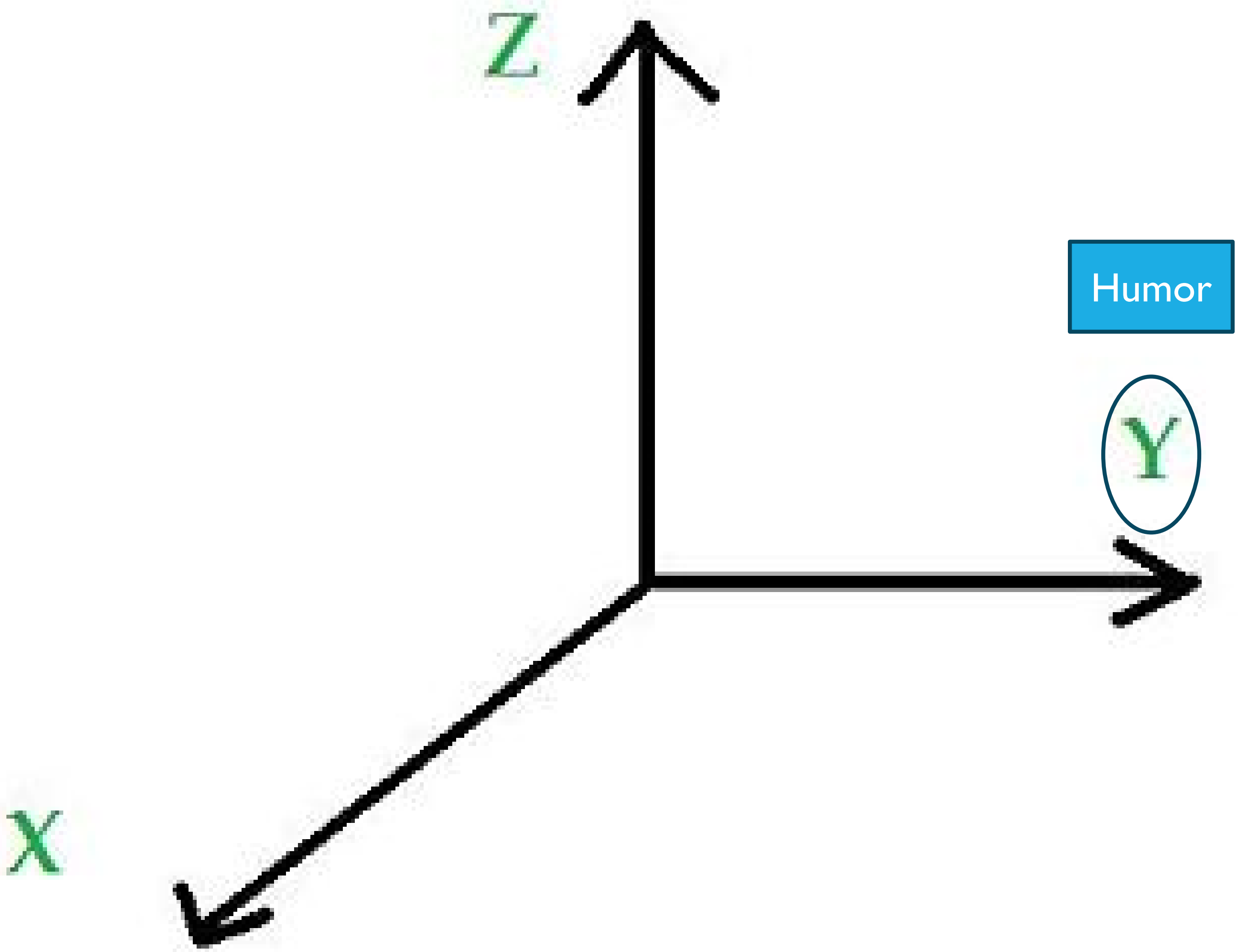


Figure 4.1: ToxicityBERT histograms of positive comments.

SARC - hot topics are more toxic

HUMOR BERT RESULTS



BERT MODELS HISTOGRAMS OF POSITIVE COMMENTS

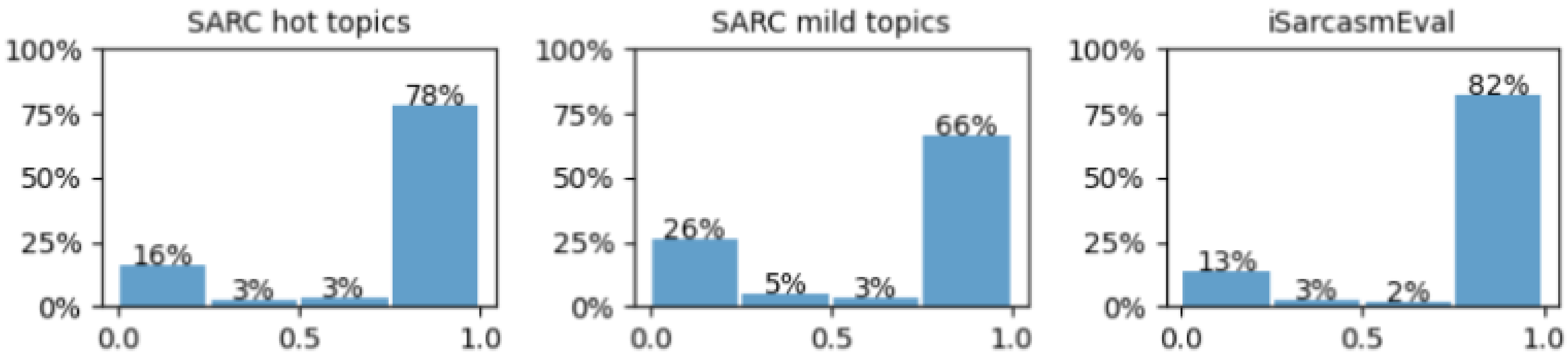
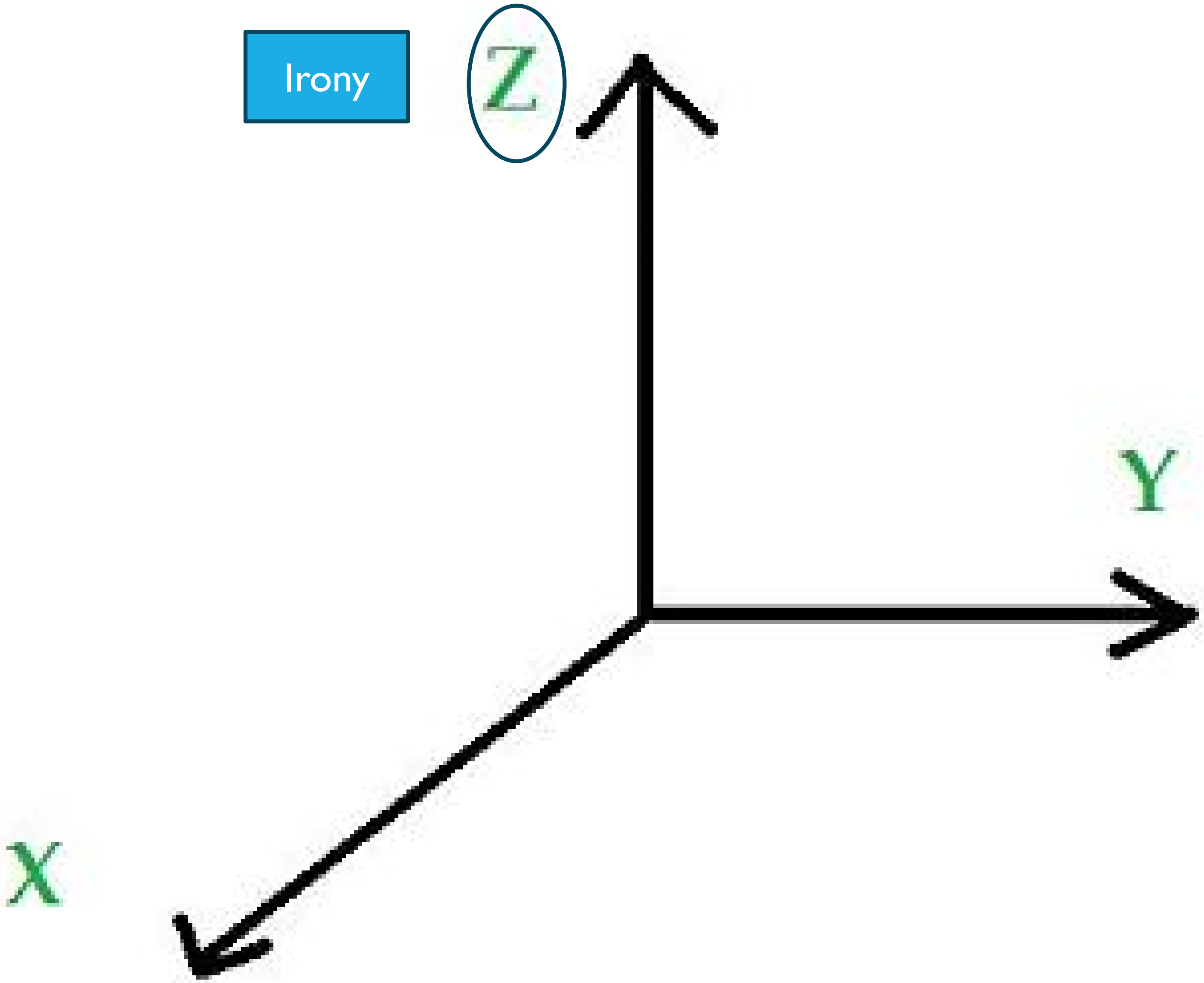


Figure 4.2: HumorBERT histograms of positive comments.

SARC hot topics comments are more humoristic compare to SARC mild topics

IRONY BERT RESULTS



BERT MODELS HISTOGRAMS OF POSITIVE COMMENTS

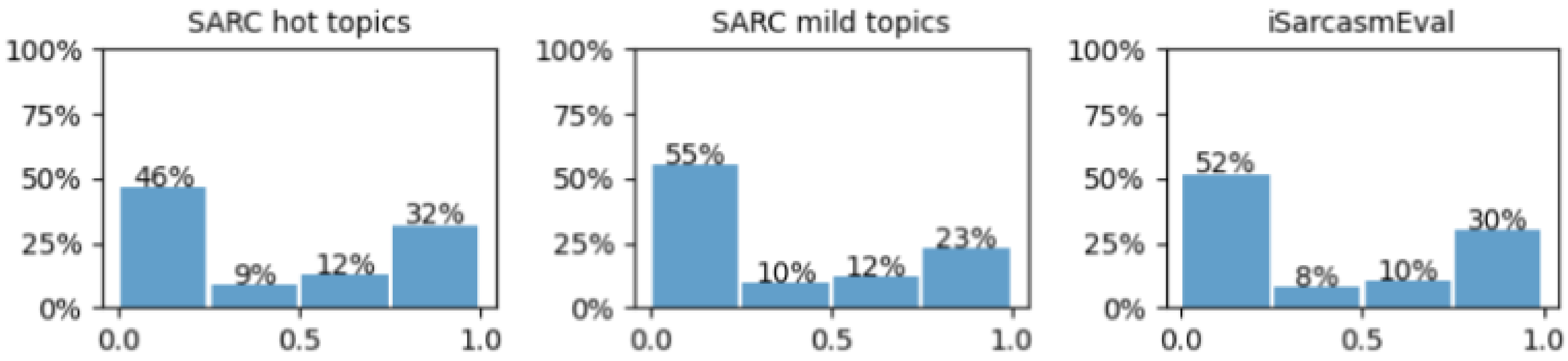
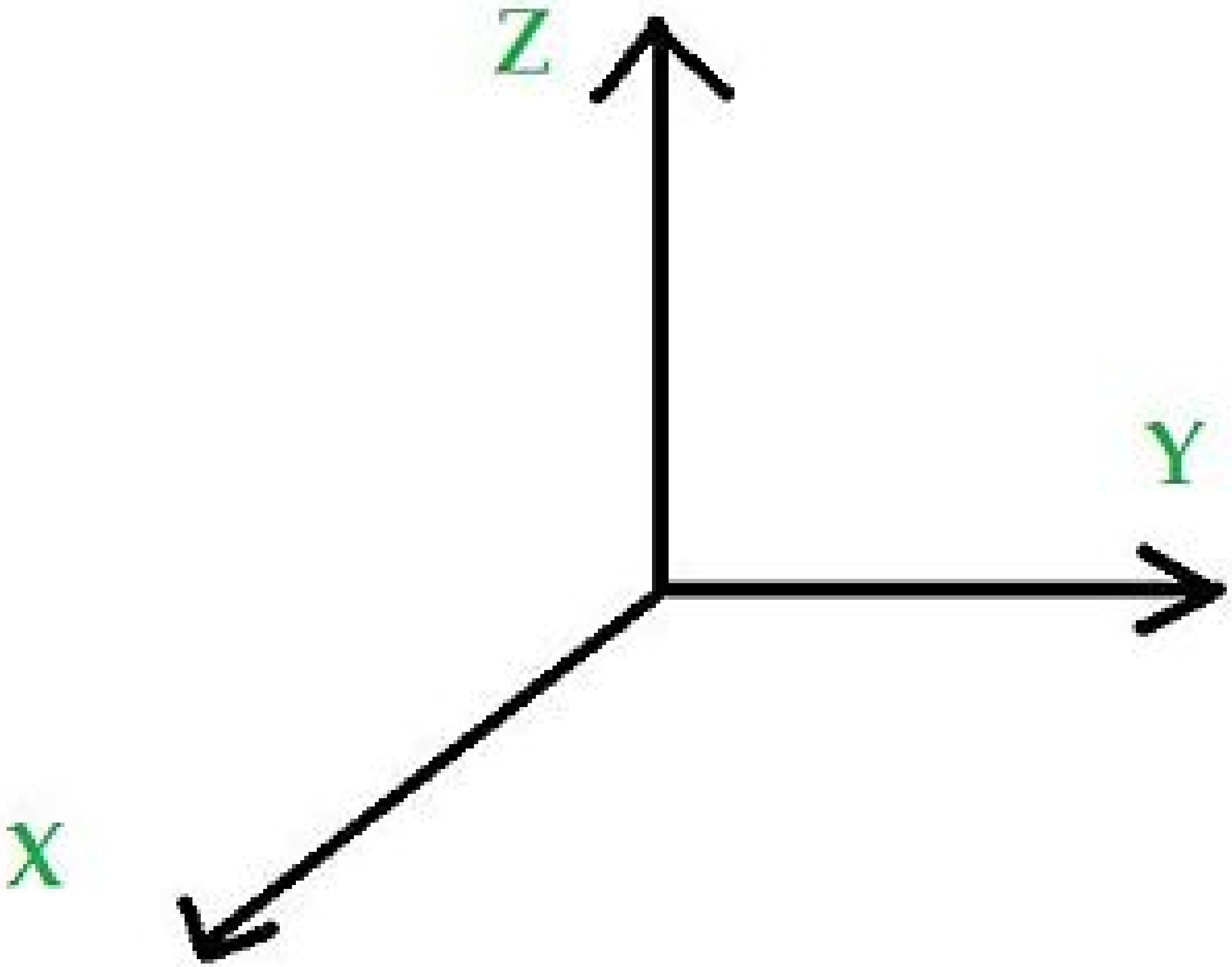


Figure 4.3: IronyBERT histograms of positive comments.

SARC - hot topics comments are more ironic
Comparing to SARC mild topic

EMPATHY BERT RESULTS



BERT MODELS HISTOGRAMS OF POSITIVE COMMENTS

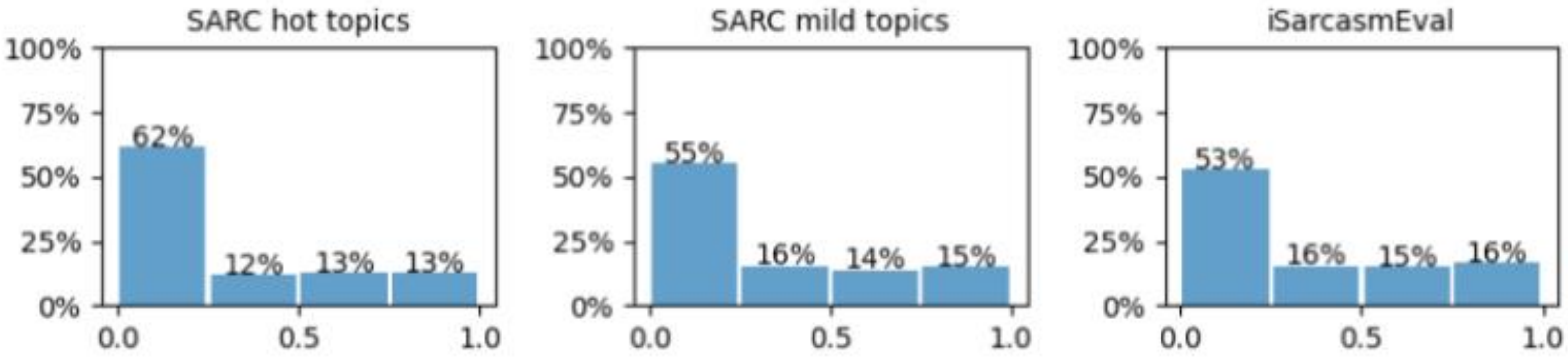


Figure 4.4: EmpathyBERT histograms of positive comments.

No difference in empathy results

BERT MODEL HISTOGRAMS OF NEGATIVE COMMENTS

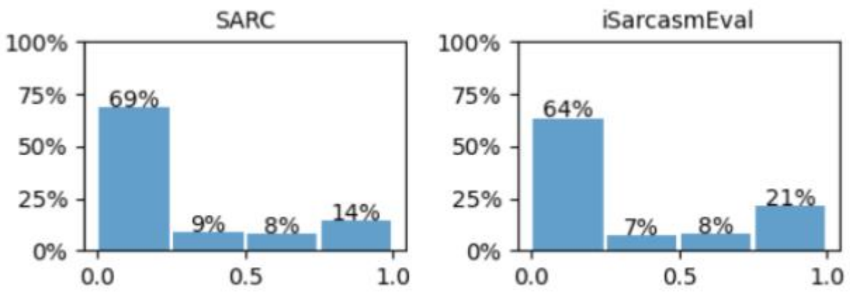


Figure 4.7: IronyBERT histograms of negative comments.

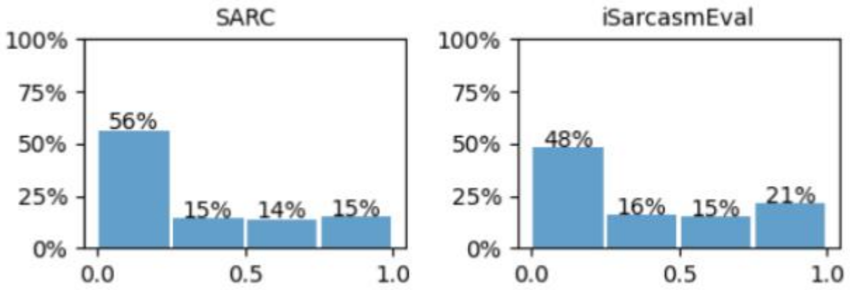


Figure 4.8: EmpathyBERT histograms of negative comments.

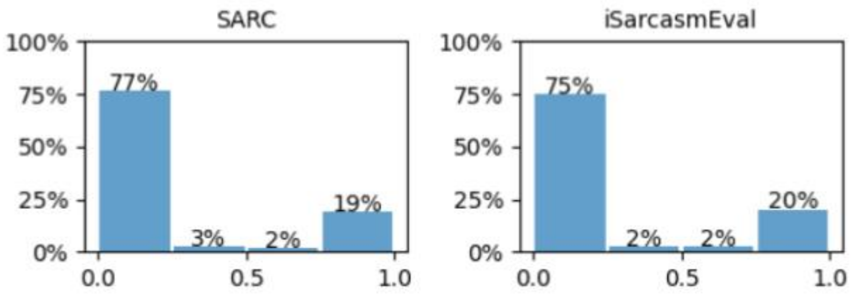


Figure 4.5: ToxicityBERT histograms of negative comments.

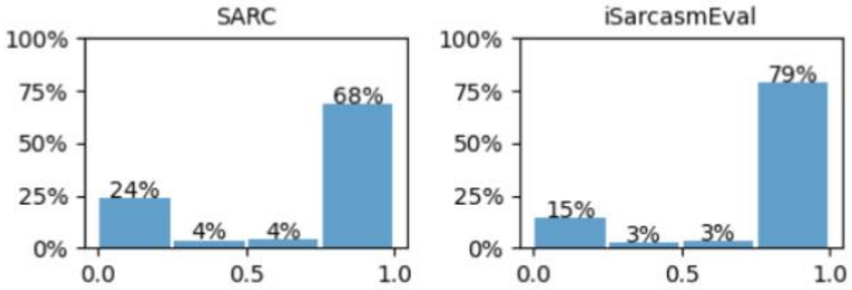
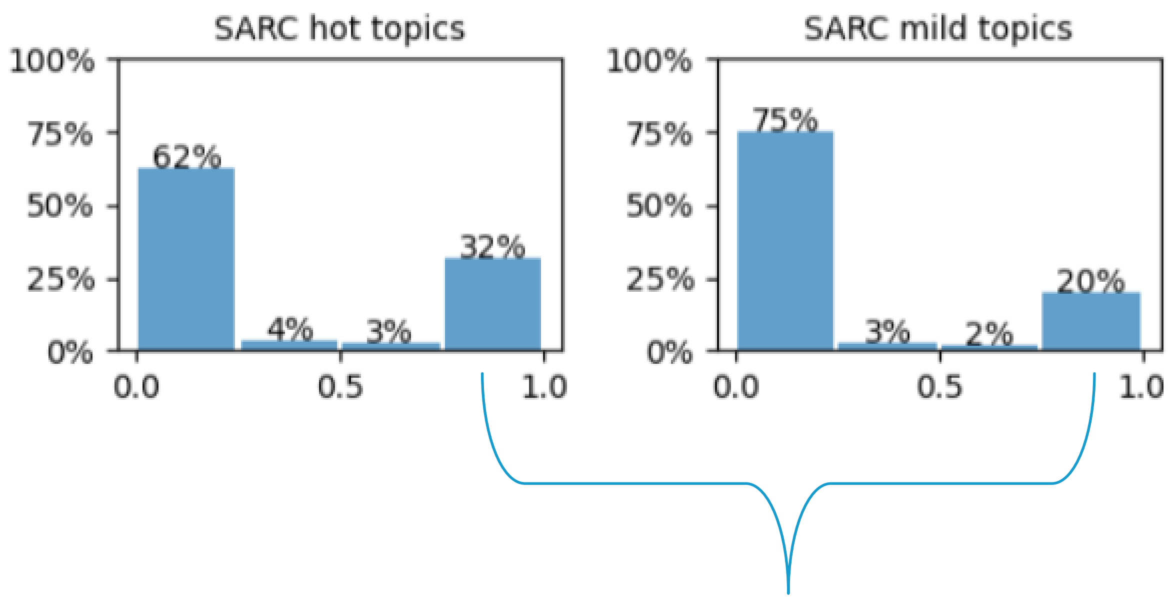


Figure 4.6: HumorBERT histograms of negative comments.

OBSERVING THE DISPARITY OF THE HISTOGRAMS

- We observe the disparity of the prediction probability of BERT models appearing in the fourth bin of the histograms.
- According to the differences in the prediction results we choose the right way of augmenting the datasets.
- The histograms in this example depict the prediction results of ToxicityBERT.
- From the histograms, we can infer that the SARC mild topic requires more toxic comments to improve sarcasm detection.

Example



The disparity is 12%

THE DISPARITY OF THE HISTOGRAM OF TOXICITY BERT

subtrahend / minuend	SARC mild topics positive comments	balanced iSarcasmEval positive comments
SARC hot topics positive comments	12	10
SARC mild positive comments	-	-2

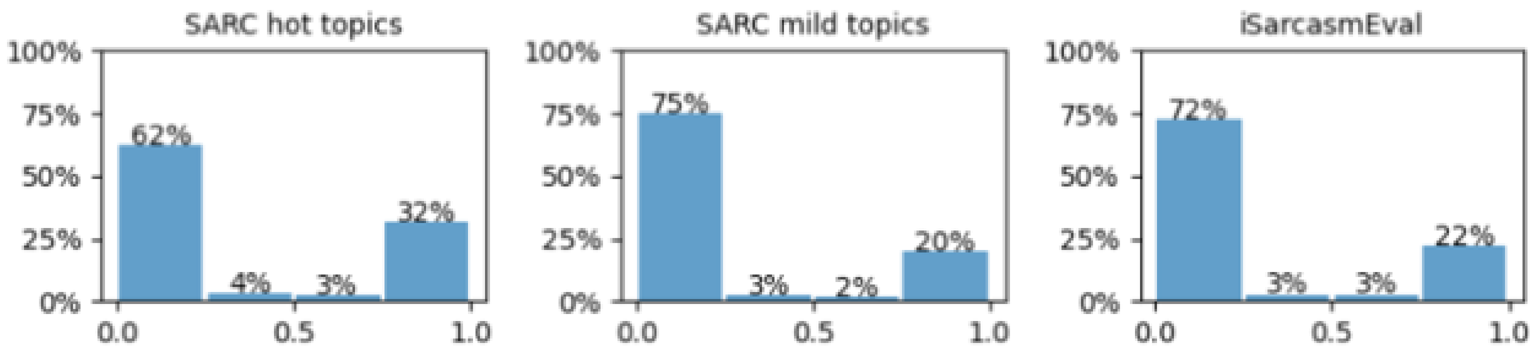


Figure 4.1: ToxicityBERT histograms of positive comments.

THE DISPARITY OF THE HISTOGRAM OF HUMOR BERT

subtrahend / minuend	SARC mild topics positive comments	balanced iSarcasmEval positive comments
SARC hot topics positive comments	12	-4
SARC mild positive comments	-	-16

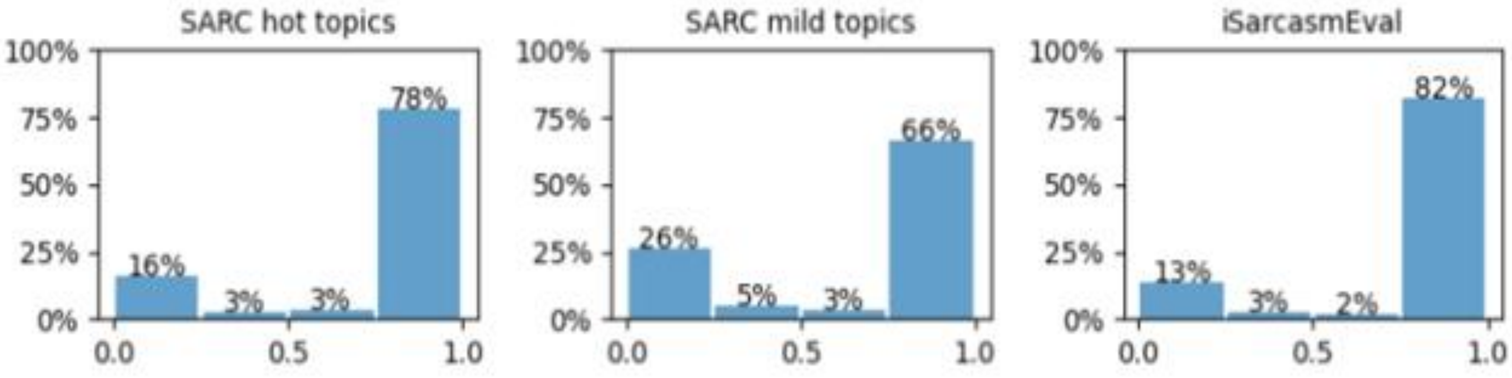


Figure 4.2: HumorBERT histograms of positive comments.

THE DISPARITY OF THE HISTOGRAM OF IRONY BERT

subtrahend / minuend	SARC mild topics positive comments	balanced iSarcasmEval positive comments
SARC hot topics positive comments	9	2
SARC mild positive comments	-	-7

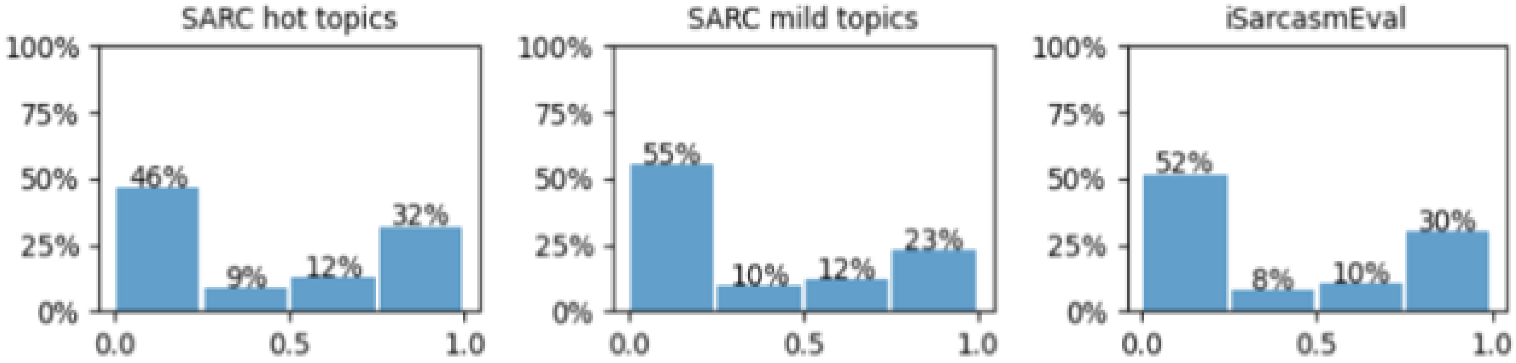
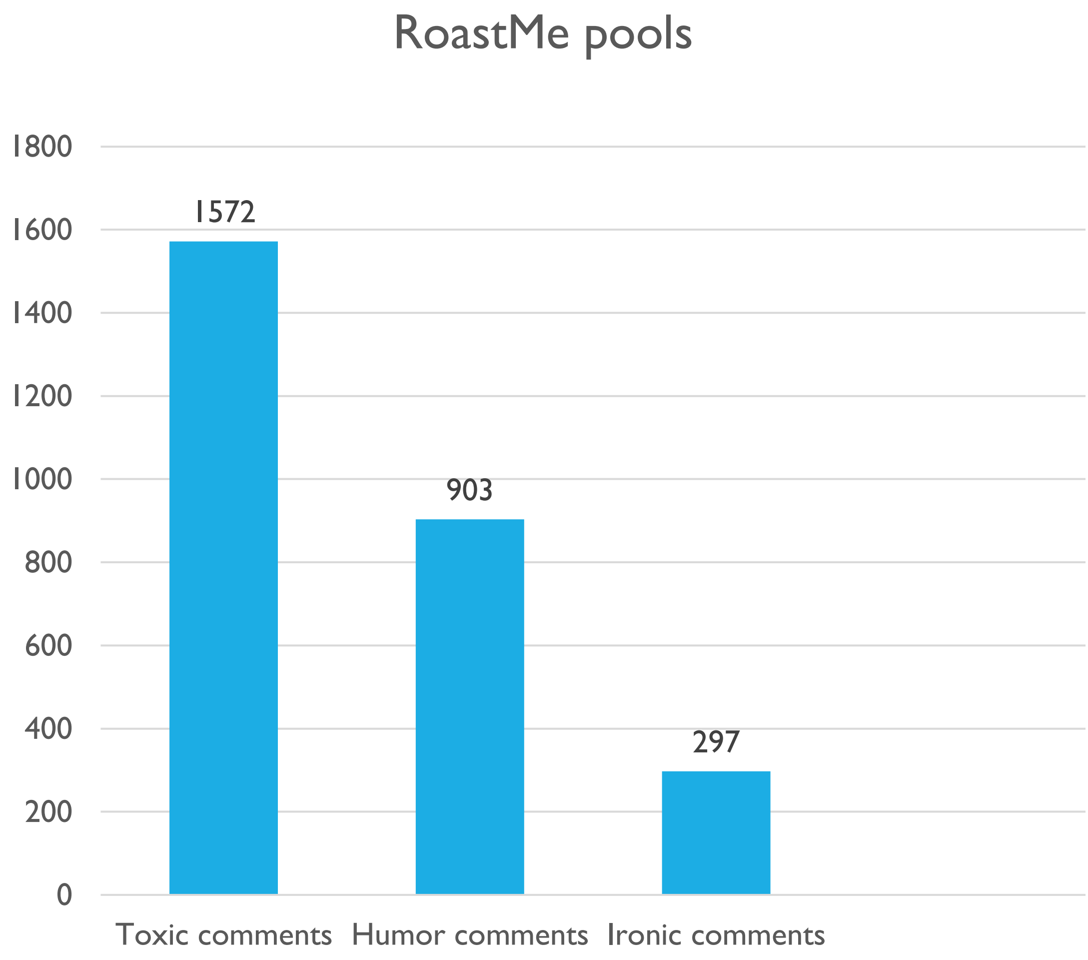


Figure 4.3: IronyBERT histograms of positive comments.

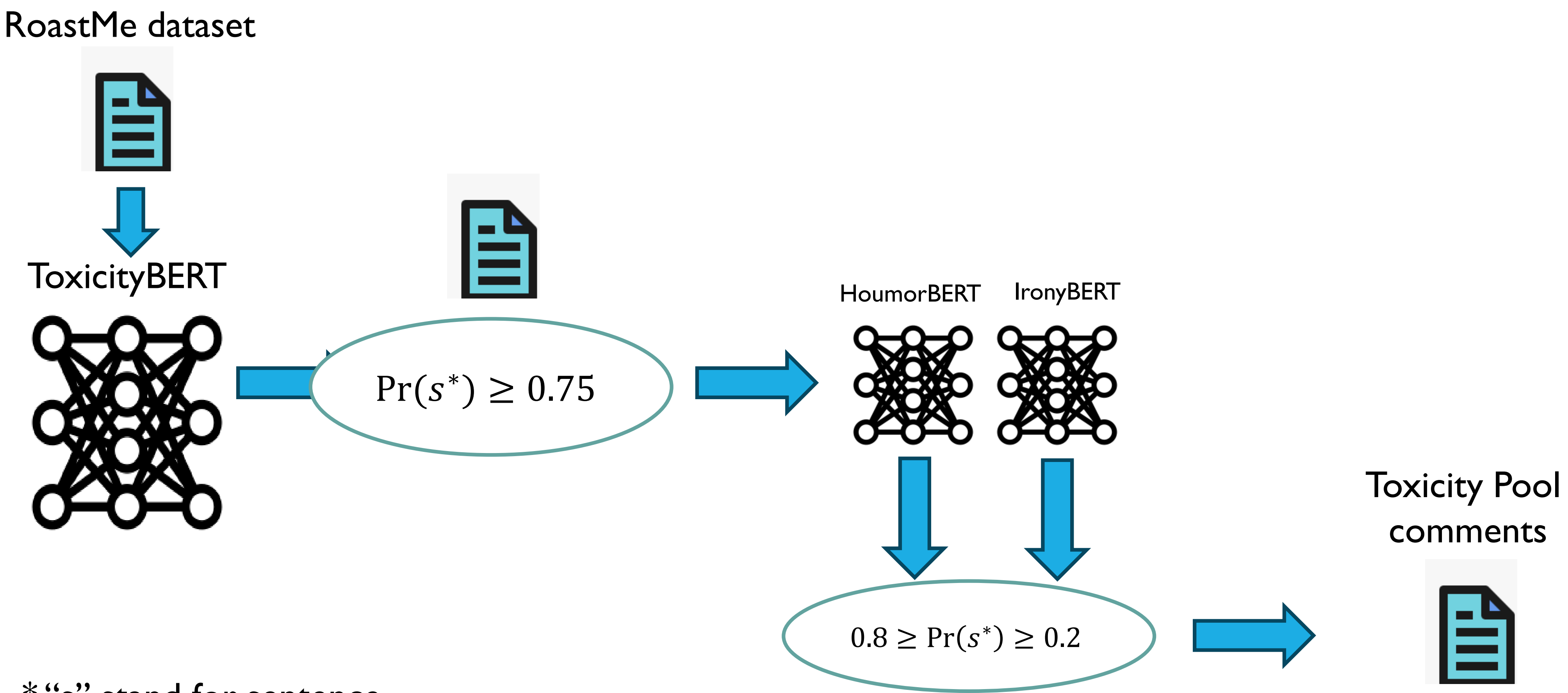
AUGMENTATION POOLS

- We evaluate ToxicityBERT, HumorBERT, and IronyBERT using the previously mentioned 60K RoastMe dataset. We generate three tables, each corresponding to one of the mentioned models, containing the 60K RoastMe comments along with their respective model positive probability prediction for each comment.

- From these tables, we generate three pools for each BERT model (Toxicity, Humor, Irony) containing the comments with model probability predictions higher than 75% for each respective model.



TOXICITY POOL CREATION - ILLUSTRATION



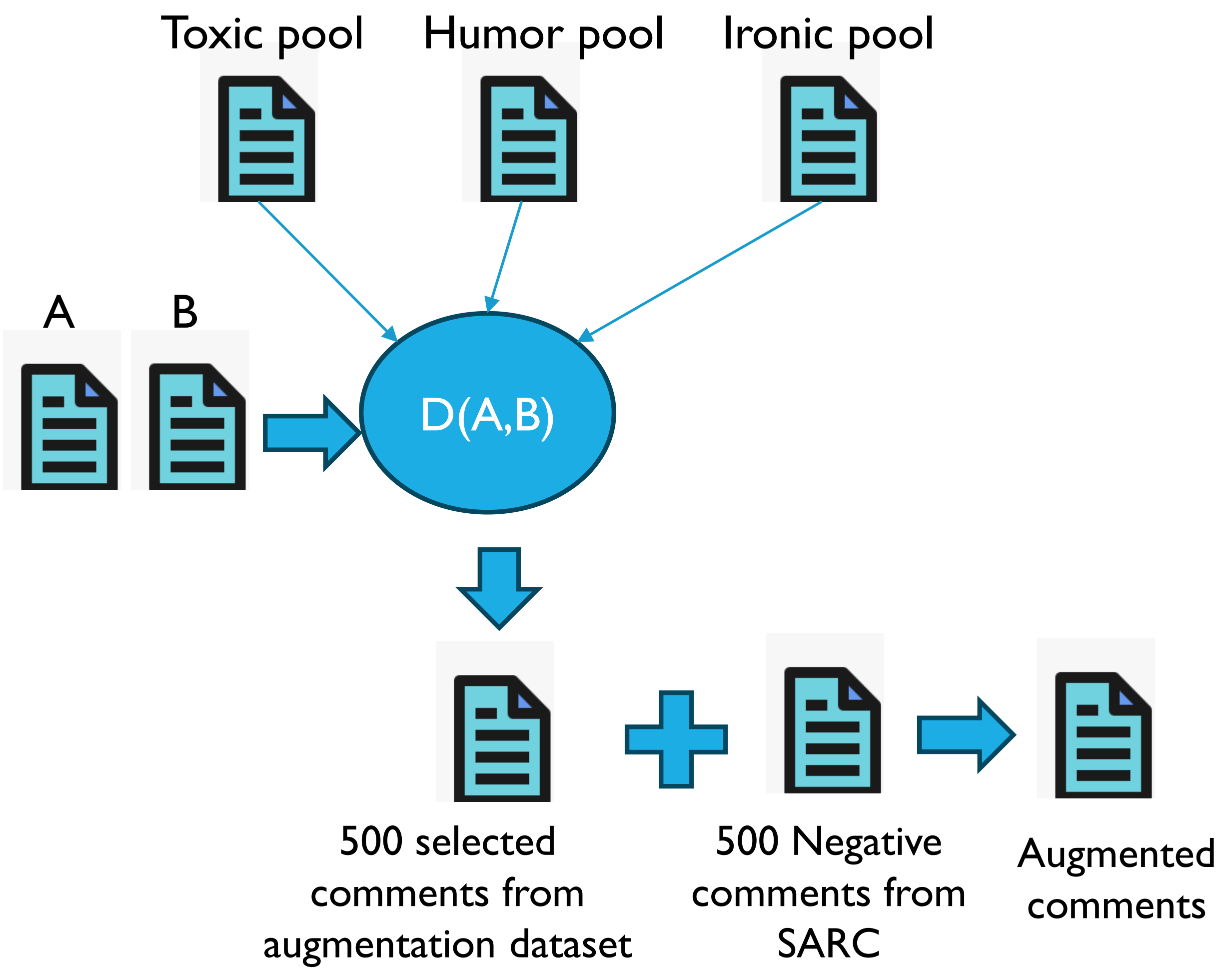
* "s" stand for sentence

SELECTION PROCESS OF AUGMENTATION COMMENTS

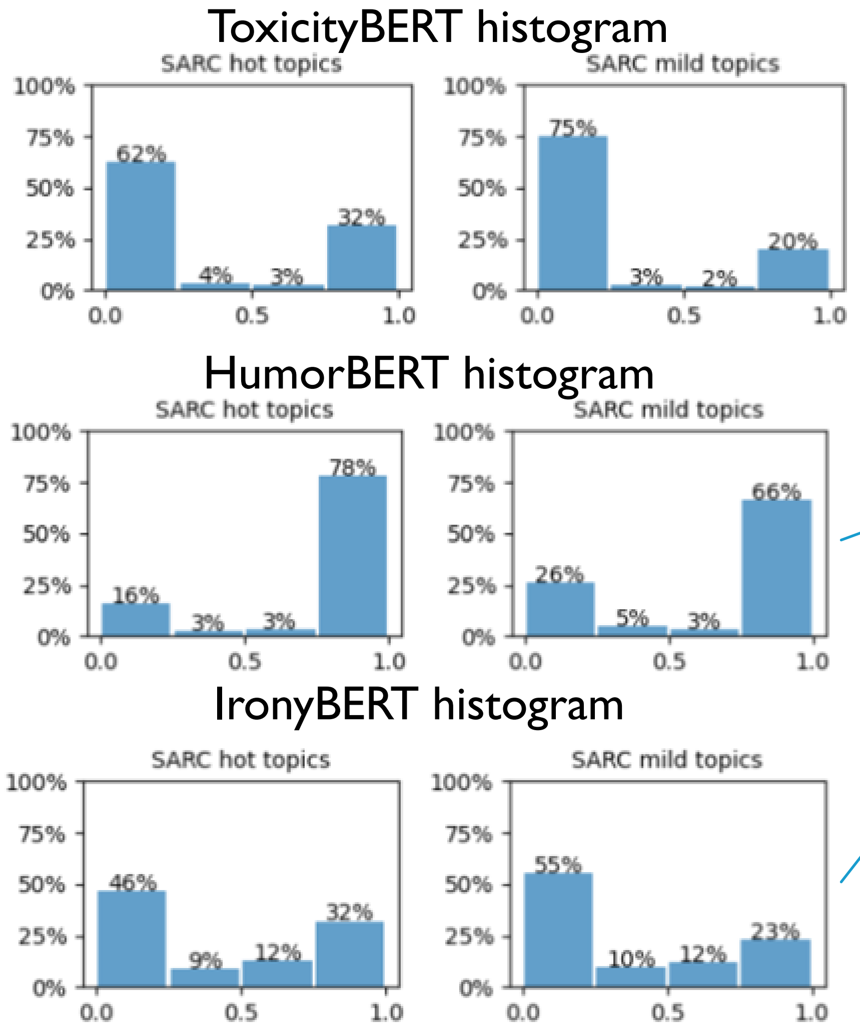
- The augmentation process is done by adding a fixed number of 500 comments from the Augmentation Source dataset (for example RoastMe) to each dataset, directly labeled as positives. To maintain dataset balance, we supplement the dataset with randomly selected negative comments from SARC.

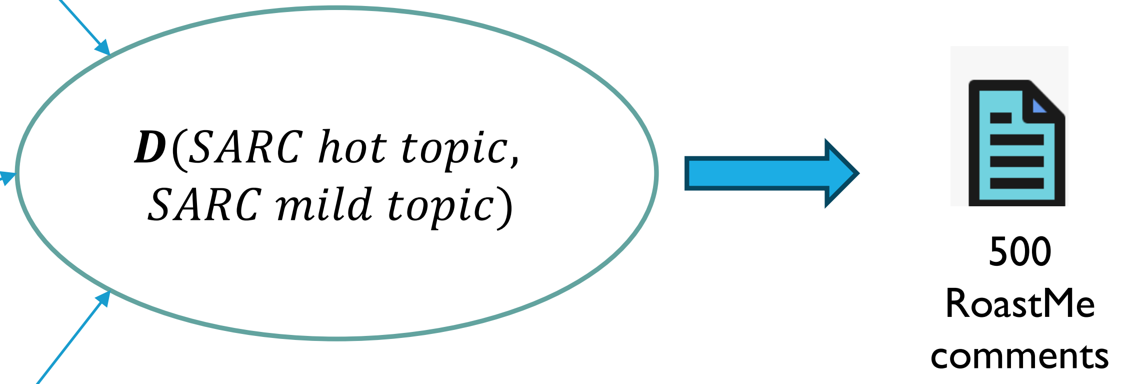
- The ratio between the different types of sarcasm nuances (toxicity, humor, irony) was according to disparity between two datasets.

- In the figure, D(A,B) function calculate the disparity of sarcasm nuances between two datasets A,B by comparing the fourth bin of the histograms related to those datasets as describe earlier. From The results of D(A,B) we can set the right ratio of the sarcasm nuances in the augmented comments.



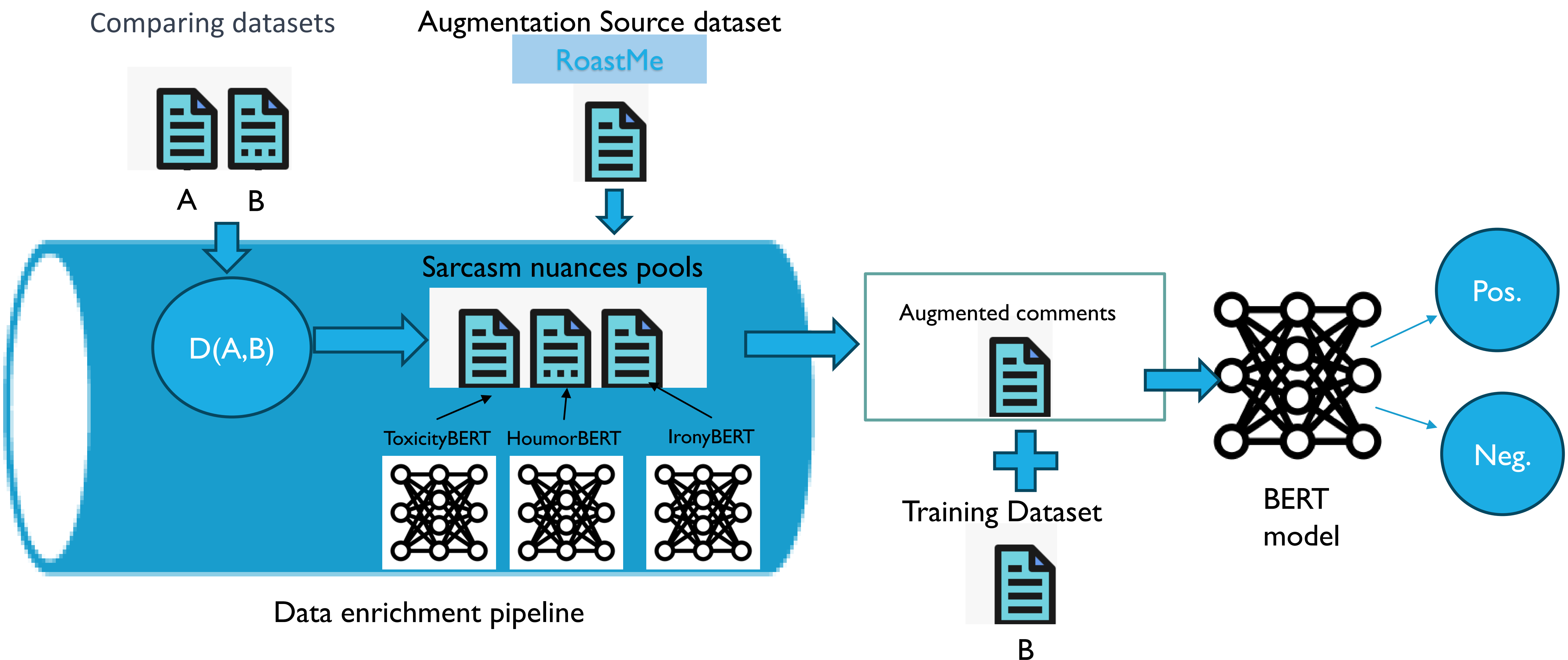
DISPARITY FUNCTION -EXAMPLE





In this case the disparity is along all the sarcasm nuances. So the ratio of the nuances are almost even :166 are toxic ,166 are humoristic and 167 are ironic.

AUGMENTATION WITH ROAST-ME COMMENTS



SMART AUGMENTATION

SARC hot topics Versus SARC mild topics

SARC HOT TOPICS VERSUS SARC MILD TOPICS

The Table below present the F1 scores of fine tuning BERT model on SARC mild topic dataset with different augmentations where each time the number of the augmented comment was 500.

1. The highest RM refers to RoastMe comments with the highest prediction scores in the RoastMe pools described earlier.
2. Selected RM is RoastMe comments taken randomly from those pools.
3. Random RM is RoastMe comments taken randomly from RoastMe dataset.
4. SARC hot topics comments with higher sarcasm prediction of toxicity, humor and irony BERT models.

F1 SCORE WITH ROAST ME AUGMENTATION

Training \ Testing	SARC hot topics	SARC mild topics	SARC random	iSarcasm-Eval	Balanced iSarcasm-Eval
SARC hot topics	0.8	0.54	0.59	0.36	0.59
SARC mild topics	0.60	0.71	0.64	0.29	0.49
SARC mild topics augmented with highest RM	0.7	0.74	0.68	0.32	0.62
SARC mild topics augmented with selected RM	0.66	0.72	0.62	0.32	0.55
SARC mild topics augmented with random RM	0.64	0.73	0.62	0.28	0.55

Table 5.4: F1 Scores for BERT model fine tuned on SARC mild topics dataset augmented with different groups of RoastMe comments.

ABLATION TEST CONCEPT

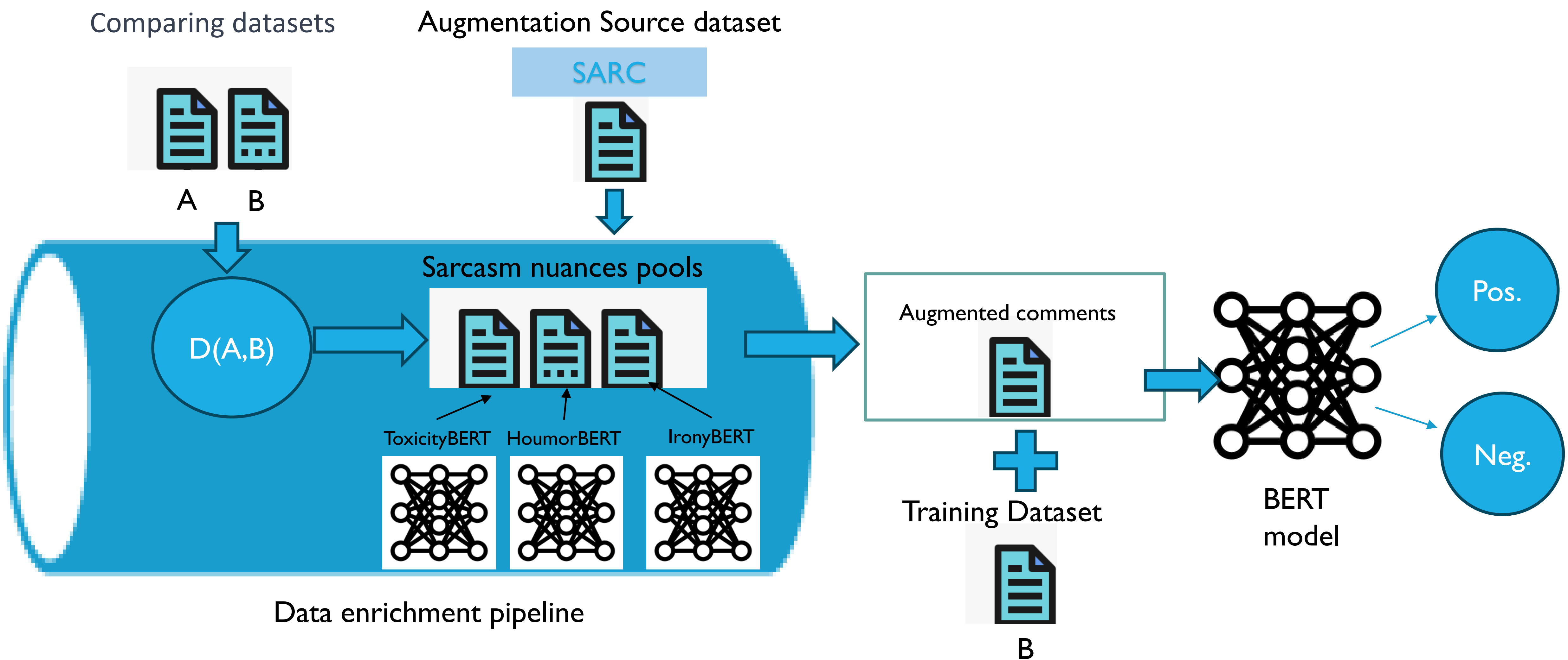
- Ablation test refers to the deliberate removal of a one or more component(s) from the augmentation comments.
- In our case the meaning is augmentation with only one or two sarcasm nuances (instead of three).
- The table below shows that the synergy of all the relevant sarcasm nuances is necessary to achieve higher sarcasm detection results.

Training \ Testing	SARC hot topics	SARC mild topics	SARC random	iSarcasm-Eval	Balanced iSarcasm-Eval test
SARC mild topics augmented with toxicity	0.57	0.65	0.55	0.25	0.4
SARC mild topics augmented with humor	0.62	0.7	0.57	0.31	0.53
SARC mild topics augmented with irony	0.61	0.69	0.56	0.26	0.4
SARC mild topics augmented with toxicity+humor	0.68	0.73	0.62	0.31	0.57
SARC mild topics augmented with toxicity+irony	0.66	0.74	0.64	0.3	0.53
SARC mild topics augmented with irony+humor	0.66	0.73	0.63	0.34	0.58

Table 5.5: F1 scores for the BERT model fine-tuned on the 'SARC mild topics' dataset augmented with different groups of RoastMe comments, as per the ablation test.

ABLATION TEST

AUGMENTATION WITH SARC COMMENTS



F1 SCORE WITH SARC AUGMENTATION

Training \ Testing	SARC hot topics	SARC mild topics	SARC random	iSarcasm-Eval	Balanced iSarcasm-Eval
SARC hot topics	0.8	0.54	0.59	0.36	0.59
SARC mild topics	0.60	0.71	0.64	0.29	0.49
SARC mild topics augmented with highest SARC	0.7	0.74	0.65	0.3	0.59
SARC mild topics augmented with random SARC	0.65	0.71	0.64	0.3	0.55
SARC mild topics augmented with SARC hot topics	0.69	0.75	0.65	0.3	0.57

Table 5.6: F1 Scores for BERT model fine tuned on SARC mild topics dataset augmented with different groups of SARC comments.

ABLATION TEST

Training \ Testing	SARC hot topics	SARC mild topics	SARC random	iSarcasm-Eval	Balanced iSarcasm-Eval
SARC mild topics augmented with toxicity	0.61	0.7	0.59	0.33	0.5
SARC mild topics augmented with humor	0.62	0.69	0.61	0.3	0.52
SARC mild topics augmented with irony	0.63	0.7	0.59	0.31	0.51
SARC mild topics augmented with toxicity+humor	0.65	0.74	0.6	0.29	0.46
SARC mild topics augmented with toxicity+irony	0.66	0.72	0.63	0.33	0.57
SARC mild topics augmented with irony+humor	0.68	0.71	0.59	0.28	0.46

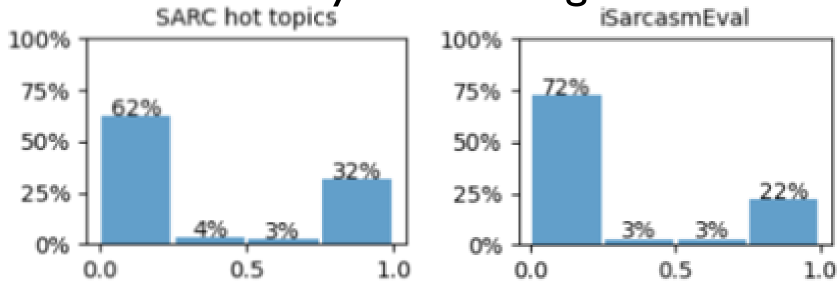
Table 5.7: F1 scores for BERT model fine-tuned on the 'SARC mild topics' dataset augmented with different groups of SARC comments, as per the ablation test.

SMART AUGMENTATION

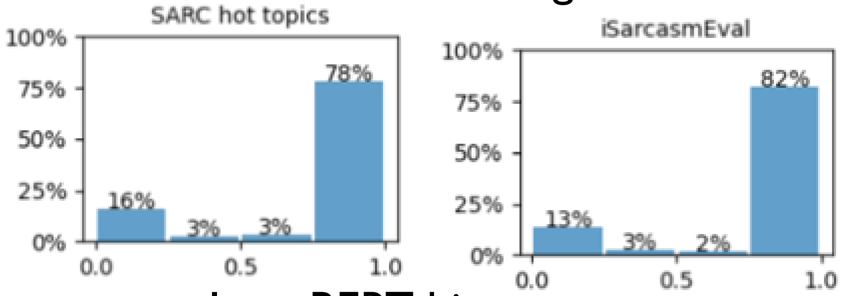
SARC hot topics Versus iSarcasmEval

DISPARITY FUNCTION -EXAMPLE

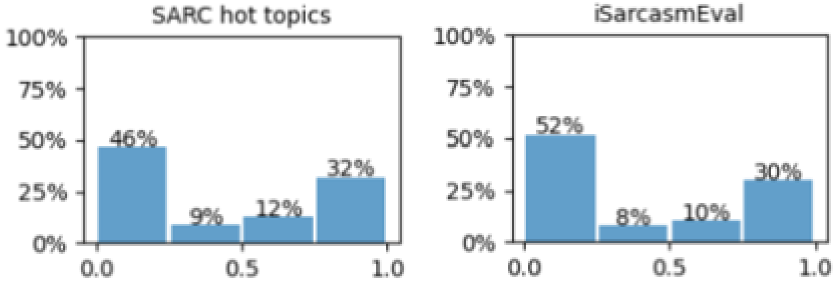
ToxicityBERT histogram



HumorBERT histogram



IronyBERT histogram



$\boldsymbol{D}(SARC\ hot\ topic,\ iSarcamEval)$

500
RoastMe
comments

In this case the disparity is in toxicity and irony nuances. So from the 500 RoastMe comments 166 are toxic ,166 are ironic and 167 are selected randomly.

F1 SCORE WITH ROAST ME AUGMENTATION

Training \ Testing	SARC hot topics	SARC mild topics	SARC random	iSarcasm-Eval	Balanced iSarcasm-Eval
SARC hot topics	0.8	0.54	0.59	0.36	0.59
iSarcasmEval	0.52	0.44	0.48	0.29	0.54
iSarcasmEval augmented with highest RM	0.62	0.56	0.55	0.34	0.6
iSarcasmEval augmented with random RM	0.48	0.43	0.49	0.3	0.49

Table 5.8: F1 Scores for BERT model fine tuned on iSarcamEval dataset augmented with different groups of RoastMe comments.

ABLATION TEST

Training \ Testing	SARC hot topics	SARC mild topics	SARC random	iSarcasm-Eval	Balanced iSarcasm-Eval
SARC hot topics	0.8	0.54	0.59	0.36	0.59
iSarcasmEval	0.52	0.44	0.48	0.29	0.54
iSarcasm-Eval augmented with toxicity	0.59	0.5	0.55	0.37	0.61
iSarcasm-Eval augmented with irony	0.54	0.48	0.54	0.36	0.63

Table 5.9: F1 scores for BERT model fine-tuned on the iSarcamEval dataset dataset augmented with different groups of RoastMe comments, as per the ablation test.

F1 SCORE WITH SARC AUGMENTATION

Training \ Testing	SARC hot topics	SARC mild topics	SARC random	iSarcasm-Eval	Balanced iSarcasm-Eval
SARC hot topics	0.8	0.54	0.59	0.36	0.59
iSarcasmEval	0.52	0.44	0.48	0.29	0.54
iSarcasm-Eval augmented with highest SARC	0.68	0.65	0.63	0.33	0.61
iSarcasm-Eval augmented with random SARC	0.6	0.56	0.52	0.32	0.55
iSarcasmEval augmented with SARC hot topics	0.71	0.62	0.61	0.38	0.67

Table 5.10: F1 Scores for BERT model fine tuned on iSarcasmEval dataset augmented with different groups of SARC comments.

Training \ Testing	SARC hot topics	SARC mild topics	SARC random	iSarcasm-Eval	Balanced iSarcasm-Eval
SARC hot topics	0.8	0.54	0.59	0.36	0.59
iSarcasmEval	0.52	0.44	0.48	0.29	0.54
iSarcasm-Eval augmented with toxicity	0.61	0.61	0.62	0.3	0.56
iSarcasm-Eval augmented with irony	0.61	0.5	0.53	0.33	0.51

Table 5.11: F1 scores for BERT model fine-tuned on the iSarcasmEval dataset augmented with different groups of SARC comments, as per the ablation test.

ABLATION TEST

CONCLUSIONS

- Every sarcastic sentence contain unique scale of sarcastic axes\genes.
- The sarcastic axes\genes can be vary including genes like toxic, humoristic, ironic and other.
- Smart augmentation of Dataset 'A' with specifics comments that contain high concentration of sarcastic genes that lack in 'A' ,can improve the sarcasm detection even in cross domain evaluation .
- The source of Augmentation pools can be vary : Roast Me SARC ,Tweets and more.

RESEARCH EXTENSION

- Expand the "genetic sequencing" of sarcastic text (suggest more axes)
- Compare BERT model sarcasm detection with new LMM on the datasets that take part in this study.
- Use another sarcastic datasets for training and augmentation the BERT model.
- Compre synthetic data vs. real data (RoastMe)