Supervised Contrastive Learning Approach for Contextual Ranking

Abhijit Anand, Jurek Leonhardt, Koustav Rudra, Avishek Anand

01

Introduction

Motivation

Contextual Models have **impressive performance** vs classical models.

But have following **drawbacks**:

- Training data requirement is large
- Fine-tuning with small amount of data does not generalise

How to use contextual model in low data regime?

Problem Statement

To come up with an effective method to improve **document ranking** performance on **smaller datasets**

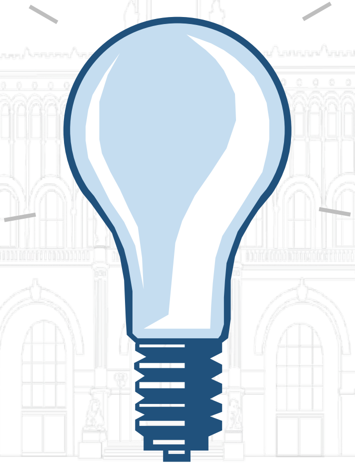
Approach

To improve document ranking performance for smaller dataset we use **data augmentation**(to increase dataset size) coupled with **supervised contrastive learning**

Research Questions

RQ1: Does **Data Augmentation** or **Supervised Contrastive Learning** help to improve document re-ranking performance for smaller datasets?

RQ2: Does the **augmentation style** impact the ranking performance?



Questions

RQ3: How does **training data size** impact ranking performance?

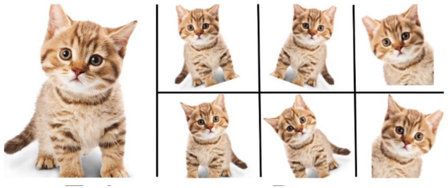
02

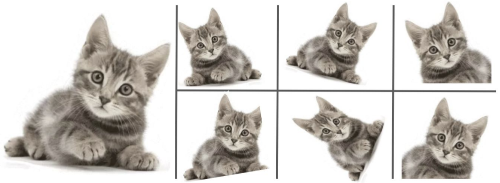
Methodology

Data Augmentation

Why do data augmentation?

- To increase the **training** data without collecting more data





How to do data augmentation?

- Create modified **copies** of existing data or create **synthetic data**

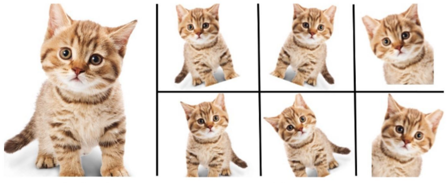
Data Augmentation

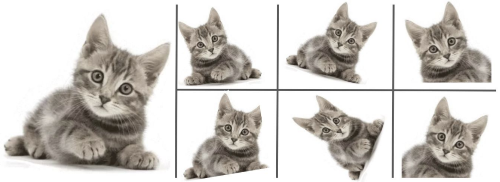
Why do data augmentation?

- To increase the **training** data without collecting more data

How to do data augmentation?

- Create modified **copies** of existing data or create **synthetic data**





Challenges in data augmentation for ranking?
- Relevance of a document is **query specific**
- Positive labels are **sparse** in most ranking datasets

Data Augmentation

Query

Is september a good time to go to aruba?

Positive Document

The Best Time to Travel to Aruba "Catch sunsets from a private beach palapa year-round in Aruba. In Aruba one can almost always count on sunny skies and calm seas. The best time to visit the island depends on the type of vacation. If looking for the cheapest hotel rooms and best travel deals, go when the trade winds stop blowing.

Query

What is a parenthesis phrase?

Positive Document

Algebraic expressions Mathematical phrases Mathematical phrases can be written as verbal sentences You should be able to:- translate verbal sentences into algebraic expressions, - translate algebraic expressions into phrases. Example: The product of two and three. Word „ product ” indicates, that there should be multiplication of these numbers (“product” is a result of multiplication).

Simple Data Augmentation Strategies

Query

Is september a good time to go to aruba?

Positive Document(D^+)

The Best Time to Travel to Aruba "Catch sunsets from a private beach palapa year-round in Aruba. In Aruba one can almost always count on sunny skies and calm seas. The best time to visit the island depends on the type of vacation. If looking for the cheapest hotel rooms and best travel deals, go when the trade winds stop blowing.

The Best Time to Travel to Aruba "Catch sunsets from a private beach palapa year-round in Aruba. **P1**

In Aruba one can almost always count on sunny skies and calm seas. **P2**

The best time to visit the island depends on the type of vacation. **P3**

If looking for the cheapest hotel rooms and best travel deals, go when the trade winds stop blowing. **P4**

Simple Data Augmentation Strategies

Query(q)

Is september a good time to go to aruba?

Score: $S(q,P_i)$

Scoring

The Best Time to Travel to Aruba "Catch sunsets from a private beach palapa year-round in Aruba.	**P1**	**S1=S(q,P1)**
In Aruba one can almost always count on sunny skies and calm seas.	**P2**	**S2=S(q,P2)**
The best time to visit the island depends on the type of vacation.	**P3**	**S3=S(q,P3)**
If looking for the cheapest hotel rooms and best travel deals, go when the trade winds stop blowing.	**P4**	**S4=S(q,P4)**

Simple Data Augmentation Strategies

Scoring

Query(q)

Is september a good time to go to aruba?

Score: $S(q,P_i)$

The Best Time to Travel to Aruba "Catch sunsets from a private beach palapa year-round in Aruba.

P1 **S1=S(q,P1)**

In Aruba one can almost always count on sunny skies and calm seas.

P2 **S2=S(q,P2)**

The best time to visit the island depends on the type of vacation.

P3 **S3=S(q,P3)**

If looking for the cheapest hotel rooms and best travel deals, go when the trade winds stop blowing.

P4 **S4=S(q,P4)**

Matching: BM25, Semantic: Glove, Random sampling

Simple Data Augmentation Strategies

Query(q)

Is september a good time to go to aruba?

Score: $S(q,P_i)$

Scoring

The Best Time to Travel to Aruba "Catch sunsets from a private beach palapa year-round in Aruba.

P1 **S1=S(q,P1)**

In Aruba one can almost always count on sunny skies and calm seas.

P2 **S2=S(q,P2)**

The best time to visit the island depends on the type of vacation.

P3 **S3=S(q,P3)**

If looking for the cheapest hotel rooms and best travel deals, go when the trade winds stop blowing.

P4 **S4=S(q,P4)**

S1>S3>S2>S4

Top 2 Passage for Augmentation

Simple Data Augmentation Strategies

The Best Time to Travel to Aruba
"Catch sunsets from a private
beach palapa year-round in Aruba.
The best time to visit the island
depends on the type of vacation

combined

The Best Time to Travel to Aruba
"Catch sunsets from a private
beach palapa year-round in Aruba.

P1

The best time to visit the island
depends on the type of vacation.

P3

Simple Data Augmentation Strategies

Query(q)

Is september a good time to go to aruba?

Augmented Positive Document(D_A^+)

The Best Time to Travel to Aruba "Catch sunsets from a private beach palapa year-round in Aruba. The best time to visit the island depends on the type of vacation

combined

The Best Time to Travel to Aruba "Catch sunsets from a private beach palapa year-round in Aruba.

P1

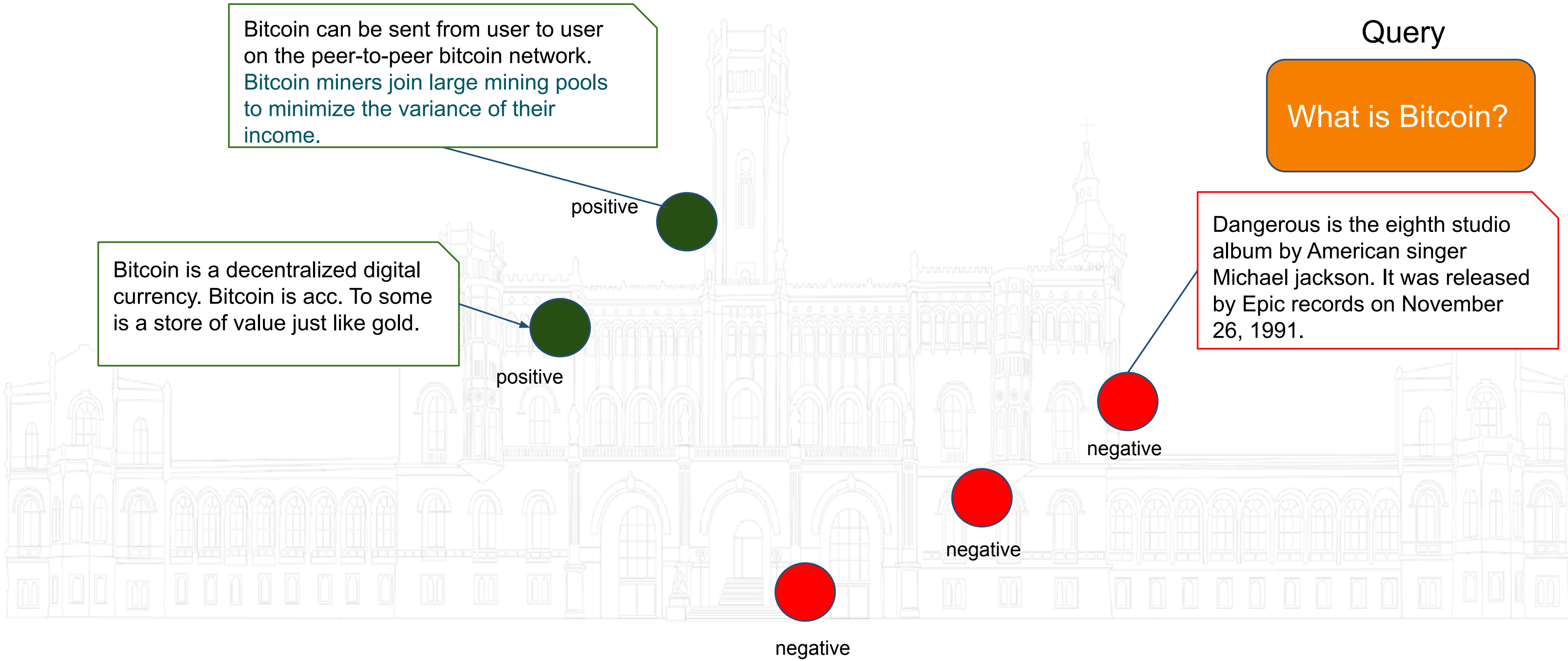
The best time to visit the island depends on the type of vacation.

P3

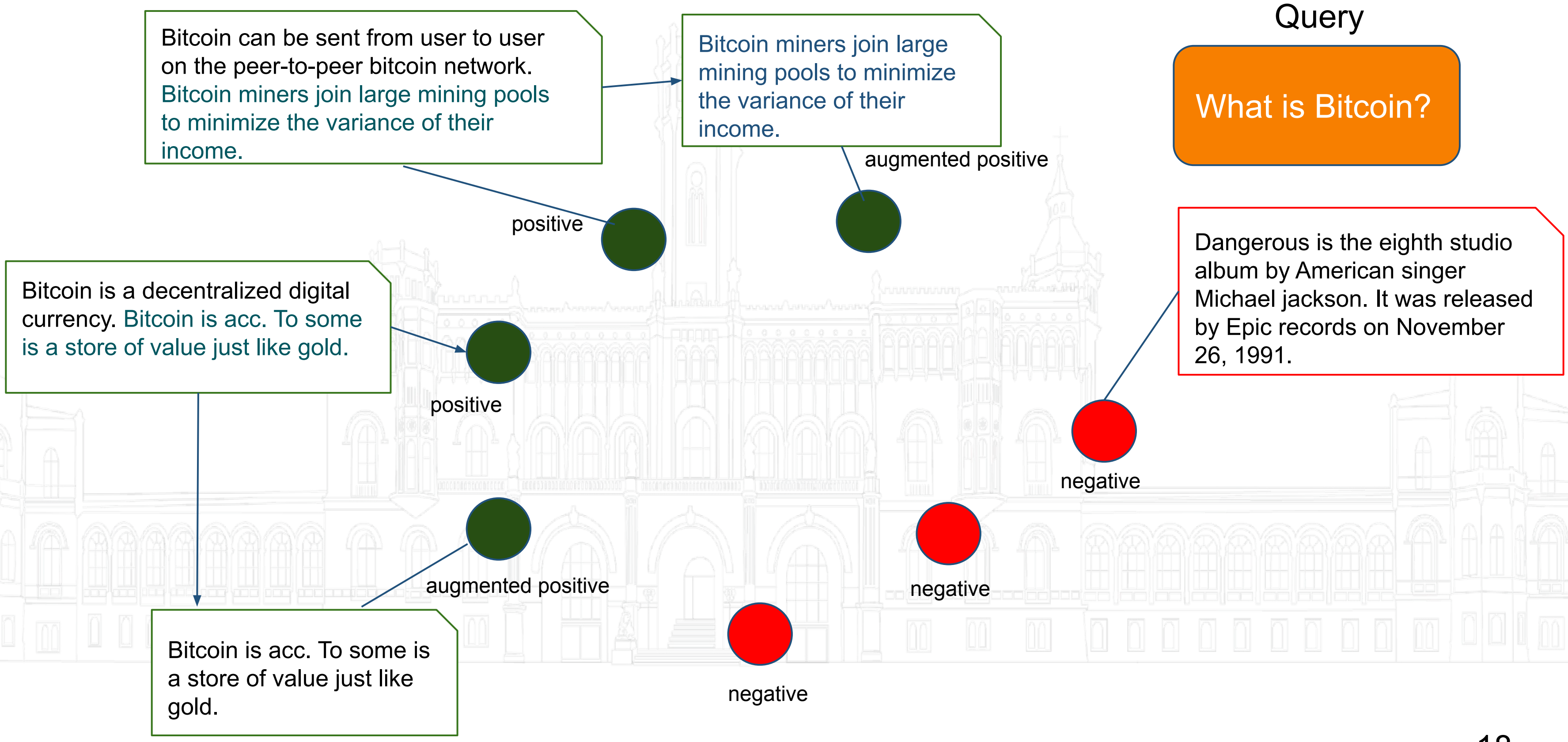
Positive Document(D^+)

The Best Time to Travel to Aruba "Catch sunsets from a private beach palapa year-round in Aruba. In Aruba one can almost always count on sunny skies and calm seas. The best time to visit the island depends on the type of vacation. If looking for the cheapest hotel rooms and best travel deals, go when the trade winds stop blowing.

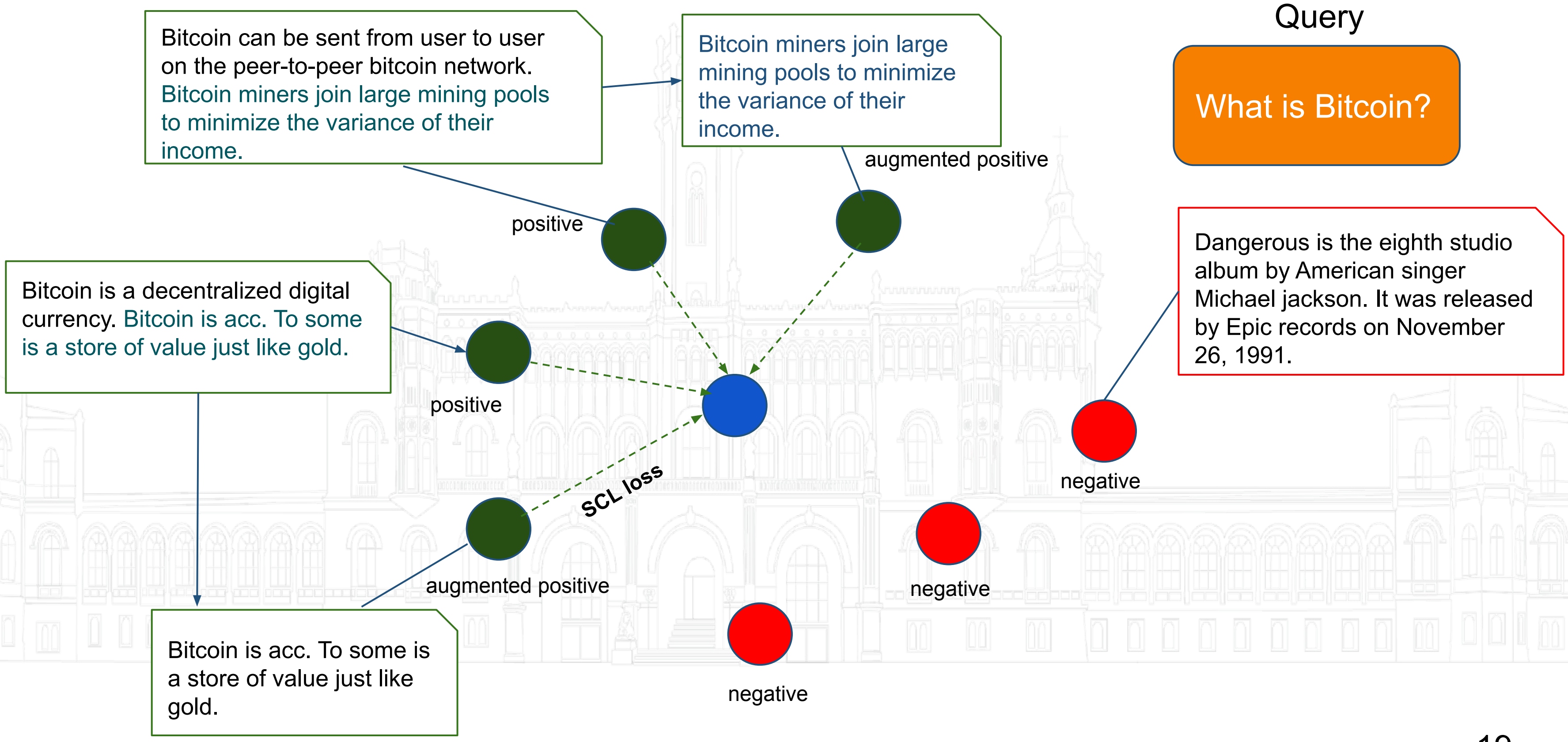
Supervised Contrastive Learning



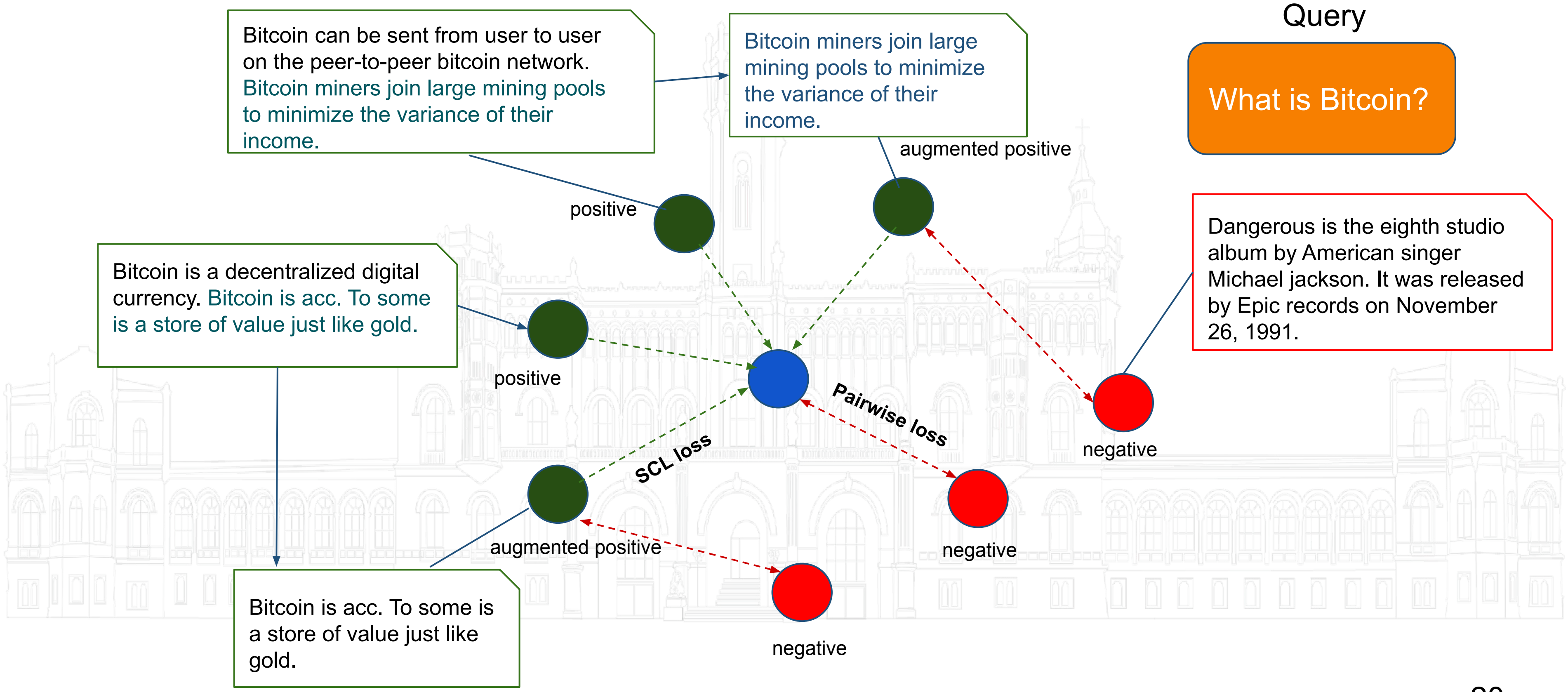
Supervised Contrastive Learning



Supervised Contrastive Learning



Ranking Supervised Contrastive Loss



Supervised Contrastive Loss

Positives

Temperature

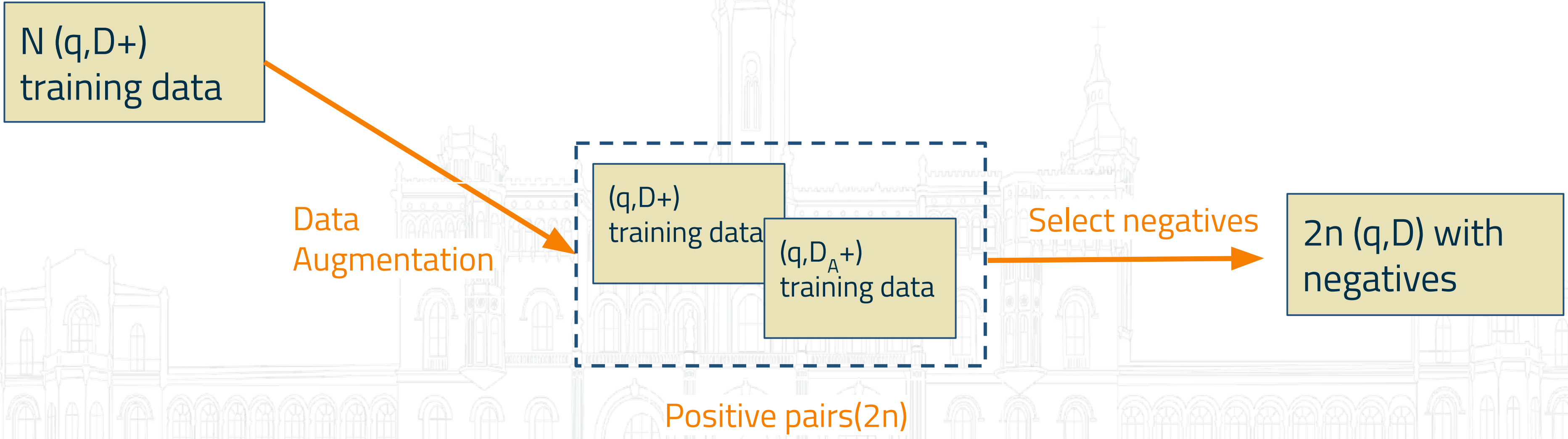
$$\mathcal{L}_{\mathrm{SCL}} = \sum_{i=1}^{N} -\frac{1}{N_+} \sum_{j=1}^{N_+} \mathbf{1}_{\substack{q_i=q_j,\\ i\neq j,\\ y_i=y_j=1}} \log \frac{\exp\left(\Phi(x_i)\cdot\Phi\left(x_j\right)/\tau\right)}{\sum_{k=1}^{N} \mathbf{1}_{i\neq k} \exp(\Phi(x_i)\cdot\Phi(x_k)/\tau)}$$

Ranking Supervised Contrastive Loss

$$\mathcal{L}_{\text{SCL}} = \sum_{i=1}^{N} -\frac{1}{N_+} \sum_{j=1}^{N_+} \mathbf{1}_{\substack{q_i=q_j,\\ i\neq j,\\ y_i=y_j=1}} \log \frac{\exp\left(\Phi(x_i)\cdot\Phi(x_j)/\tau\right)}{\sum_{k=1}^{N} \mathbf{1}_{i\neq k} \exp(\Phi(x_i)\cdot\Phi(x_k)/\tau)}$$

$$\boxed{\mathcal{L}_{\texttt{RankingSCL}} = (1-\lambda)\mathcal{L}_{\texttt{Ranking}} + \lambda\mathcal{L}_{\texttt{SCL}}}$$

Augmented Datasets and Models



03

Experimental Setup

Large Experimental Space

3 Models X 3 Augmentation X 3 Loss functions X 4 datasets

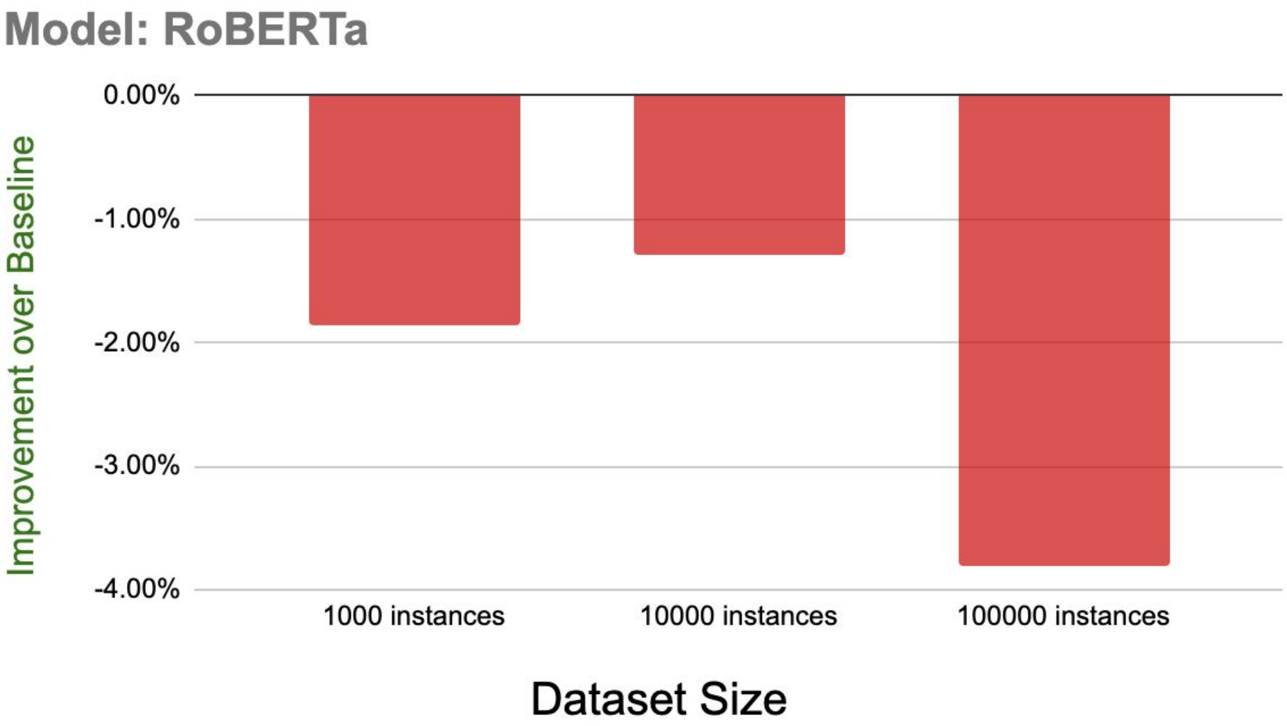
BERT RoBERTa DistillBERT	BM25 GloVe Random	Pointwise Loss Pairwise Loss SCL Ranking Loss	MsMarco Doc ROBUST FiQA SciFact

04

Results

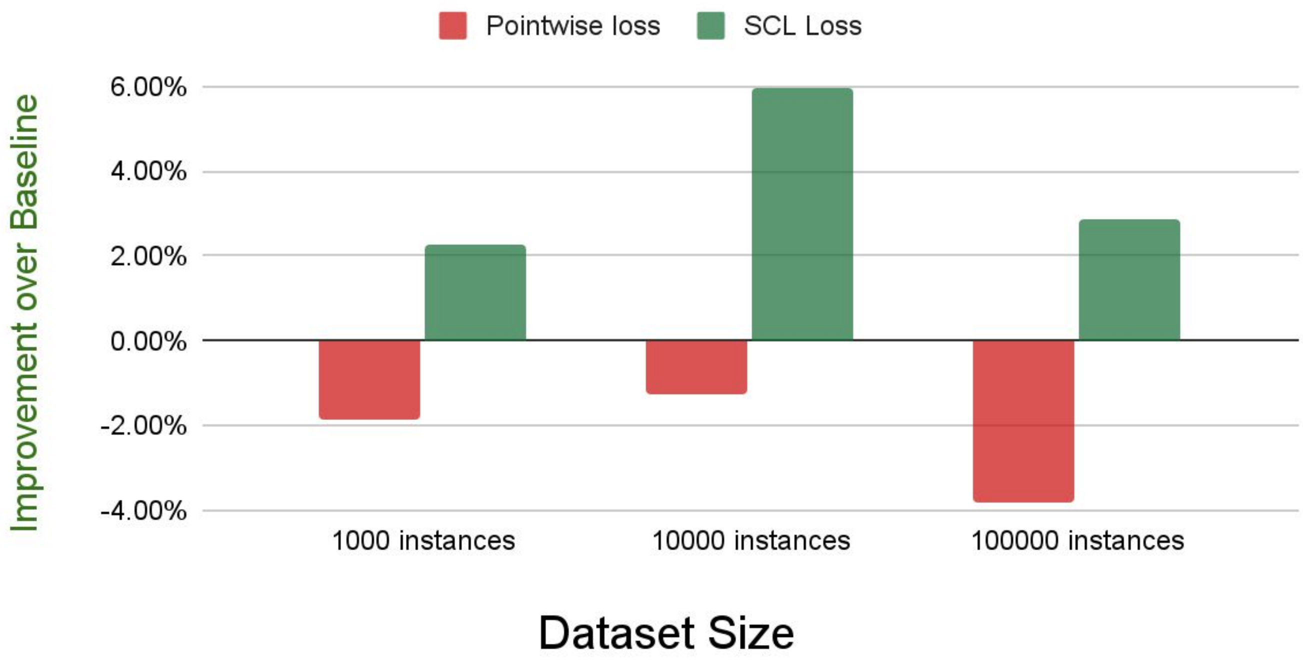
RQ I. Does **data augmentation** or **Supervised Contrastive Learning** help to improve document re-ranking performance for smaller datasets?

RQ I. Does **data augmentation** or **Supervised Contrastive Learning** help to improve document re-ranking performance for smaller datasets?



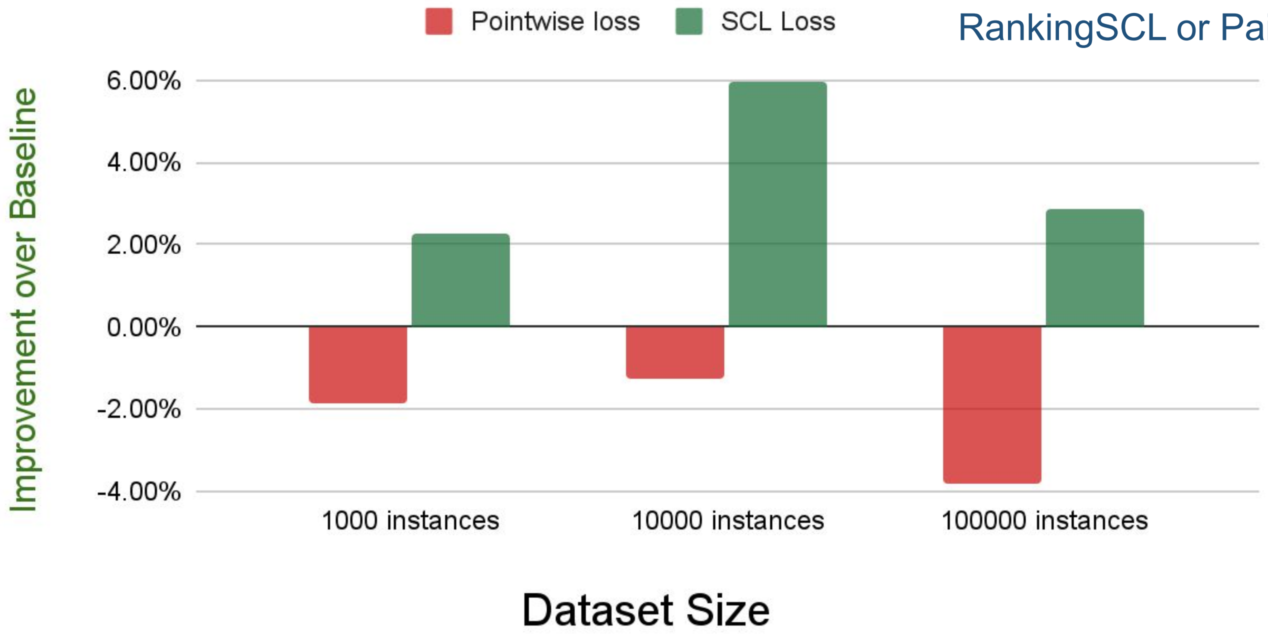
RQ I. Does **data augmentation** or **Supervised Contrastive Learning** help to improve document re-ranking performance for smaller datasets?

Model: RoBERTa



RQ I. Does **data augmentation** or **Supervised Contrastive Learning** help to improve document re-ranking performance for smaller datasets?

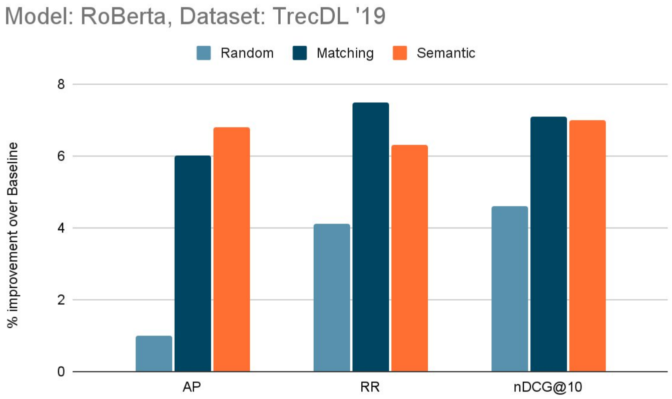
Model: RoBERTa



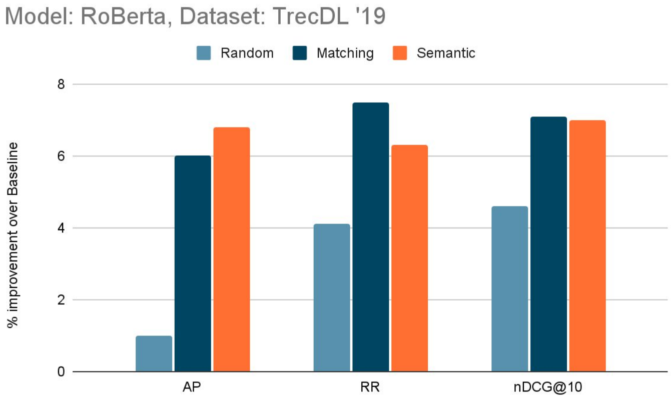
Data augmentation is useful only when **a proper loss function** is used in conjugation, i.e.Pointwise RankingSCL or Pairwise RankingSCL loss

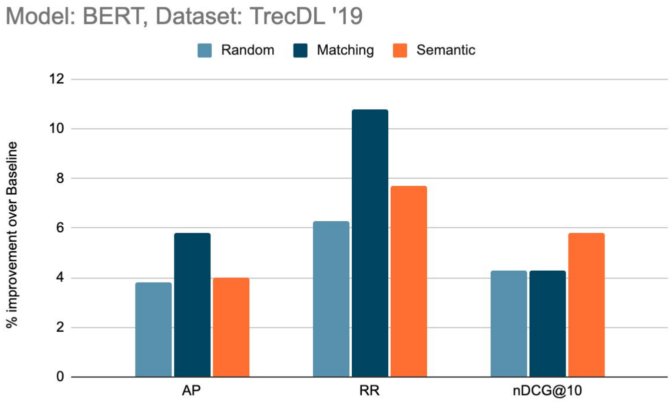
RQ II. Does the **augmentation style** impact the ranking performance?

RQ II. Does the **augmentation style** impact the ranking performance?

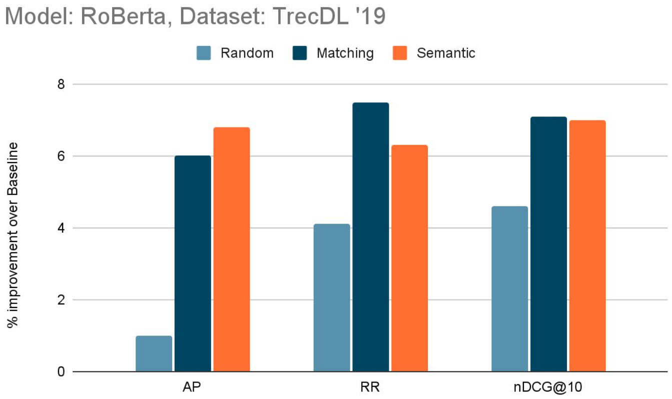


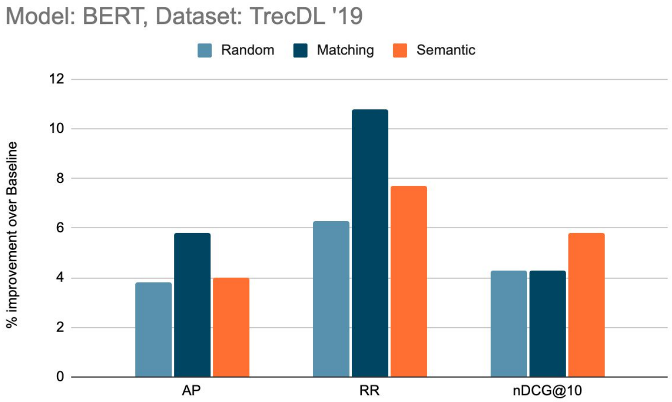
RQ II. Does the **augmentation style** impact the ranking performance?

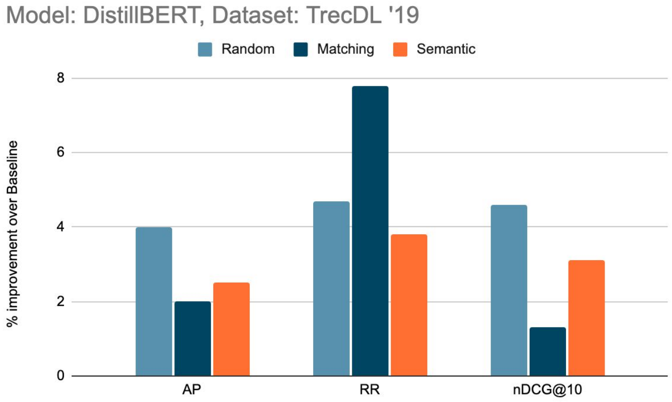




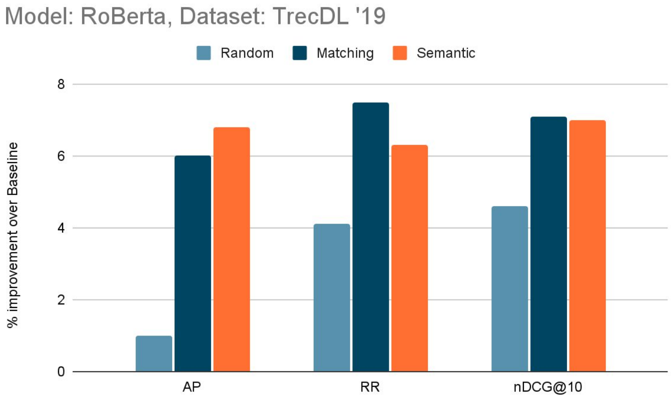
RQ II. Does the **augmentation style** impact the ranking performance?

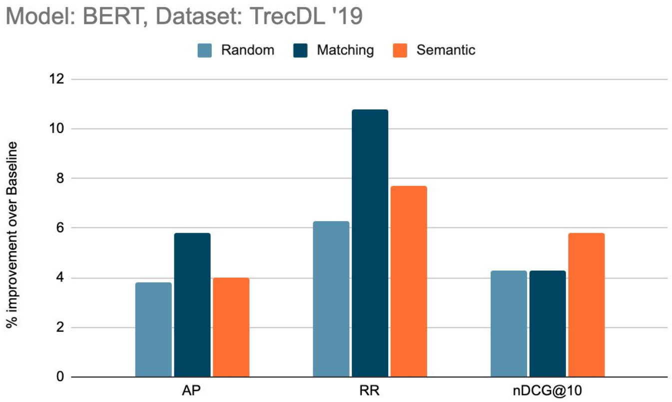


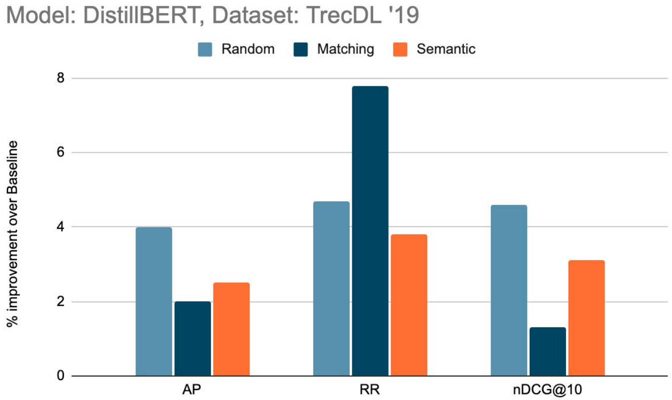




RQ II. Does the **augmentation style** impact the ranking performance?



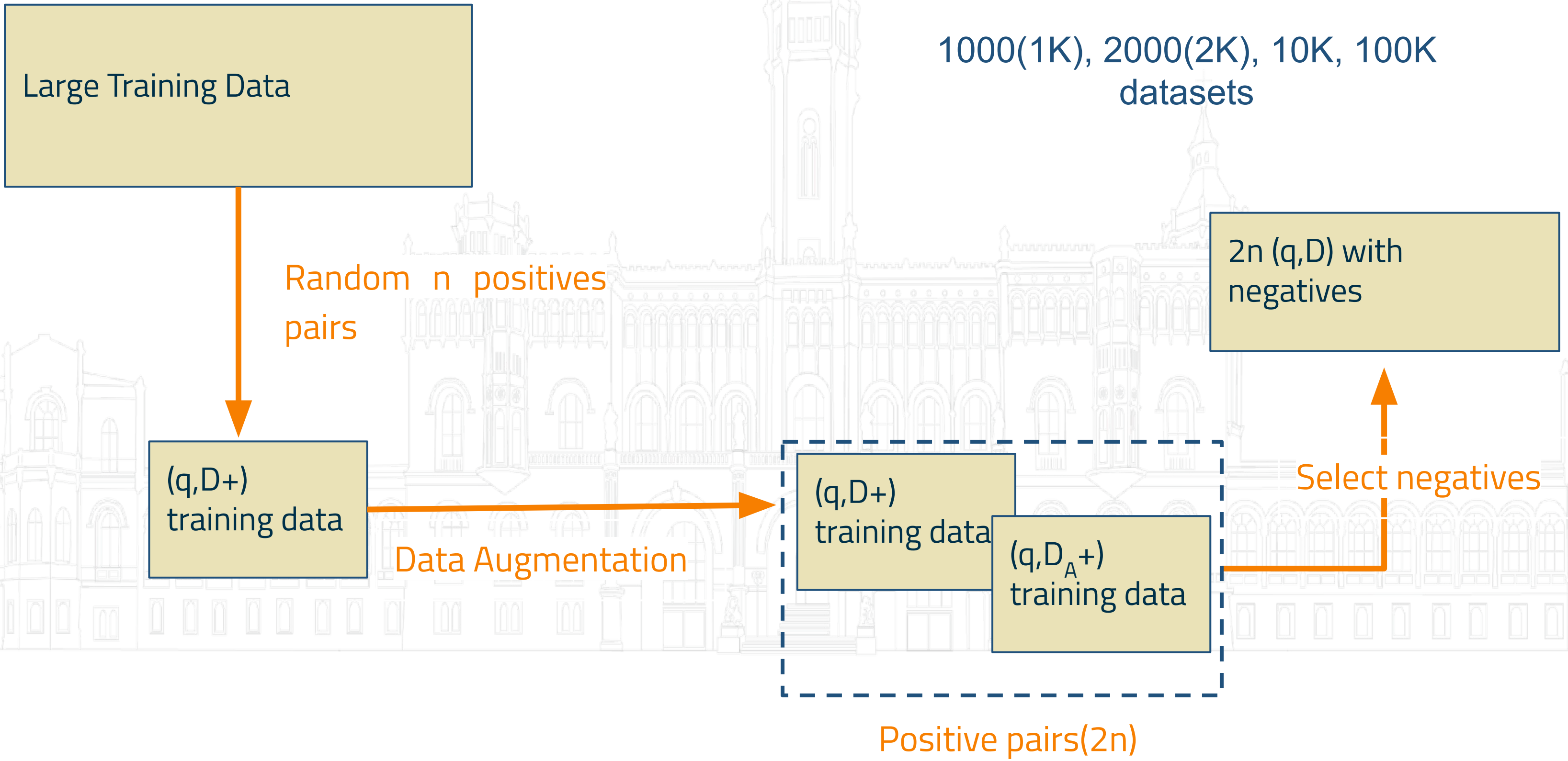




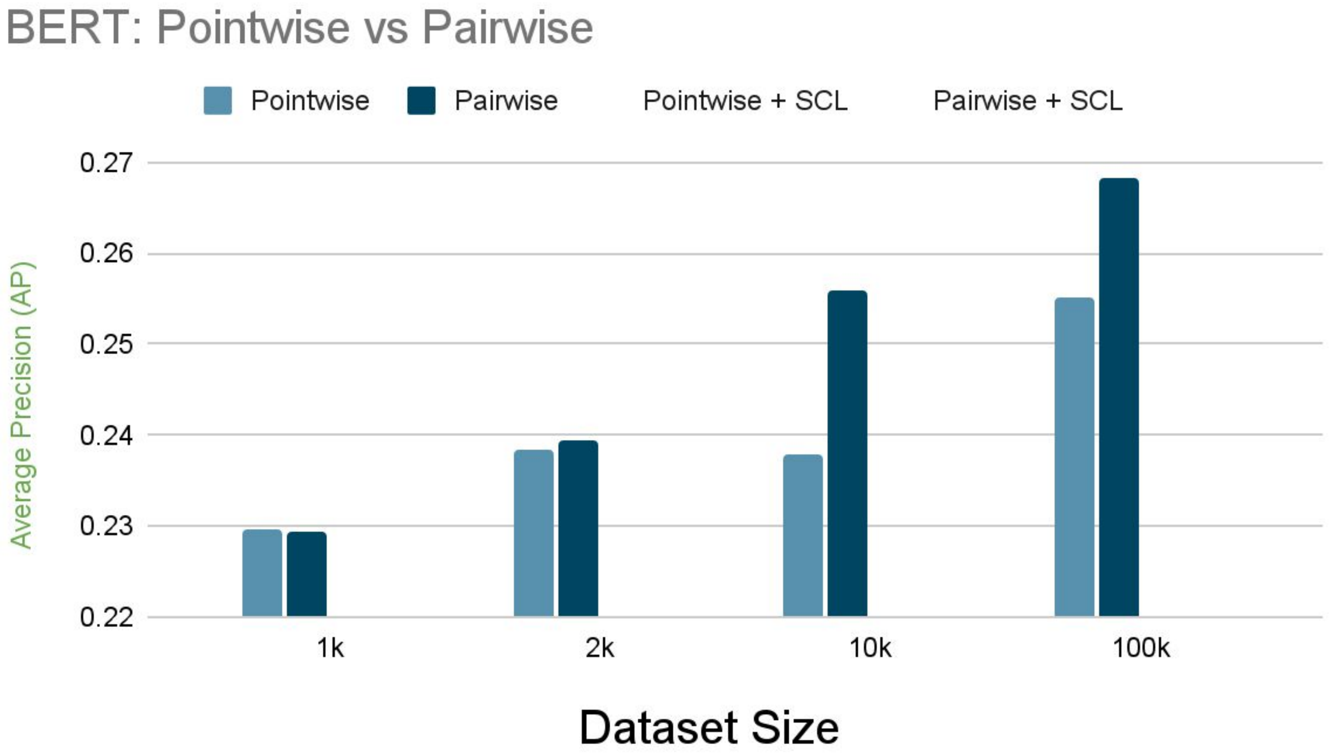
Simple data augmentation strategies **do not have a big impact** on the ranking performance

RQ III. How does **training data size** impact ranking performance?

Augmented Datasets and Models

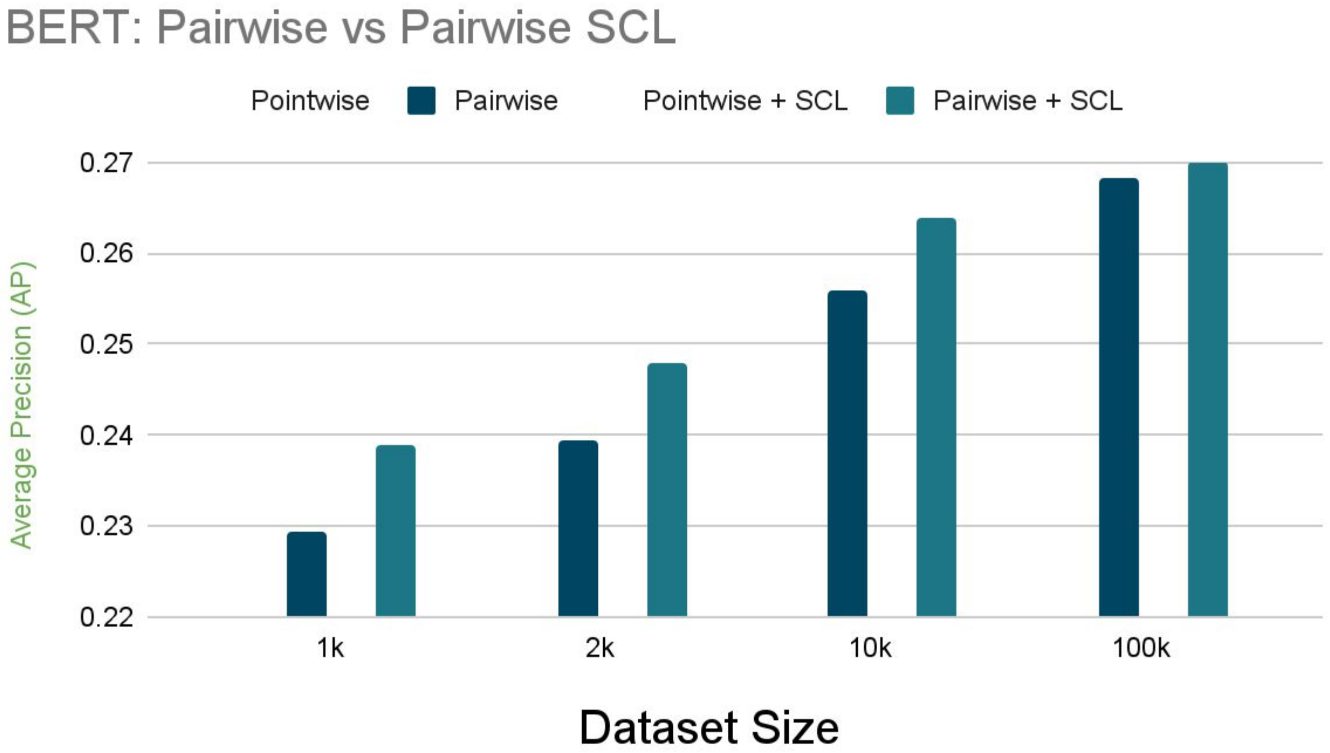


RQ III. How does **training data size** impact ranking performance?



RQ III. How does **training data size** impact ranking performance?

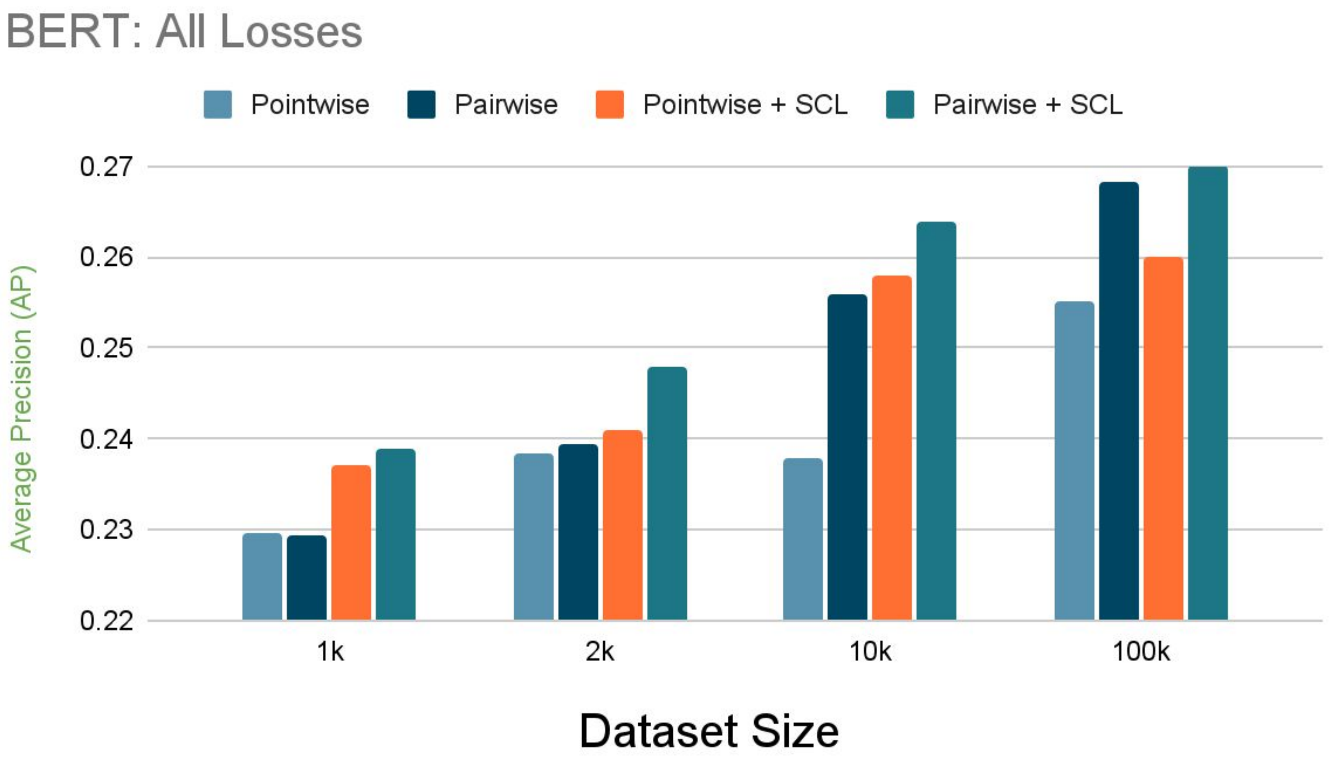
RQ III. How does **training data size** impact ranking performance?



RankingSCL has the **highest marginal utility** when the dataset sizes are **small**

The utility diminishes with increasing dataset size

RQ III. How does **training data size** impact ranking performance?

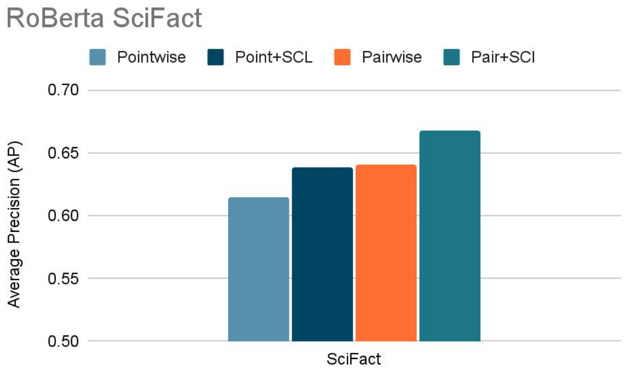


RankingSCL has the **highest marginal utility** when the dataset sizes are **small**

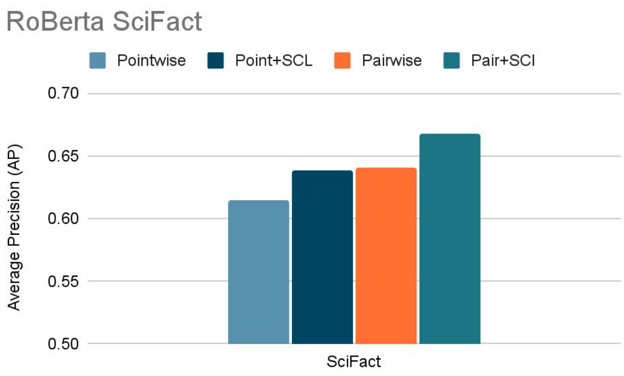
The utility diminishes with increasing dataset size

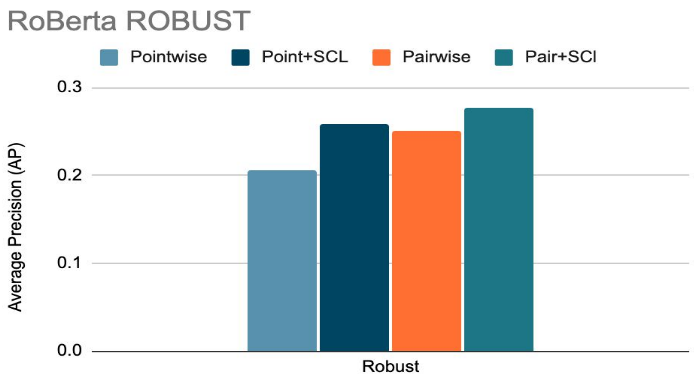
Can we **replicate** the performance on small datasets

Can we **replicate** the performance on small datasets

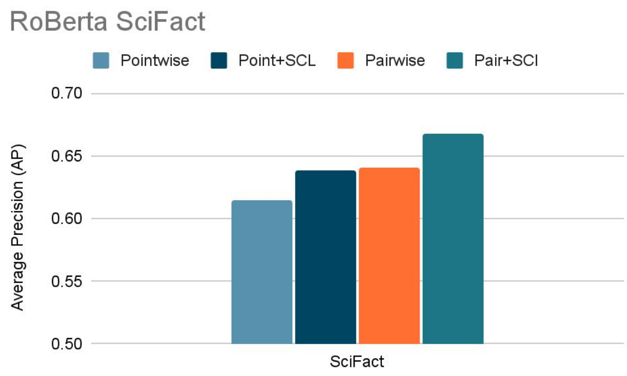


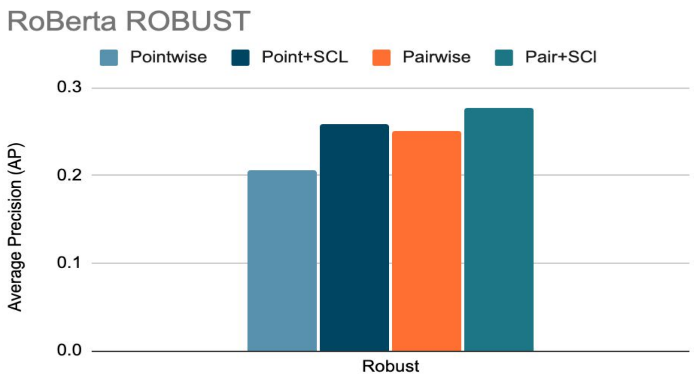
Can we **replicate** the performance on small datasets

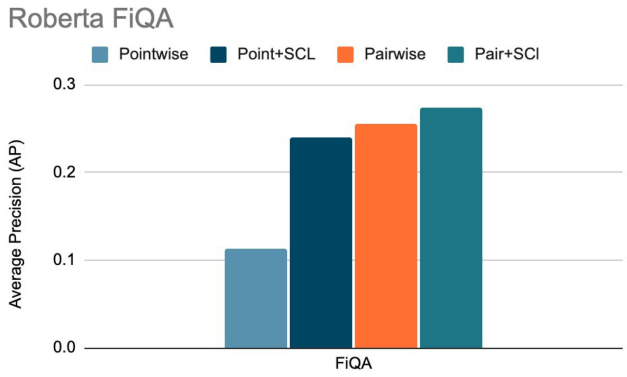




Can we **replicate** the performance on small datasets







RankingSCL results in **large performance gains** on a variety of **small ranking datasets**

05

Conclusion

Conclusion

- Data augmentation is useful only when a **proper loss function** is used in conjugation, i.e.Pointwise RankingSCL or Pairwise RankingSCL loss
- Choice of simple **data augmentation strategies do not have a big impact** on the ranking performance when using RankingSCL (Pointwise or Pairwise).
- RankingSCL has the **highest marginal utility** when the dataset sizes are **small**. The utility diminishes with increasing dataset size.
- RankingSCL results in **large performance gains** on a variety of **small ranking datasets.**

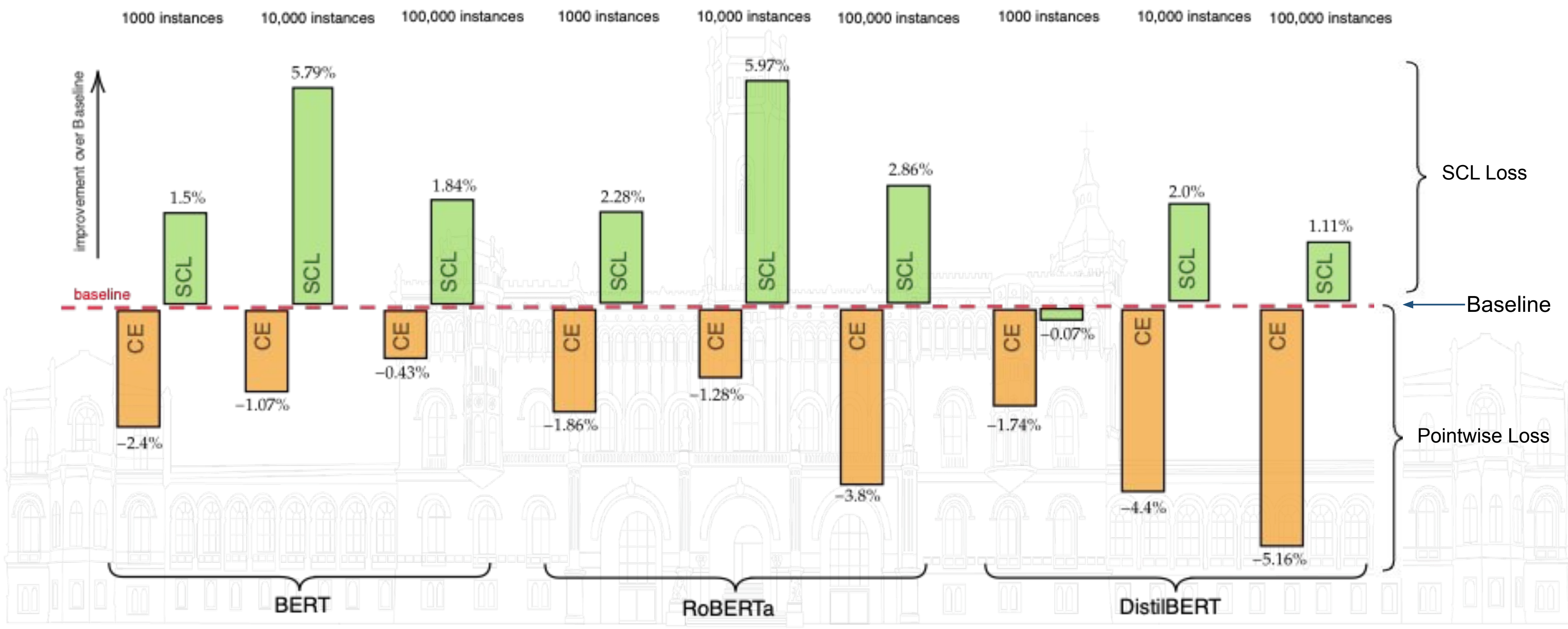
Thank You SIGIR for the Student Travel Grant

Additional Slides

Experiments Conducted

For 1 dataset(Doc’19): 3 Models * 4 datasizes * 3 Augmentation type * 2 Loss function = 72 models combination
Each combination has varying **lambda** and **Temperature**

Total model combination: 72 + 72 + 18+ 18 + 18 (robust, fiqa,scifact)
= 198 + (24+24+6+6) baselines = 258
Total experiments = 198*25+60 = 5010
Results shown for = 78 models


1000 instances
10,000 instances
100,000 instances
1000 instances
10,000 instances
100,000 instances
1000 instances
10,000 instances
100,000 instances
improvement over Baseline
baseline
1.5%
SCL
5.79%
SCL
1.84%
SCL
2.28%
SCL
5.97%
SCL
2.86%
SCL
2.0%
SCL
1.11%
SCL
CE
−2.4%
CE
−1.07%
CE
−0.43%
CE
−1.86%
CE
−1.28%
CE
−3.8%
CE
−1.74%
−0.07%
CE
−4.4%
CE
−5.16%
BERT
RoBERTa
DistilBERT
SCL Loss
Baseline
Pointwise Loss

	Doc'19			Doc'20			Robust04		
	AP	RR	$nDCG_{10}$	AP	RR	$nDCG_{10}$	AP	RR	$nDCG_{10}$
BERT									
Baseline	0.244	0.834	0.592	0.373	0.891	0.547	0.264	0.763	0.506
Sampling	0.253(▲3.8%)	0.886(▲6.3%)	0.617(▲4.3%)	0.391(▲4.8%)	0.941(▲5.6%)	0.594(▲8.6%)*	0.276(▲4.7%)	0.797(▲4.5%)	0.537(▲6%)
BM25	0.258(▲5.8%)	0.924(▲10.8%)	0.617(▲4.3%)	0.378(▲1.3%)	0.944(▲6.0%)	0.562(▲2.8%)	0.273(▲3.1%)	0.793(▲3.9%)	0.533(▲5.3%)
GloVe	0.253(▲4.0%)	0.898(▲7.7%)	0.626(▲5.8%)	0.387(▲3.8%)	0.940(▲5.6%)	0.566(▲3.5%)	0.278(▲5.2%)	0.799(▲4.7%)	0.541(▲6.8%)
RoBERTa									
Baseline	0.243	0.812	0.557	0.307	0.725	0.470	0.205	0.594	0.378
Sampling	0.245(▲1.0%)	0.878(▲4.1%)	0.583(▲4.6%)	0.365(▲18.8%)	0.922(▲27.2%)	0.557(▲18.5%)*	0.257(▲25.8%)	0.746(▲25.5%)	0.496(▲37.4%)
BM25	0.257(▲6.0%)	0.873(▲7.5%)	0.597(▲7.1%)*#	0.362(▲18.1%)	0.922(▲27.2%)	0.548(▲16.7%)*	0.265(▲29.7%)	0.766(▲28.8%)	0.509(▲34.9%)
GloVe	0.259(▲6.8%)	0.863(▲6.3%)	0.596(▲7.0%)*#	0.354(▲15.3%)	0.870(▲20.0%)	0.550(▲17%)*	0.267(▲30.4%)	0.787(▲32.4%)	0.519(▲37.3%)
DistilBERT									
Baseline	0.244	0.843	0.565	0.322	0.849	0.515	0.201	0.614	0.395
Sampling	0.253 (▲4%)	0.883(▲4.7%)	0.591(▲4.6%)#	0.350(▲8.8%)	0.919(▲8.2%)	0.557(▲8.1%)	0.213 (▲6.3%)	0.713(▲16.2%)	0.480(▲21.6%)
BM25	0.248(▲2.0%)	0.909(▲7.8%)	0.573(▲1.3%)#	0.346(▲7.6%)	0.915(▲7.7%)	0.538(▲4.4%)	0.211(▲5.3%)	0.704(▲14.7%)	0.505(▲27.8%)
GloVe	0.250(▲2.5%)	0.872(▲3.8%)	0.583(▲3.1%)	0.338(▲5.1%)	0.907(▲6.8%)	0.505(▼−1.9%)	0.210(▲3.9%)	0.681(▲11.0%)	0.509(▲2[illegible].9%)

	Pointwise			Pairwise		
Ranking Models	AP	RR	$nDCG_{10}$	AP	RR	$nDCG_{10}$
BERT						
1k	0.237(▲1.5%)	0.868(▲3.7%)	0.551(▲3.1%)	0.239(▲4.3%)	0.851(▲6.2%)	0.576(▲5.7%)
2k	0.241(▲1.9%)	0.916(▲12.9%)	0.592(▲5.2%)	0.248(▲3.6%)	0.892(▼ − 0.4%)	0.603(▲1.5%) *
10k	0.258(▲5.8%)	0.924(▲10.8%)	0.617(▲4.3%)	0.264(▲3.1%)	0.926(▲3.9%)	0.627(▲7.5%)*
100k	0.260(▲1.8%)	0.942(▲4.3%)	0.653(▲6.3%)	0.270(▲0.6%)	0.959(▲2.7%)	0.666(▲3.4%)
RoBERTa						
1k	0.170(▲2.3%)	0.697(▲25.9%)	0.319(▲7.4%)	0.228(▲25.9%)	0.803(▲15.7%)	0.533(▲59.8%)
2k	0.171(▲1%)	0.670(▲12.4%)	0.322(▲9.5%)	0.236(▲4.4%)	0.871(▲4.7%)	0.587(▲7.4%)
10k	0.257(▲6%)	0.873(▲7.5%)	0.597(▲7.1%)*#	0.261(▲3.5%)	0.914(▲3.8%)	0.633(▲3.5%)*
100k	0.263(▲2.9%)	0.946(▲4.7%)	0.646(▲11.7%)	0.270(▲1.2%)	0.955(▲1.4%)	0.6667(▲0.3%)
DistilBERT						
1k	0.150(▲0%)	0.553(▲14.3%)	0.239(▲9.2%)	0.208(▲33.9%)	0.802(▲35.8%)	0.471(▲61.4%)
2k	0.164(▲2.3%)	0.589(▲0.6%)	0.304(▲9.2%)	0.231(▲15%)	0.862(▲13.1%)	0.526(▲19.4%)
10k	0.248(▲2.0%)	0.909(▲7.8%)	0.573(▲1.3%) #	0.253(▲5.1%)	0.893(▲3.9%)	0.613(▲7.7%)*
100k	0.255(▲1.1%)	0.942(▲3.1%)	0.641(▲5.7%)	0.270(▲3.3%)	0.927(▲2.9%)	0.645(▲1.5%)*

Can we replicate the performance on small datasets

	Robust04			SciFact			FiQA		
	AP	RR	$nDCG_{10}$	AP	RR	$nDCG_{10}$	AP	RR	$nDCG_{10}$
BERT									
Base-pointwise	0.264	0.763	0.506	0.312	0.32	0.383	0.140	0.221	0.187
Pointwise	0.276(▲4.7%)	0.797(▲4.5%)	0.537(▲6%)	0.434(▲39%)	0.448(▲40%)	0.466(▲22%)	0.141(▲0.8%)	0.221(▲3.4%)	0.187(▼−1.5%)
Base-pairwise	0.195	0.599	0.382	0.454	0.466	0.504	0.136	0.205	0.174
Pairwise	0.200(▲2.7%)	0.601(▲0.4%)	0.388(▲1.6%)	0.562(▲33.6%)	0.575(▲23.5%)	0.616(▲29%)*	0.221(▲63%)	0.343(▲67%)	0.277(▲59%)
RoBERTa									
Base-pointwise	0.205	0.594	0.3776	0.615	0.626	0.668	0.113	0.173	0.146
Pointwise	0.258(▲26%)	0.746(▲25.5%)	0.496(▲37.4%)	0.638(▲3.7%)	0.649(▲3.7%)	0.687(▲2.8%)*	0.240(▲112%)	0.365(▲111%)	0.300(▲108%)
Base-pairwise	0.250	0.762	0.460	0.641	0.652	0.685	0.255	0.382	0.316
Pairwise	0.277(▲13.9%)	0.529(▲11.65%)	0.766(▲6.1%)*	0.668(▲4.2%)	0.681(▲4.5%)	0.712(▲3.8%)*	0.274(▲7.6%)	0.412(▲7.9%)	0.339(▲7.4%)*
DistilBERT									
Base-pointwise	0.201	0.614	0.395	0.551	0.567	0.595	0.111	0.188	0.132
Pointwise	0.258(▲28.5%)	0.688(▲12.1%)	0.480(▲21.6%)	0.532(▼−3.5%)	0.558(▼−3.3%)	0.574(▼−3.6%)	0.170(▲54%)	0.269(▲43%)	0.216(▲64%)*
Base-pairwise	0.186	0.372	0.576	0.538	0.554	0.577	0.235	0.362	0.288
Pairwise	0.182(▼−1.9%)	0.617(▲7%)	0.375(▲0.7%)*	0.558(▲3.8%)	0.573(▲3.4%)	0.599(▲3.8%)	0.238(▲1.2%)	0.366(▲1.2%)	0.319(▲10.8%)*

Supervised Contrastive loss

$$\mathcal{L}_{\mathrm{SCL}} = \sum_{i=1}^{N} -\frac{1}{N_+} \sum_{j=1}^{N_+} \mathbf{1}_{\substack{q_i=q_j,\\ i\neq j,\\ y_i=y_j=1}} \log \frac{\exp\left(\Phi(x_i)\cdot\Phi\left(x_j\right)/\tau\right)}{\sum_{k=1}^{N} \mathbf{1}_{i\neq k} \exp(\Phi(x_i)\cdot\Phi(x_k)/\tau)}$$

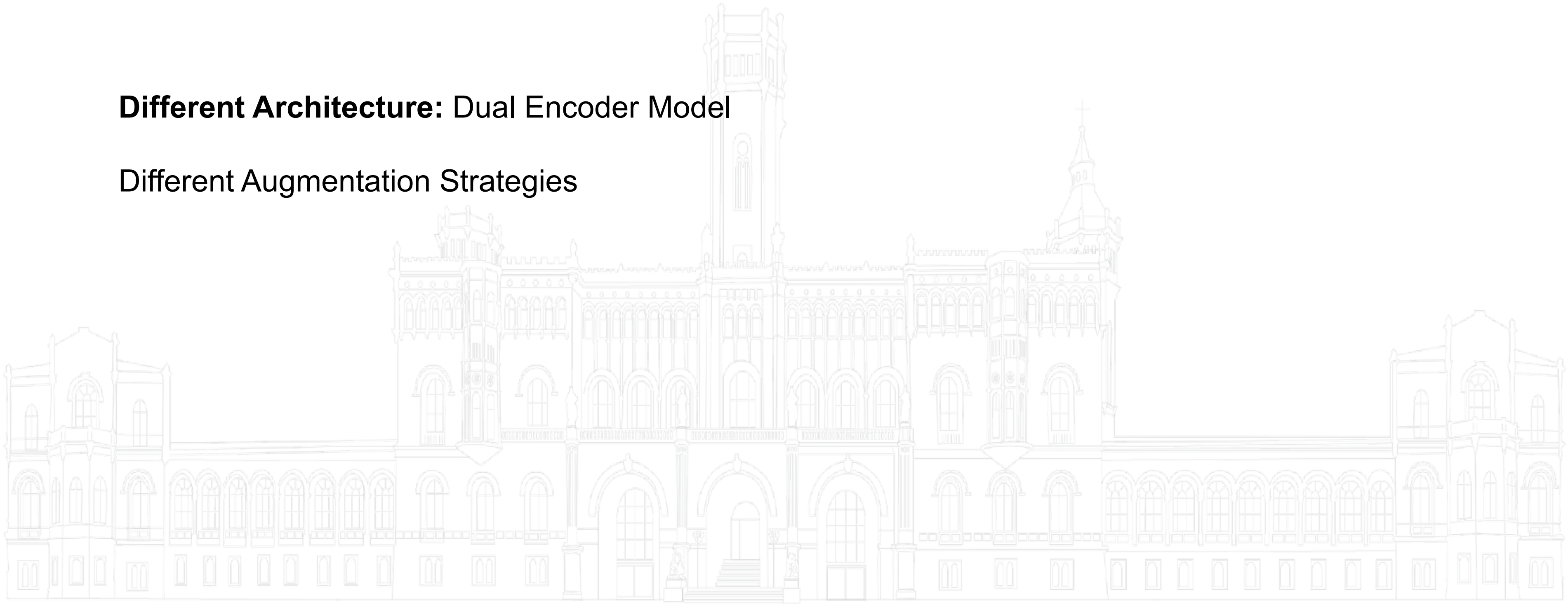
$$\boxed{\mathcal{L}_{\mathtt{Point}} = -\frac{1}{N} \sum_{i=1}^{N} (y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i))}$$

$$\boxed{\mathcal{L}_{\mathtt{Pair}} = \frac{1}{N} \sum_{i=1}^{N} \max\left\{0, m - \hat{y}_i^{+} + \hat{y}_i^{-}\right\}}$$

Future Work

Different Architecture: Dual Encoder Model

Different Augmentation Strategies



Datasets

MsMarco Document Collection
Queries: **367K**
Corpus: **3.2 million**

Document ranking dataset with long documents.

Used as a Dev and training set. TrecDL’19 & TrecDL’20 used as test sets.

ROBUST
Queries: 250
Corpus: 528K

News related dataset with long documents.

We focus on the re-ranking scenario

Datasets

MsMarco Document Collection
Queries: **367K**
Corpus: **3.2 million**

FiQA
Queries: **6650**
Corpus: **57K**

→ Question Answering dataset over Financial text.

ROBUST
Queries: **250**
Corpus: **528K**

SciFact
Queries: **1110**
Corpus: **5K**

→ Fact checking dataset.