

Twitter IR with BM25 and re-ranking with BERT

TU Graz, Advanced Information Retrieval (706.705)
Group 09

Ángela González de Diego - 12307642

Claudia Meana Iturri

Cristina Vera Zambrano - 12312000

<https://github.com/cverazam/AIR-Project.git>

Motivation

“Analysis of the accuracy and relevance of tweet search results through the application of a BM25-based information retrieval model followed by re-ranking with DistilBERT”

Twitter's
influence



+500 mill
tweets/day



Retrieval
challenges



Twitter dataset

Signal-1M Related Tweets

BelR/signal1m-generated-queries

- English language
- 8.4 million relationships between queries and tweets
- Dataset cleaning: Preliminary reduction of the dataset.
Elimination of queries with less than 5 docs associated

Methods:



Data
Analysis

1



Ranking and
Re-Ranking

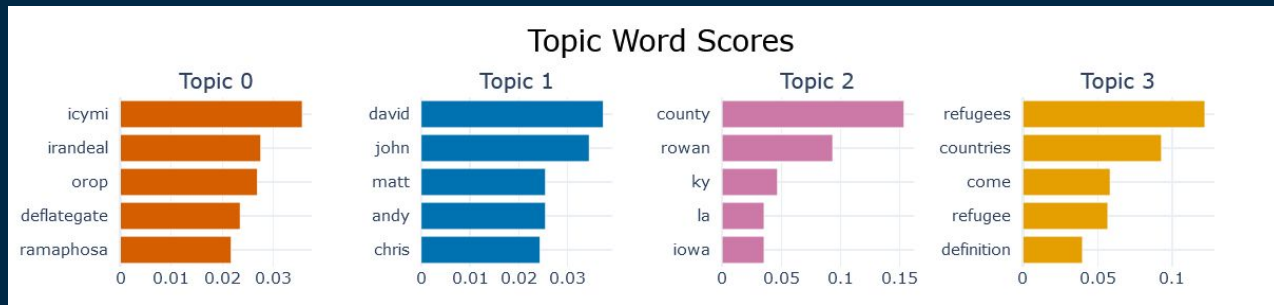
2

1

DATA ANALYSIS: topic modelling (BERTopic)

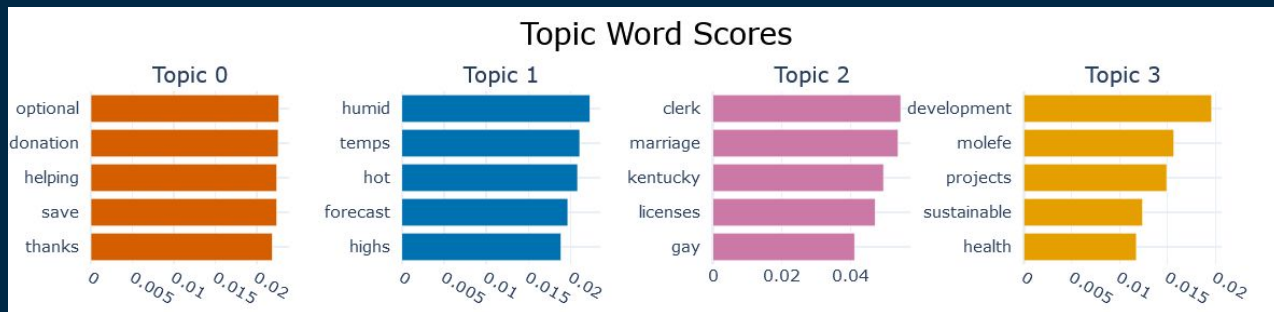
QUERIES

- 7989 queries
- 240 topics



TWEETS

- 40000 docs
- 436 topics

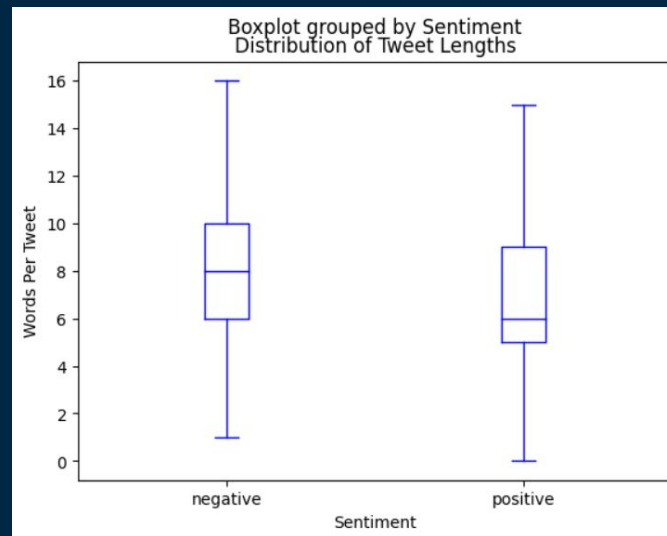
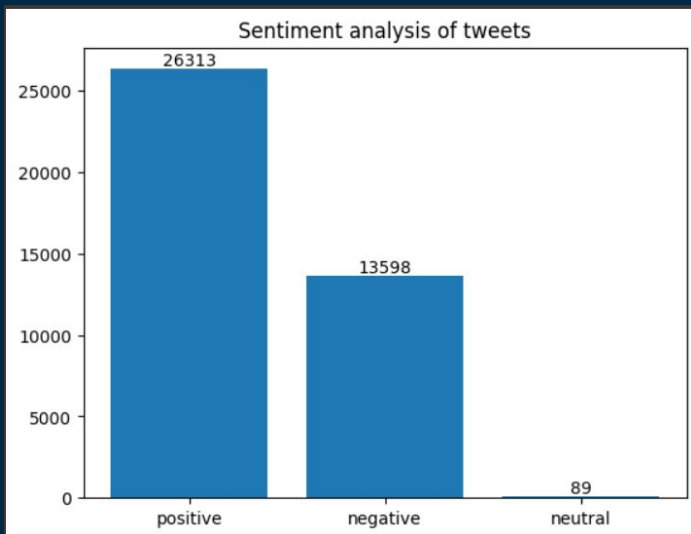


1

DATA ANALYSIS: sentiment analysis (DistilBERT)

TWEETS
40000 docs

- 26313 positive
- 13598 negative
- 89 neutral

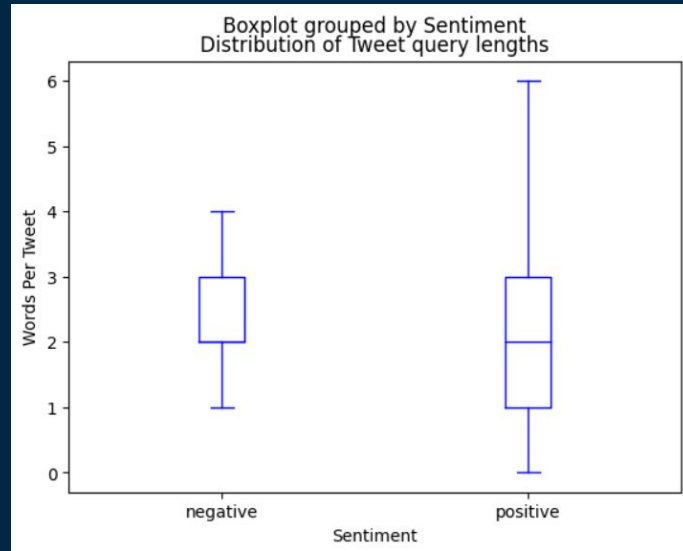
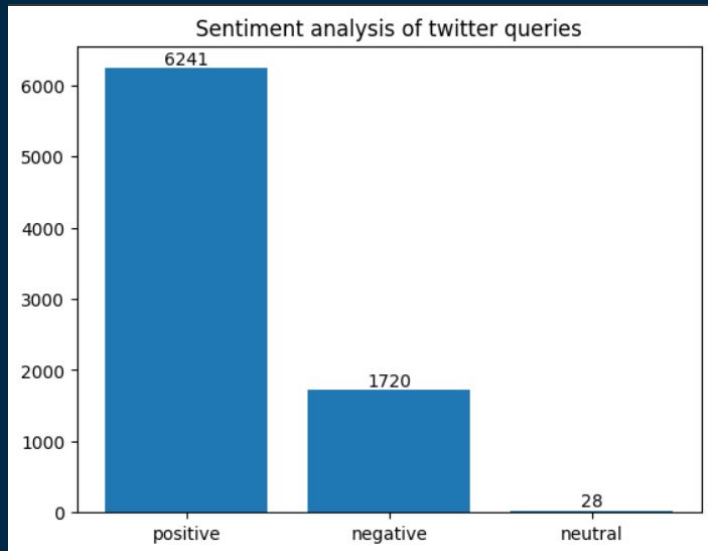


1

DATA ANALYSIS: sentiment analysis (DistilBERT)

QUERIES
7989 queries

- 6241 positive
- 1720 negative
- 28 neutral



DATA ANALYSIS: sentiment analysis (DistilBERT)

- 6241 positive
- 1720 negative
- 28 neutra

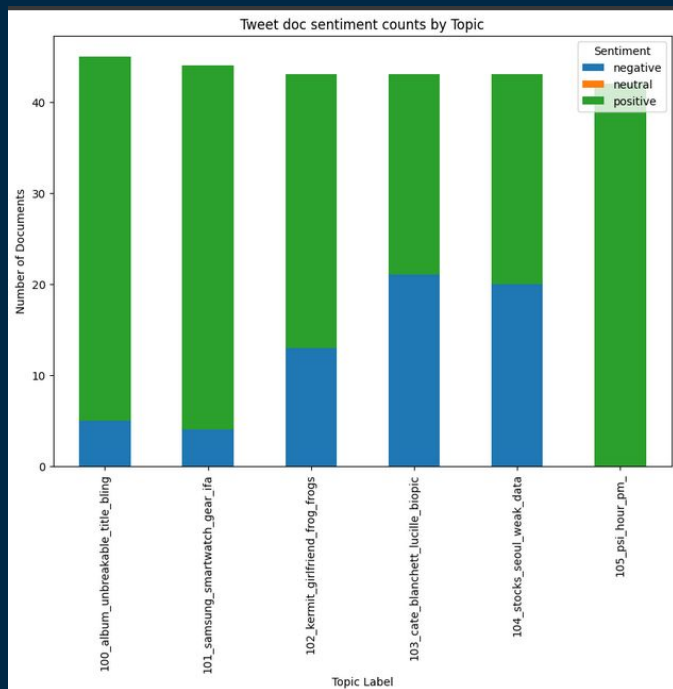
Most frequent negative words



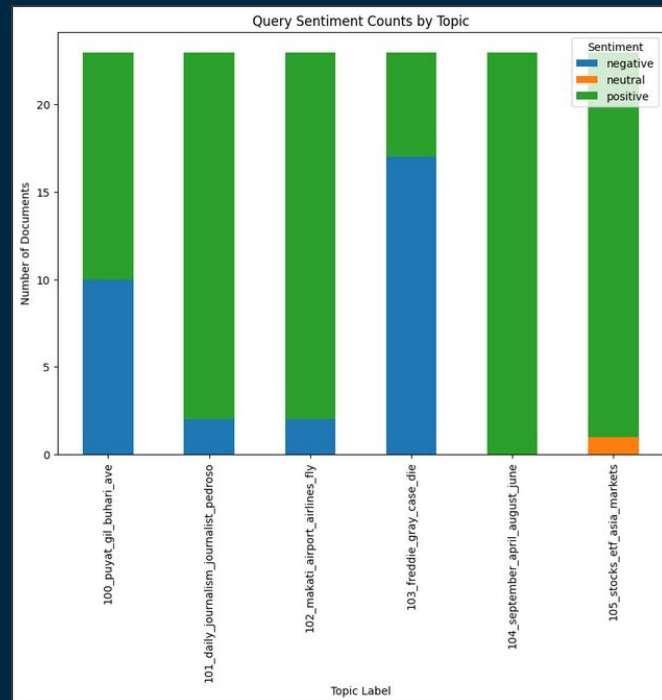
1

DATA ANALYSIS: correlation topic-sentiment

TWEETS



QUERIES



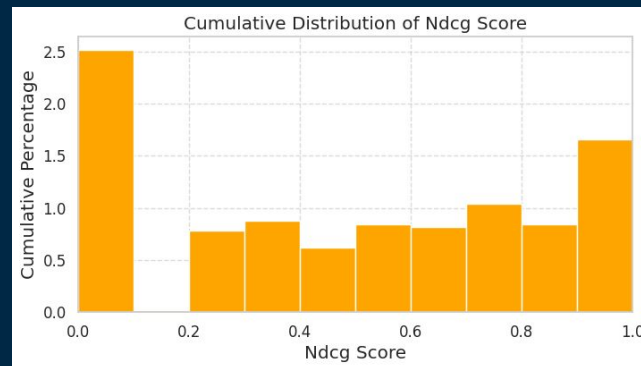
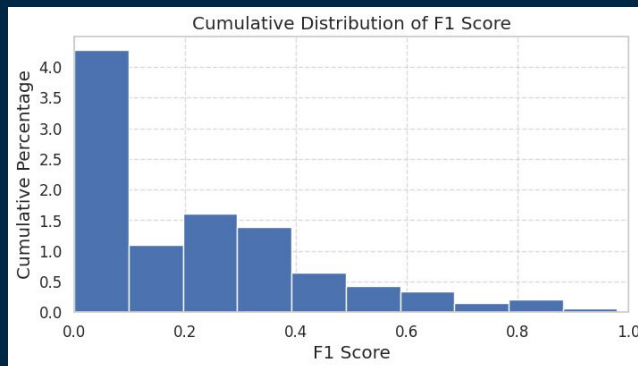
2

RANKING WITH BM25 AND RE-RANKING WITH BERT

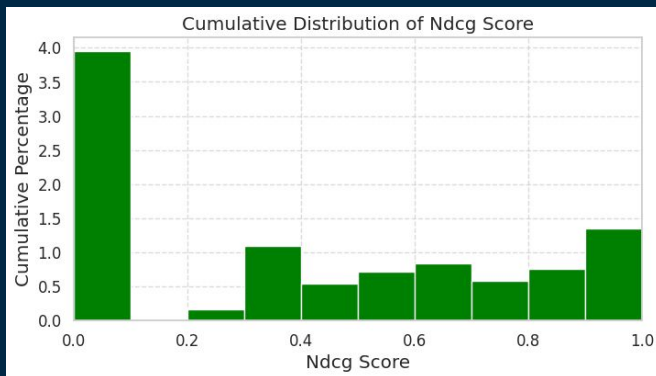
Sentence-transformers model

It maps sentences & paragraphs to a 768 dimensional dense vector space. It can be used for tasks like clustering or semantic search

BM25



BM25 AND DISTILBERT



2

THE DATASET

```

▶ filtered_queries[filtered_queries['text'] == "A hot start to September tomorrow! I'll show you how long the heat streak will last from tonight's CTV News at Six."]

```



_id title

text

query

_id	title	text	query
1210	638502758006566912	A hot start to September tomorrow! I'll show y...	what's the weather like in september

```

[ ] filtered_queries[filtered_queries['query'] == "when is september"]

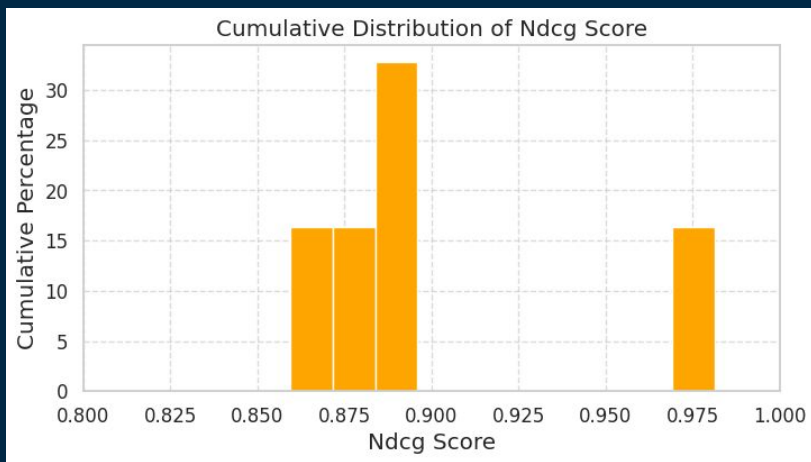
```

	_id	title	text	query
51	638501434560876544	September Starts With A Sizzle: We're putting ...	when is september	
13967	638519212491804672	September is National Preparedness Month. 2015...	when is september	
46417	638555941454000128	Good morning! Happy September everyone! Have a ...	when is september	
52348	638564084586315776	We made it. It's September.	when is september	
54844	638567334148505600	September, then. Let's give September a try.	when is september	
55792	638568397048020993	RT @MRotellaWx: Happy September!	when is september	
56271	638569352934072320	September is here and so am I at the trav desk...	when is september	
57358	63856963670769668	Come September!! Gud morn	when is september	
94647	63861176663073797	#RipLesVacances #SeptemberIssue #augustisgone ...	when is september	
100544	638618208367431681	Happy September 1st to all my sweet tweeters ,...	when is september	
128128	638644881838227457	September whaaaaat?	when is september	
136242	638652269559443456	Wow! Can you believe it is already September? ...	when is september	
137035	638653125289750528	Can you believe it's already September?! Star...	when is september	
137299	638653496791707648	Yeah. Sure. That looks just like September 1.	when is september	
141885	638657491530817536	SEPTEMBER ALREADY? Seems like a good day to pl...	when is september	
142530	638657896859987973	5 things to know for Tuesday, September 1, 2015	when is september	
152879	638667507218411520	Welcome to September! It's a month of change a...	when is september	

2

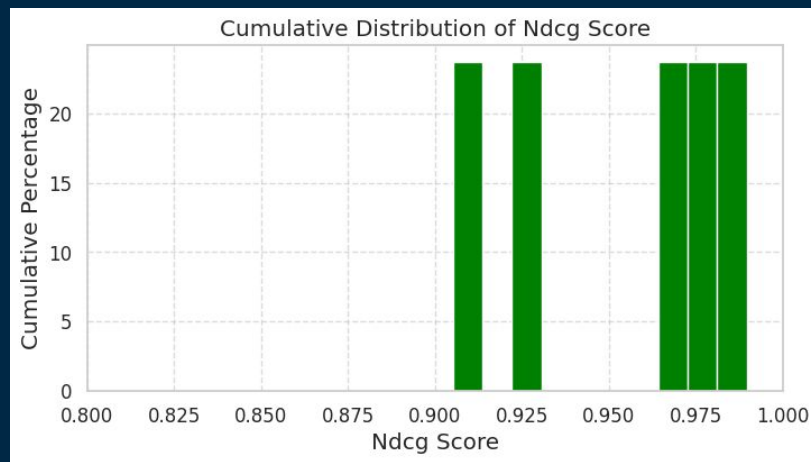
RANKING WITH BM25 AND RE-RANKING WITH BERT (REDUCED)

BM25




Mean: 0.90

BM25 AND DISTILBERT



Mean: 0.952

Conclusions



Topic modelling and **sentiment analysis** for better contextual understanding

Understanding user intent

More accurate and meaningful retrieval



Computational considerations

Careful consideration not only to performance requirements but also to resource limitations

Challenges in Query-Document Assignment

Continuous evaluation → Effectiveness

BERT Reranking Unveils Semantic Context

High efficacy in capturing semantic context





Do you have any questions?

THANKS

TU Graz, Advanced Information Retrieval (706.705)

Group 09

Ángela González de Diego - 12307642

Claudia Meana Iturri

Cristina Vera Zambrano - 12312000