Fact Finder - Fact Search Engine

A Comparative Study of Retrieval Techniques for Fact-Checking

Information Retrieval
Academic Year 2023-2024

Nicolò Urbani 856213
Mattia Piazzalunga 851931

Fact or not?

Elon Musk

stated on February 16, 2024 in a post

> **There is “clear scientific consensus” that “hormonal birth control makes you fat, doubles risk of depression and triples risk of suicide.”**

By Samantha Putterman • February 23, 2024

Elise Stefanik

stated on January 28, 2024 in a television interview

> **“We’ve seen an 800% increase in the Swanton sector, which is the part of the northern border that I represent, in illegal crossings.”**

By Jill Terreri Ramos • February 23, 2024

Source: https://www.politifact.com/

Fact and the problem of fake news

FACT. "something that is known to have happened or to exist, especially something for which proof exists, or about which there is information" - Cambridge Dictionary

Social Media

Social Increase the diffusion of fake news

Absence of Verification

Just 23 % of young people verify the Truthfulness news

Big influence on Society

Influence public opinion, harm the reputation, and even pose a threat to national security.

Fake news

Source: https://ec.europa.eu/eurostat/

Project Purpose

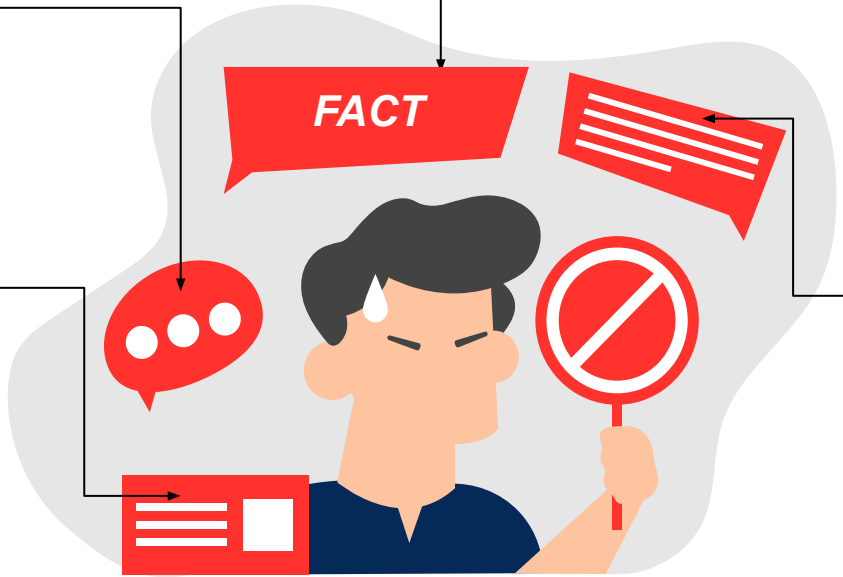
The objective of this project is to compare and evaluate the performance of different Basic and Neural Retrieval techniques and determine which technique is the most effective for fact retrieval

Analysis Fact Language

Fake Fact use a specific language?

Preprocessing Impact

Does preprocessing improve the performance?



Compare Retrieval Models

"Which model achieves the best effectiveness?"

Neural Models Impact

Does the use of neural model impact?

The original FEVER (Fact Extraction and VERification)

FEVER (Fact Extraction and Verification) consists of **185,445 claims** generated by altering sentences extracted from Wikipedia and subsequently verified without knowledge of the sentence they were derived from. The claims are classified by expert in:

- Supported (supported by documents)
- Refuted (supported by documents)
- Not Enough Info

The dataset contains the following attributes:

- id: the identifier of claim
- calim: the text of the claim; could be incomplete
- label: Supported, Refuted, NotEnoughInfo
- Evidence: a set of document wich support or refute the claim

Source: 2017, Thorne et al. - https://aclanthology.org/N18-1074/

Example

Supports

```
{
    "id": 62037,
    "label": "SUPPORTS",
    "claim": "Oliver Reed was a film actor.",
    "evidence": [
        [
            [<annotation_id>, <evidence_id>,
"Oliver_Reed", 0]
        ],
        [
            [<annotation_id>, <evidence_id>,
"Oliver_Reed", 3],
            [<annotation_id>, <evidence_id>,
"Gladiator_-LRB-2000_film-RRB-", 0]
        ],

    ]
}
```

Refuted

```
{
    "id": 78526,
    "label": "REFUTES",
    "claim": "Lorelai Gilmore's father is named
Robert.",
    "evidence": [
        [
            [<annotation_id>, <evidence_id>,
"Lorelai_Gilmore", 3]
        ]
    ]
}
```

Not Enough Info

```
{
    "id": 137637,
    "label": "NOT ENOUGH INFO",
    "claim": "Henri Christophe is recognized
for building a palace in Milot.",
    "evidence": [
        [
            [<annotation_id>, <evidence_id>,
null, null]
        ]
    ]
}
```

BEIR datasets

A robust and heterogeneous evaluation benchmark for information retrieval.

Problem

IR models have often been studied in homogeneous and narrow settings

Datasets

They leverage a careful selection of 18 publicly available datasets

Extra results

BM25 is a robust baseline and re-ranking and late-interaction based models on average achieve the best zero-shot performances, however, at high computational costs.

BEIR

Source: 2018, Thakur et al. - https://aclanthology.org/N18-1074/

How FEVER dataset for IR was built?

Random Wikipedia Claim

Each Claim

Manual Annotation

Label + Eventual Evidence

FEVER Complete Dataset

185,445 claims + Evidence

FEVER IR

FEVER IR- DATASET

DOCS

QUERIES

QRELS

Using Supported and Refuted

SPLITTING

TRAIN

VALIDATION

TEST

FEVER - a look at the size of the dataset

		DOCUMENTS	QUERIES	QRELS
FEVER IR DATASET	TRAIN	5.4M	110K	140K
	VALIDATION	5.4M	6.7K	8.1K
	TEST	5.4M	6.7K	7.9K

Dataset Sample - Test

QUERIES

query_id	
55426	Firefox is a computer game.

RELEVANCE

The documents are relevant for the proposed query

QRELS

query_id	doc_id	relevance
55426	Web_browser	1
55426	Firefox	1

doc_id	text
Web_browser	A web browser (commonly referred to as a browser) is a software application for retrieving , presenting and tra Uniform Resource Identifier (URI/URL) that may be a web page , image , video or other piece of content . Hype Although browsers are primarily intended to use the World Wide Web , they can also be used to access informa browsers are Google Chrome , Microsoft Edge (preceded by Internet Explorer) , Safari , Opera and Firefox .

doc_id	text
Firefox	Mozilla Firefox (or simply Firefox) is a free and open-source web browser developed by the Mozilla Foundation and systems , with its Firefox for Android available for Android (formerly Firefox for mobile , it also ran on the discontinu implements current and anticipated web standards . An additional version , Firefox for iOS , was released in late 201 WebKit-based layout engine built into iOS . Firefox was created in 2002 under the name `` Phoenix '' by Mozilla com Even during its beta phase , Firefox proved to be popular with its testers and was praised for its speed , security , ar 2004 , and was highly successful with 60 million downloads within nine months , which was the first time that Interne Navigator , as the Mozilla community was created by Netscape in 1998 before their acquisition by AOL . Firefox usa browser . Usage then declined in competition with Google Chrome . , Firefox has between 9 % and 16 % of worldwi most popular desktop browser in Cuba (even most popular overall) , Eritrea , and Germany , with 85.93 % , 79.39 other African countries . According to Mozilla , there were half a billion Firefox users around the world .

DOCS

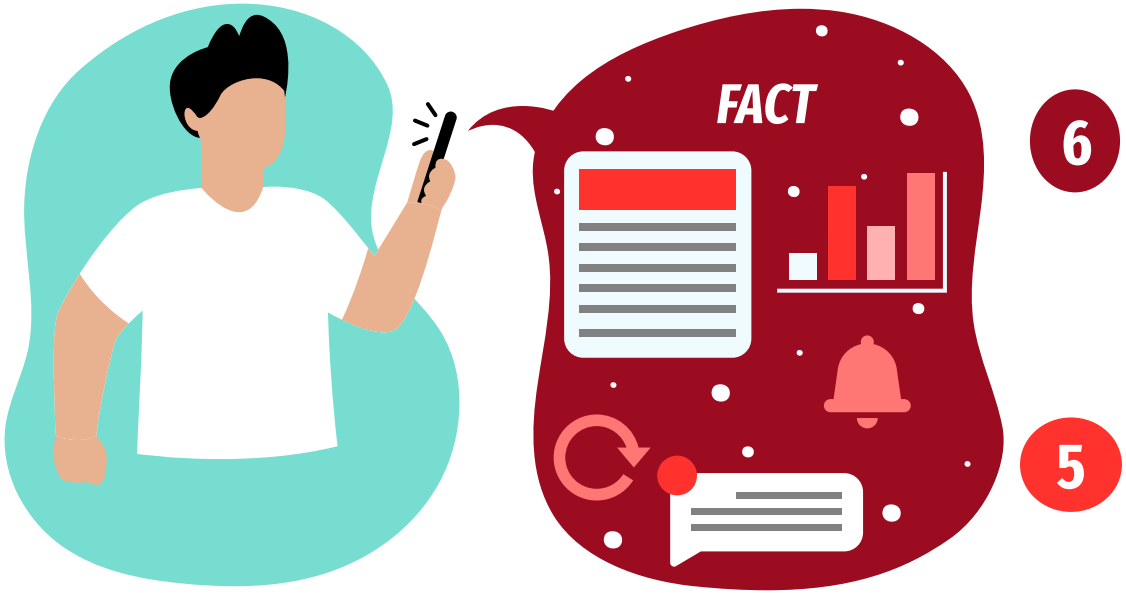
Project Steps

Explorative Analysis

Problem Understanding and first analysis

Documents Indexing

Document Indexing and related preprocessing



Basic Retrieval Models

Definition of basic Models for IR

3

4

Query Expansion

Expansion of the query using different models

Analysis of Results

Analysis of performance

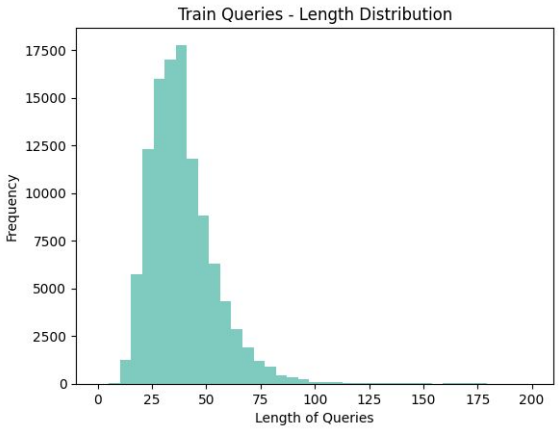
Neural Retrieval Model

Introduction of Advanced Models

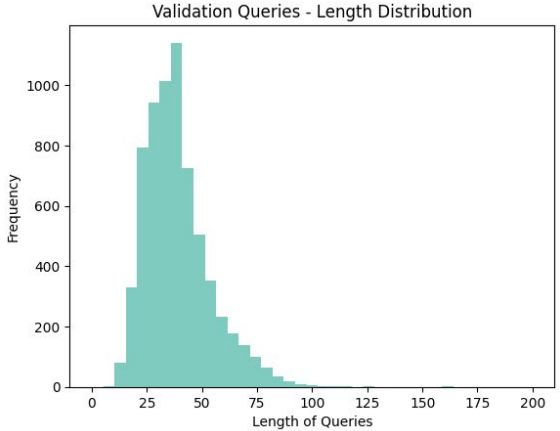
1

Explorative Analysis

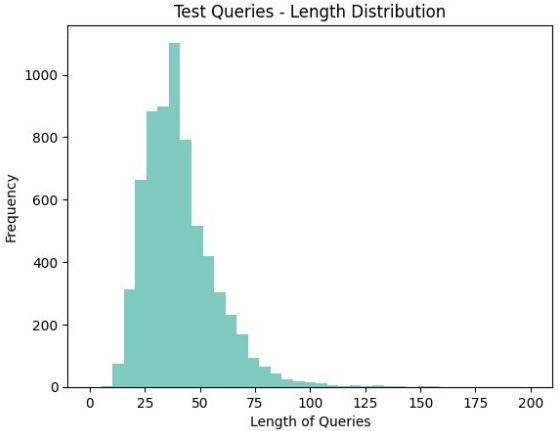
Queries Length



mean	39.038184
std	16.618665
min	9.000000
25%	28.000000
50%	36.000000
75%	47.000000
max	516.000000

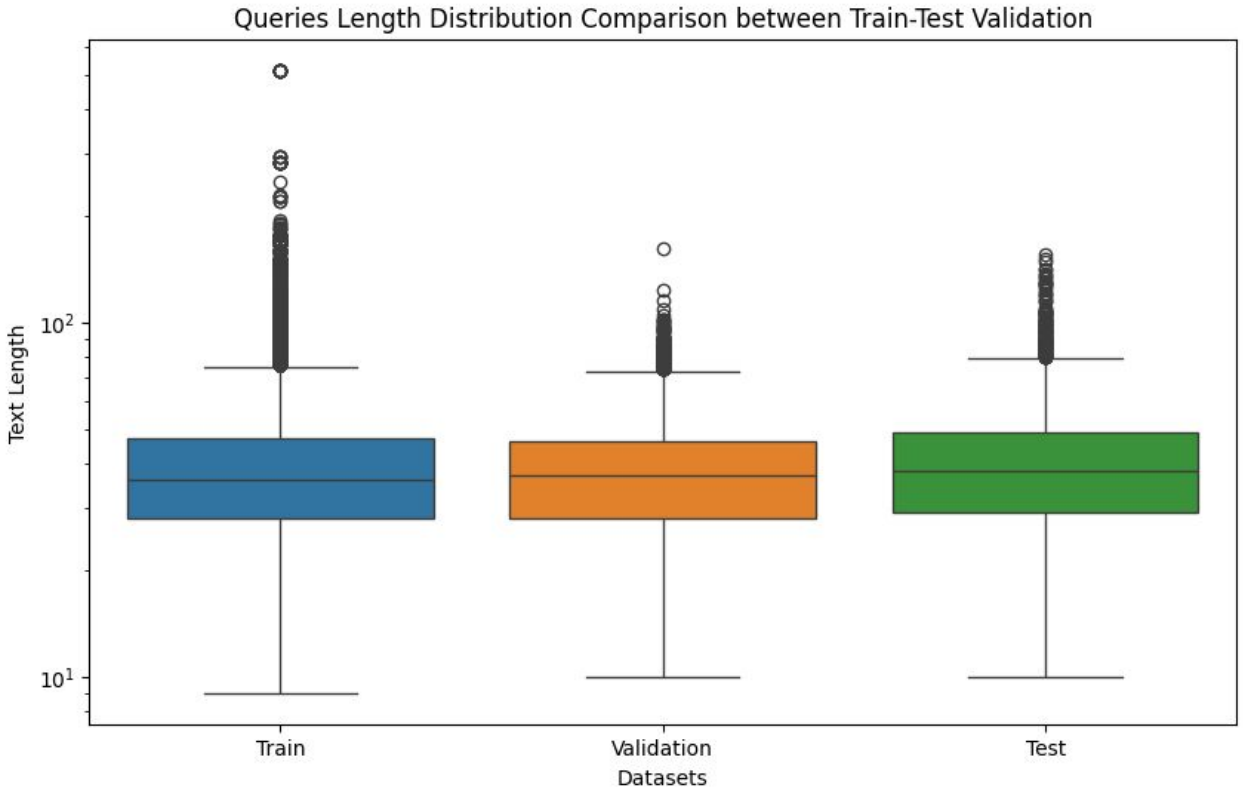


mean	38.776628
std	14.795147
min	10.000000
25%	28.000000
50%	37.000000
75%	46.000000
max	162.000000



mean	40.722622
std	16.452091
min	10.000000
25%	29.000000
50%	38.000000
75%	49.000000
max	155.000000

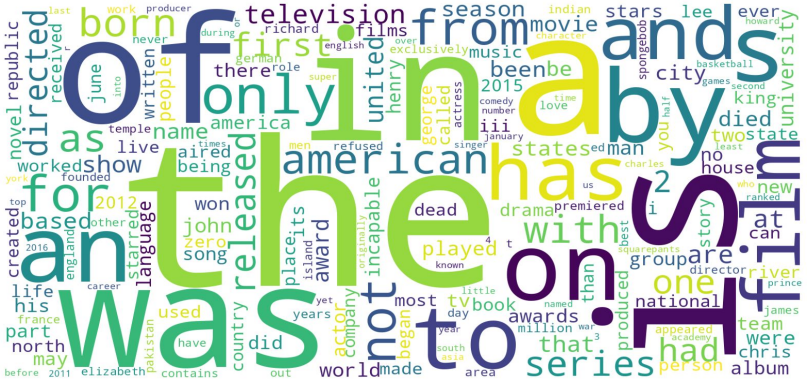
Queries Comparison



Most Frequent Words in Queries

Train

Test



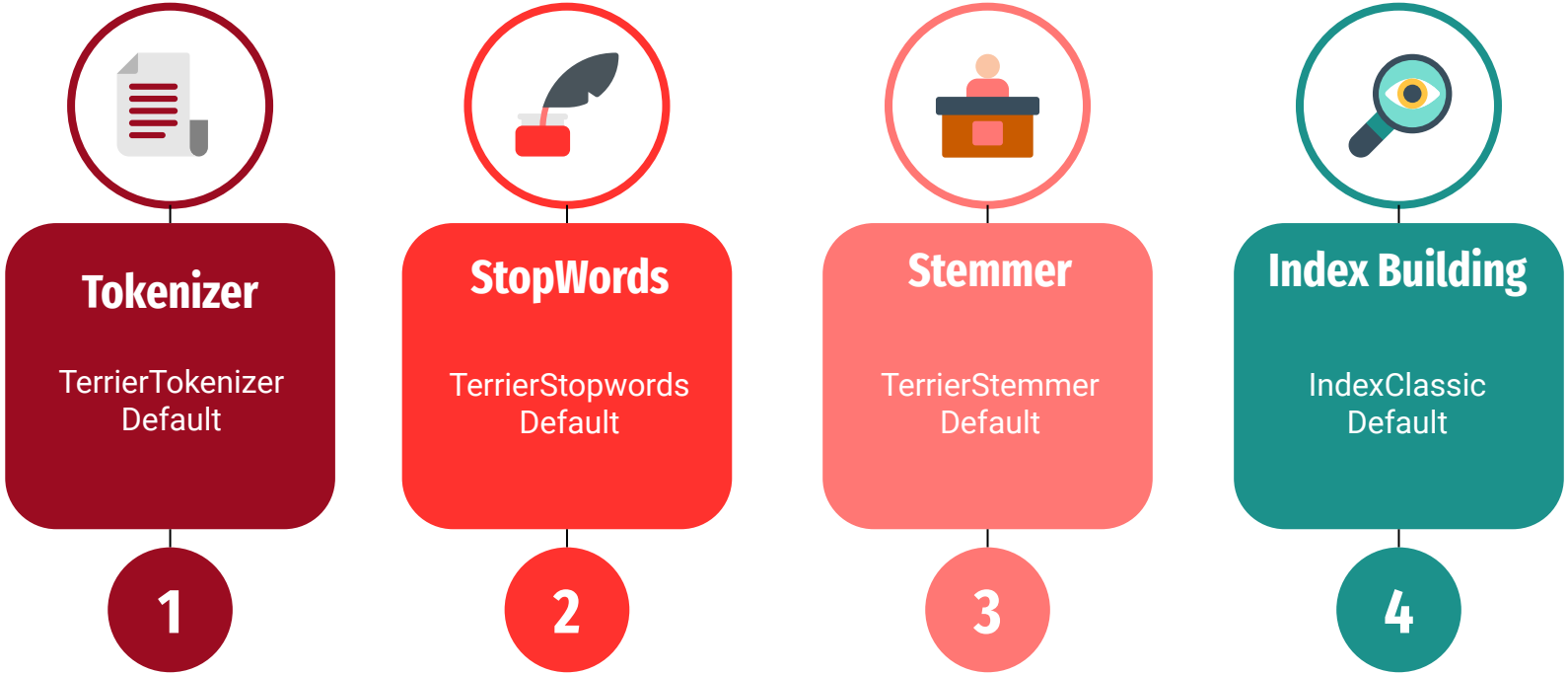
Validation

2

Documents Indexing

Dataset Preprocessing

NOT USED IN INDEX FOR NEURAL MODEL



Preprocessing & Tokenization

Basic text processing is often the first step in any text mining application and consists of several levels. The document is to be made computable by representing it in a formal way: it is necessary to access the words

The tokens

First, the text is divided into tokens, which are the basic units. The tokens are candidates to be meaningful for our studies

Problems

The tokenizer, therefore, is highly dependent on the language, documents, and context

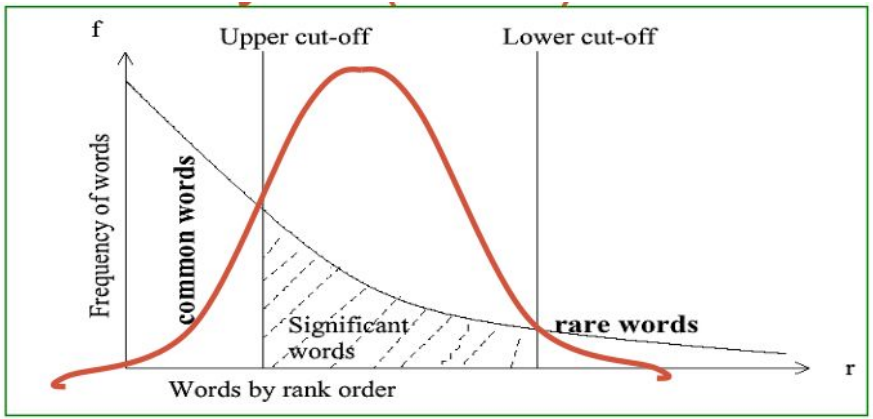
Our tokenizer

By default PyTerrier uses space-based tokanization. For languages that use space between words, such as Western languages, it is an excellent split point.

Tokenization

Source: https://pyterrier.readthedocs.io/en/latest/text.html

Stop Words Removal - Zipf's Law & Luhn's Analysis



Why Stop Word Removal is used?
Source: 2019, Quiao et al. - https://arxiv.org/abs/1904.07531

What is Zipf's Law ?
Source: 1949, Zipf - https://psycnet.apa.org/record/1950-00412-000

What is Luhn's analysis?
Source: 1958, Luhn - https://ieeexplore.ieee.org/document/5392672

Our decision

Very frequent words are removed from the indexes (upper cut-off). Stopwords removal, however, is not applied on index for neural models considering that context becomes central in neural model as cited in *2019, Quiao et al*

Stemming

Stemming is the process of reducing the inflected form of a word to its root form

VS lemmatization

lemmatization produces better, but not enough to overlook the computational cost.

Benefits

Decreases dictionary size, improves recall, but worsens precision.

Porter stemmer

The algorithm is based on analyzing patterns of vowel-consonant sequences and is at least as good as all the other stemming options.

Stemming

Source: 1980, Porter - https://www.emerald.com/insight/content/

Stemming

Stemming is applied in **both basic and neural models** considering that it could improve the performance.
2016, Flores et al. - https://doi.org/10.1016/j.ipm.2016.03.004

BEFORE

law enforcement alert threats cops whites by black lives matter terrorists
comment expected barack obama fukyoflag black lives matter movements called lynching hanging white people cops encouraged radio tuesday night tide kill white people cops send message killing black people america fyoflag organizers called sunshine radio blog hosted texas called sunshine fing opinion radio snapshot twitter fiftythree aftermidday urging supporters…….

AFTER

law enforcement alert threat cop white black lives matter terrorist

comment expect barack obama fukyoflag black lives matter movement call lynching hanging white people cop encourage radio tuesday night tide kill white people cop send message kill black people america fyoflag organizer call sunshine radio blog hosted texas call sunshine fing opinion radio snapshot fyf911 twitter fiftythree aftermidday urge supporter….

Bag-of-Words & Inverted file structure

Bag-of-Word representation aims to convert text into a numerical form understandable by machine learning algorithms.
BoW treats text as an unordered set of words.

BoW problems

Polysemic words, not considering the context of the word, are a problem.

Why Inverted file

We cannot have a static document/term structure because it is highly scattered and access is inefficient.

Inverted file

The inverted file structure associates term/document and organizes them into a dynamic structure consisting of dictionary and posting list

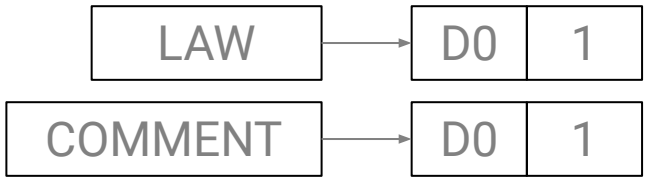
What about

Preprocessing, inverted file structure - Example

	title	text
0	LAW ENFORCEMENT ON HIGH ALERT Following Threat...	No comment is expected from Barack Obama Member...
1	NaN	NaN
2	NaN	Did they post their votes for Hillary already? ....

INDEXING AND PREPROCESS

Title + Text



Our Indexes

FEVER DOCUMENTS COLLECTION 5.4M

DOC1

DOC2

DOC3

DOC4 ....

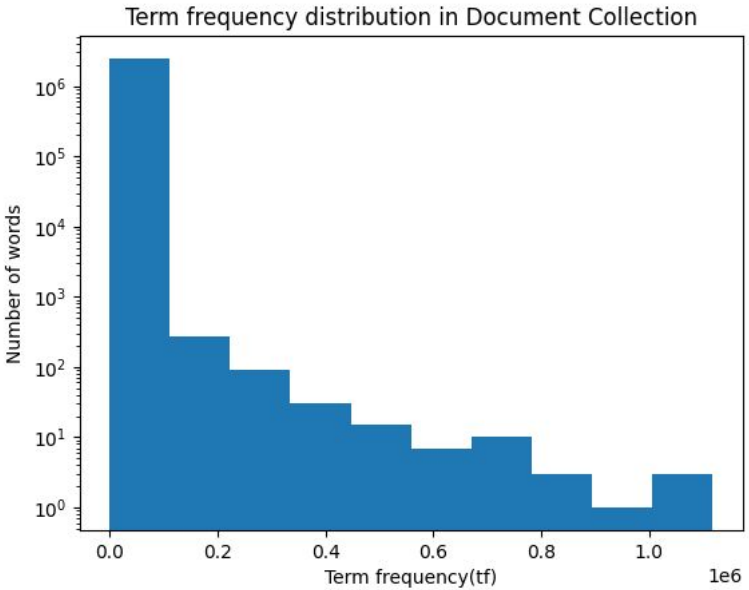
```
indexer = pt.IterDictIndexer(index_path,
meta={'docno': 300})
index_ref =
indexer.index(dataset.get_corpus_iter(),
fields=['text'])
```

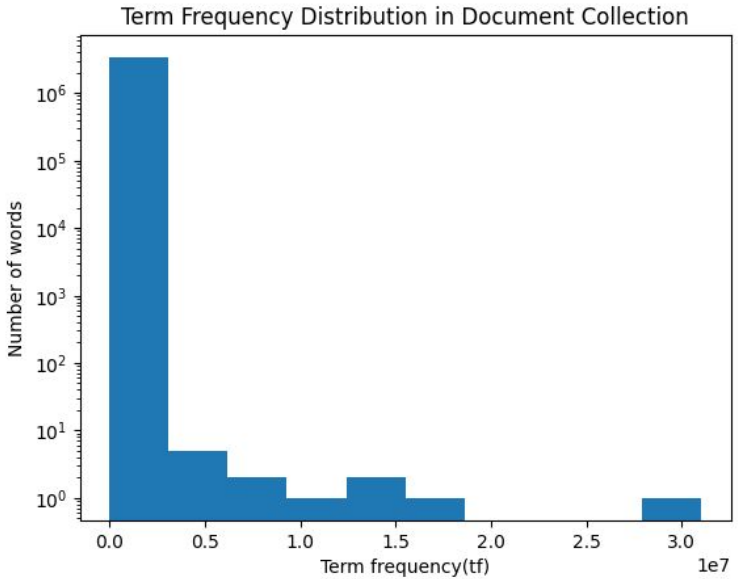
INDEX FOR BASIC MODELS

```
indexer = pt.IterDictIndexer(index_path2,
meta={'docno': 300}, stopwords=None)
index_ref =
indexer.index(dataset.get_corpus_iter(),
fields=['text'])
```

INDEX FOR NEURAL MODELS

Words Distribution in Documents Collection





Index 1

Number of documents: 5416568
Number of terms: 2471240
Number of postings: 203556545
Number of fields: 1
Number of tokens: 269889695
Field names: [text]

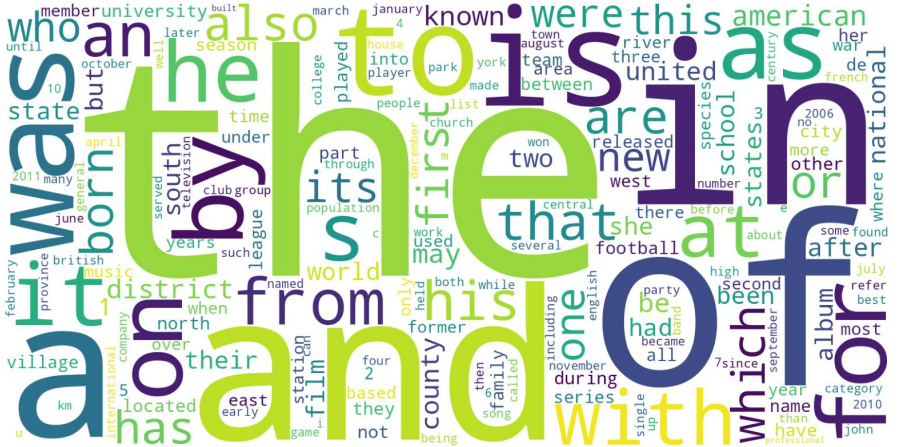
← 152 STOP WORDS →

Index 2

Number of documents: 5416568
Number of terms: 2471392
Number of postings: 284943850
Number of fields: 1
Number of tokens: 453148161
Field names: [text]

Word-Cloud in Documents Collection

Index 1



Index 2

2

Basic Retrieval Models

IR Models

Information Retrieval (IR) models are used to represent and retrieve documents or information from a collection based on a user's query. A model is a formal, concise representation.

The query

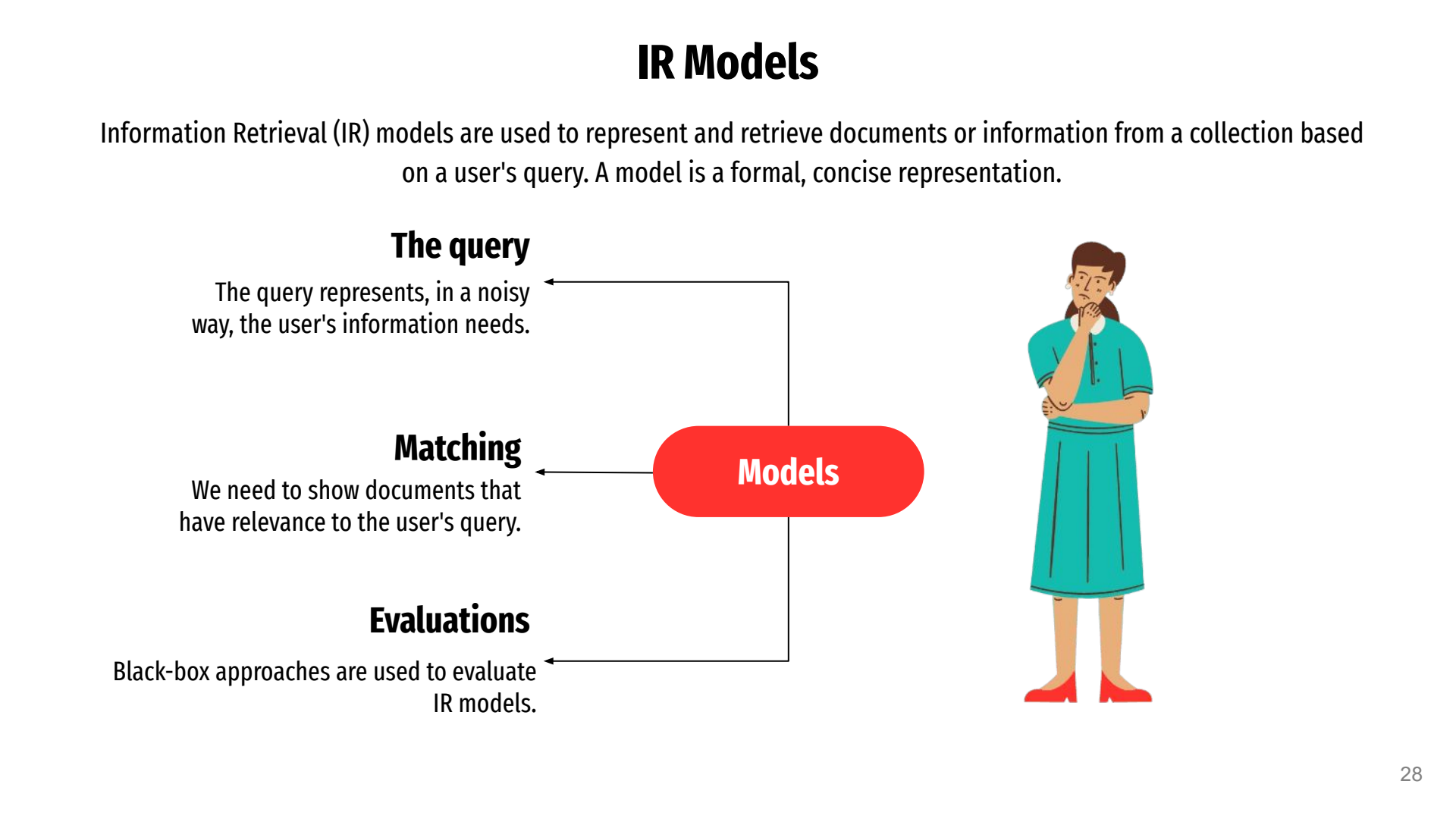
The query represents, in a noisy way, the user's information needs.

Matching

We need to show documents that have relevance to the user's query.

Evaluations

Black-box approaches are used to evaluate IR models.



Workflow

Indexing	Using the whole collection of 5.4M Documents	1
Model Definition	Definition of the model and hyperparameter tuning	2
Experiments	Experiment using test to evaluate the models	3

Methodological Approach

Does Query Expansion and Neural Models improve the performance ?

1

Basic Retrieval Models

Evaluate Basic Retrieval models

2

Query Expansion

Evaluate the models adding related terms to a user's search query.

3

Neural Retrieval Models

Evaluate the models using neural models

Does each classifier perform well improving the search query?

How does it perform using neural models?

Simple Retrieval Models

Each model will be tested using
BATCH SIZE = 264

1

TF-IDF

Basic Model

2

BM25

Basic Model considering the Saturation

3

TUNED BM25

Tuned Version of BM25 to improve performance

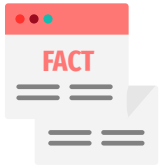
4

QUERY EXPANSION

Use of BM25 + B01 Query Expansion

1

TF-IDF



TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents. It is composed by two component: TF (Document Wise) & IDF (Corpus Wise). Effective for small to medium-sized document collections. **Does not consider term saturation**, which can lead to skewed relevance scores for longer documents.

Firefox is a computer game

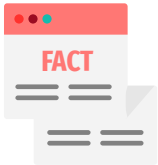
$$w_{t,d} = \frac{tf_{t,d}}{\max_{t_i \in d} tf_{t_i,d}} \qquad idf_t = \log\left(\frac{N}{df_t}\right)$$

	qid	docid	docno	rank	score	query
0	1	1661512	Firefox_Portable	0	33.148275	Firefox is a computer game
1	1	1643760	Firefox_for_Android	1	31.226661	Firefox is a computer game
2	1	1693421	Firefox_Sync	2	29.974616	Firefox is a computer game
3	1	1705595	FireTune	3	28.465207	Firefox is a computer game
4	1	4212581	River_Trail_(JavaScript_engine)	4	27.062045	Firefox is a computer game
5	1	1687831	Firefox_3.6	5	26.267953	Firefox is a computer game

Source: 1972, Jones - https://www.staff.city.ac.uk/~sbrp622/

1

Result

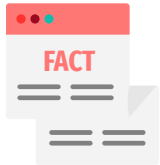


River_Trail_(JavaScript_engine)	River Trail (also known as Parallel JavaScript) is an open source software engine designed by Intel for executing JavaScript code using parallel computing on multi-core processors . River Trail was announced at the Intel Developer Forum in September 2011 , and demonstrated using a Firefox extension developed by Intel . Brendan Eich , the original author of JavaScript , promised that he would promote River Trail within Ecma International , saying `` The demo shows a 15x speedup over serial JavaScript . It lights up the ridiculously parallel hardware in modern CPUs and GPUs , for audio , video , image processing , automated voice response , computer vision , 3D gaming , etc. -- all written in memory-safe , clean , functional JavaScript , without threads and their data races and deadlocks . ....

Fire Tune	FireTune is a Firefox add-on , which aims at optimizing the speed of the browser by changing its settings based on the user 's preferences . The user is first invited to choose from a list what best describes his/her computer configuration such as `` fast computer / fast connection '' , `` slow computer / fast connection '' , etc. . FireTune then attempts to adjust the settings of Firefox to best matches the user 's computer - thus improving the performances of the browser . As of 17 December 2009 , Totalidea no longer distributes or supports FireTune . According to the Totalidea website : `` Because the Mozilla Foundation disallows us to show the Firefox logo within our FireTune software , we are no longer able to distribute FireTune , otherwise we would face legal actions initiated by Mozilla . Because of that we have removed the FireTune product from our product catalogue and do no longer offer it for download . Downloads of FireTune from third party websites are out of our control . ''

1

Expected Outcome



TEST SET

	query_id	text
4266	55426	Firefox is a computer game.

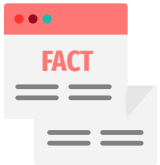
QRELS

	query_id	doc_id	relevance	iteration
5072	55426	Web_browser	1	0
5073	55426	Firefox	1	0

Web_browser	A web browser (commonly referred to as a browser) is a software application for retrieving , presenting and traversing information resources on the World Wide Web . An information resource is identified by a Uniform Resource Identifier (URI/URL) that may be a web page , image , video or other piece of content . Hyperlinks present in resources enable users easily to navigate their browsers to related resources . Although browsers are primarily intended to use the World Wide Web , they can also be used to access information provided by web servers in private networks or files in file systems . The most popular web browsers are Google Chrome , Microsoft Edge (preceded by Internet Explorer) , Safari , Opera and Firefox .
Firefox	Mozilla Firefox (or simply Firefox) is a free and open-source web browser developed by the Mozilla Foundation and its subsidiary the Mozilla Corporation . Firefox is available for Windows , macOS and Linux operating systems , with its Firefox for Android available for Android (formerly Firefox for mobile , it also ran on the discontinued Firefox OS) ; where all of these versions use the Gecko layout engine to render web pages , which implements current and anticipated web standards . An additional version , Firefox for iOS , was released in late 2015 , but this version does not use Gecko due to Apple 's restrictions limiting third-party web browsers to the WebKit-based layout engine built into iOS . Firefox was created in 2002 under the name `` Phoenix '' by Mozilla community .....

2

BM25



BM25 is an extension of the TF-IDF model and addresses some of its limitations, such as term saturation. BM25 is a probabilistic model, scalable and performs well on larger collections of documents compared to TF-IDF. BM25 is so **effective as well as efficient** that it is used as a **baseline in many IR tasks**.

Firefox is a computer game

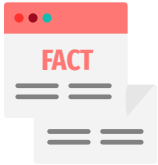
$$RSV_d = \sum_{t \in q} \left[\log \frac{N}{\mathrm{df}_t} \right] \cdot \frac{(k_1+1)\mathrm{tf}_{td}}{k_1((1-b)+b \times (L_d/L_{\mathrm{ave}}))+\mathrm{tf}_{td}} \cdot \frac{(k_3+1)\mathrm{tf}_{tq}}{k_3+\mathrm{tf}_{tq}}$$

	qid	docid	docno	rank	score	query
0	1	1661512	Firefox_Portable	0	18.107317	Firefox is a computer game
1	1	1643760	Firefox_for_Android	1	17.049271	Firefox is a computer game
2	1	1693421	Firefox_Sync	2	16.364326	Firefox is a computer game
3	1	1705595	FireTune	3	15.552895	Firefox is a computer game
4	1	4212581	River_Trail_(JavaScript_engine)	4	14.816217	Firefox is a computer game
5	1	1687831	Firefox_3.6	5	14.337230	Firefox is a computer game

Source: 1992, Robertson - https://nvlpubs.nist.gov/

3

Tuned BM25



BM25 provides better tuning options, allowing for more precise control over the retrieval process. Solutions to **mitigate the saturation problem** include parameter tuning, such as adjusting the **k1 and b parameters**, or considering alternative weighting schemes.

k1: This parameter **controls the scaling of term frequency (TF)** within the BM25 formula. A higher value of k1 increases the impact of term frequency on the relevance score. Values for k1 range between 0 and 2.0 (0 binary model, large value raw term frequency)

b: This parameter controls the scaling of document **length normalization**. A higher value of b means more aggressive length normalization, while a lower value results in less normalization. Common values for b range between 0.5 and 0.8.

k3 (optional): Some implementations of BM25 include an additional parameter k3, which is used to **scale the term frequency in query**

```
pt.GridSearch(
    BM25,
    {BM25: {"bm25.b"   : [0,   0.5, 1 ],
            "bm25.k_1": [0.3,   1.2, 2],
            "bm25.k_3": [0.5, 10, 20]
    }},
    dataset_validation.get_topics(),
    dataset_validation.get_qrels(),
    "map")
```

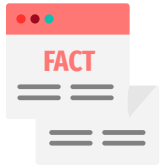
TUNED BM25

```
pt.BatchRetrieve(index_ref, wmodel="BM25",

controls={"bm25.b" : 0.5, "bm25.k_1": 0.3,
"bm25.k_3": 20})
```

3

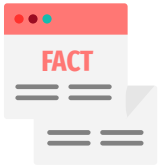
Tuned BM25



Firefox is a computer game

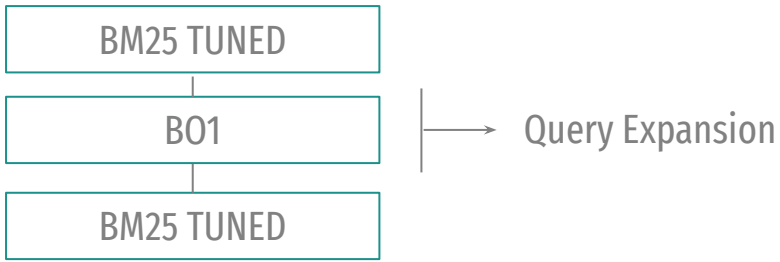
	qid	docid	docno	rank	score	query
0	1	4212581	River_Trail_(JavaScript_engine)	0	26.035720	Firefox is a computer game
1	1	4536369	Spatial_navigation	1	24.418233	Firefox is a computer game
2	1	1696167	FF2	2	24.252484	Firefox is a computer game
3	1	1661512	Firefox_Portable	3	24.179637	Firefox is a computer game
4	1	3816930	PlayCanvas	4	24.160745	Firefox is a computer game

4 QUERY EXPANSION - BO1



Query Expansion (QE) is a method used for improving Information Retrieval (IR) by **adding the terms** that are almost selected from **feedback documents**, and similar to the user query terms, in order to improve the effectiveness of the search ranking.

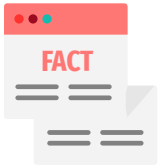
Bo1 (Bo1 (Bag of One) divergence from Randomness query expansion model to rewrite the query based on the occurrences of terms in the feedback documents provided for each query (qrels)



Source: 2005, Ounis et al. - https://link.springer.com/

4

QUERY EXPANSION - BO1



Firefox is a computer game

	qid	docid	docno	rank	score	query_0	query
0	1	4536369	Spatial_navigation	0	29.472420	Firefox is a computer game	applypipeline:off firefox^1.652677526 comput^1...
1	1	1643760	Firefox_for_Android	1	27.511093	Firefox is a computer game	applypipeline:off firefox^1.652677526 comput^1...
2	1	1696167	FF2	2	27.051403	Firefox is a computer game	applypipeline:off firefox^1.652677526 comput^1...
3	1	3816930	PlayCanvas	3	26.450392	Firefox is a computer game	applypipeline:off firefox^1.652677526 comput^1...
4	1	1657754	Firefox_OS	4	26.086535	Firefox is a computer game	applypipeline:off firefox^1.652677526 comput^1...

Evaluations

Usually to evaluate an IR system we work on **efficiency** (optimized use of resources in terms of space and time) and **effectiveness** (how well our system works to provide an output).

Subjective, yes but

Relevance is subjective, but that does not mean it is not measurable.

Cranfield experiments

We have a black-box approach to evaluating the effectiveness of a search engine based on three elements: a document collection, a suite of queries, and an assessment of relevant or not relevant.

Cranfield assumptions

The relevance of a document to a user is binary and independent of the relevance of other documents. The user is enabled to discover relevant documents in the collection without using the system

Cranfield

Evaluations

P@K

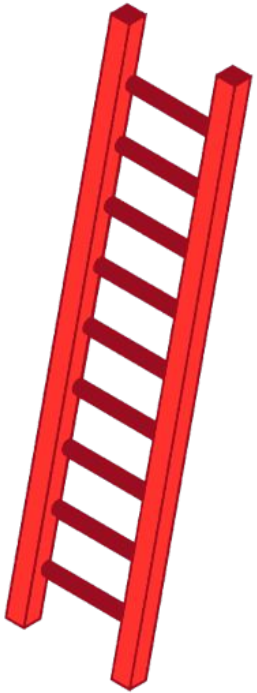
A K is set and the % of relevant documents among the first K documents returned is calculated.

R@K

A K is set and the % of relevant documents among the first K documents returned compared to the total of relevant documents is calculated

MAP

To calculate average precision, we consider the position k(i) of all relevant documents for a single query and calculate the P@K for each one. The average precision is the average of the P@K(i).



MRR

We consider the first position K of a relevant document, the Reciprocal Rank Score is 1/k.

NDCG

In the DCG we have a measure of the degree of relevance. We can express it in correlation with a discount on the profit depending on the position of the document.
Normalization, compared to the ideal ranking, NDCG allows to obtain values in a range [0, 1] and contrast queries with a variable number of relevant results

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

* Metrics averaged over all queries

Comparison Simple Retrieval Models

		P@5	P@10	R@5	R@10	MAP	NDCG	MRR
SIMPLE MODEL	TF-IDF	0.11	0.06	0.50	0.60	0.37	0.47	0.39
	BM-25	0.11	0.07	0.49	0.60	0.37	0.46	0.39
	TUNED BM25	0.14	0.08	0.63	0.71	0.50	0.60	0.53
	TUNED BM25 + BO1	0.14	0.07	0.64	0.72	0.50	0.60	0.52

5

Neural Retrieval Models

Neural Retrieval Models

Reranking refers to the process of reordering the initially retrieved documents based on additional criteria or features beyond those used in the initial retrieval stage. We use **less efficient**, but **more effective ML techniques**.

1

KNRM

Basic Reranker not re-trained

2

BI-ENCODER

Sentence Transformer Model

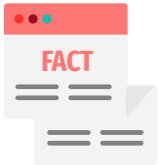
3

EPIC

Bert Lazy Epic reranker

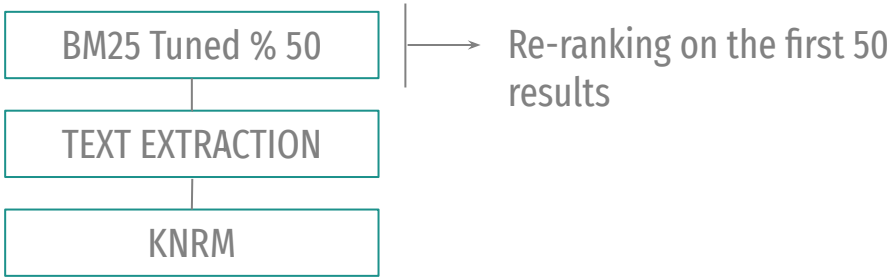
1

KNRM



The KNRM (Kernelized Neural Ranking Model) reranker is a technique used in information retrieval and search engine systems. It's designed to improve the relevance of search results by re-ranking them based on a neural network model that **captures semantic similarities between queries and documents**.

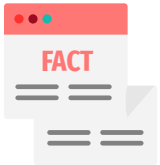
KNRM, like many neural ranking models, **requires training on labeled data** to learn the parameters that define its ranking function, in this case training is not applied, considering the limited computational resources.



Source: 2017, Xiong et al. - https://arxiv.org/abs/1706.06613

1

KNRM

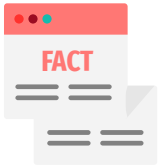


Firefox is a computer game

	qid	docid	docno	rank	score	query	text
0	1	4212581	River_Trail_(JavaScript_engine)	0	-6.700102	Firefox is a computer game	River Trail (also known as Parallel JavaScrip...
1	1	1696167	FF2	1	-6.992773	Firefox is a computer game	FF2 may refer to : Final Fantasy II , a 1988 ...
2	1	1637797	FF3	2	-6.660470	Firefox is a computer game	FF3 may refer to : Mozilla Firefox 3 Fatal F...
3	1	3816930	PlayCanvas	3	-8.257060	Firefox is a computer game	PlayCanvas is an open source 3D game engine/in...
4	1	3455306	MojoPac	4	-8.328350	Firefox is a computer game	MojoPac was an application virtualization prod...
5	1	1661512	Firefox_Portable	5	-9.664987	Firefox is a computer game	Mozilla Firefox , Portable Edition (formerly ...

1

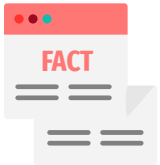
Sample Result



River_Trail_(JavaScript_engine)	River Trail (also known as Parallel JavaScript) is an open source software engine designed by Intel for executing JavaScript code using parallel computing on multi-core processors . River Trail was announced at the Intel Developer Forum in September 2011 , and demonstrated using a Firefox extension developed by Intel . Brendan Eich , the original author of JavaScript , promised that he would promote River Trail within Ecma International , saying `` The demo shows a 15x speedup over serial JavaScript . It lights up the ridiculously parallel hardware in modern CPUs and GPUs , for audio , video , image processing , automated voice response , computer vision , 3D gaming , etc. -- all written in memory-safe , clean , functional JavaScript , without threads and their data races and deadlocks .
FF2	FF2 may refer to : Final Fantasy II , a 1988 console role-playing game for the Family Computer Final Fantasy IV , retitled Final Fantasy II in North America , a 1992 console role-playing game for the Super NES Fatal Fury 2 , a 1992 competitive fighting game for the Neo-Geo Fatal Frame II , a 2003 horror adventure game for the PlayStation 2 and Xbox Final Fight 2 , a 1993 side-scrolling action game for the Super NES Fantastic Four : Rise of the Silver Surfer , the sequel to the 2005 Fantastic Four film 2 Fast 2 Furious , a 2003 film Mozilla Firefox 2 , a web browser released in 2006
FF3	FF3 may refer to : Mozilla Firefox 3 Fatal Frame III : The Tormented , a 2005 horror adventure game for the PlayStation 2 Fatal Fury 3 : Road to the Final Victory , a 1995 competitive fighting game for the Neo-Geo Final Fantasy III , a 1990 console role-playing game for the Family Computer Final Fantasy VI , retitled Final Fantasy III in North America , a 1994 console role-playing game for the Super NES Final Fight 3 , a 1995 side-scrolling action game for the Super NES The Fast and the Furious : Tokyo Drift , a 2006 film . Freedom Flotilla III , a maritime activism project regarding the blockade of the Gaza Strip Fantastic Four , a 2015 film and the third film in the Fantastic Four franchise .

1

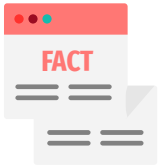
Sample Result



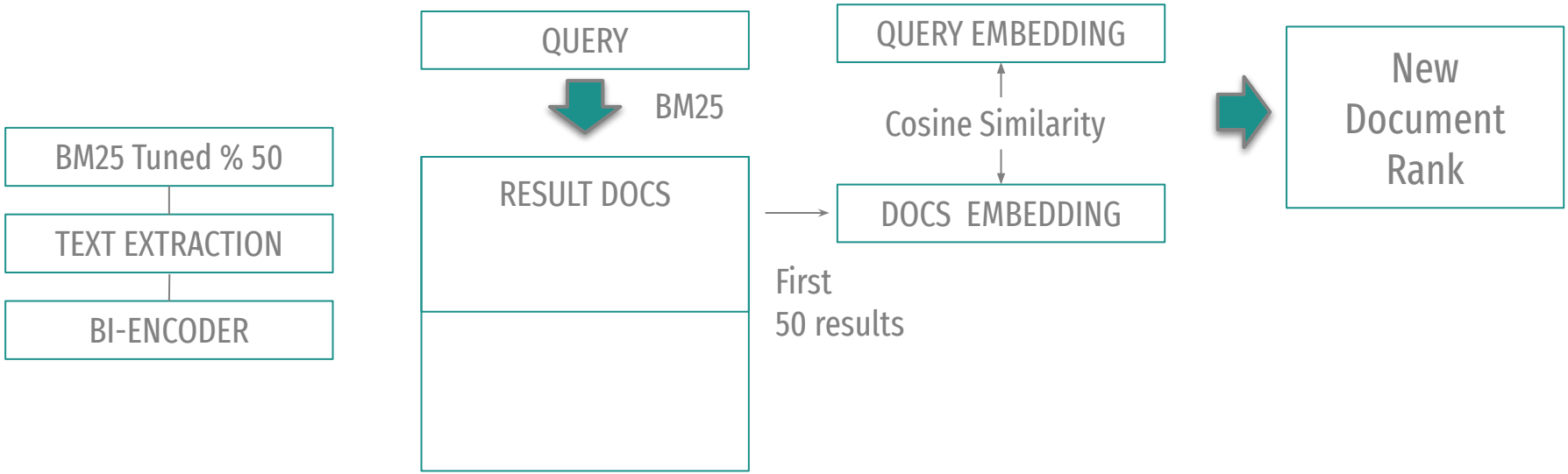
PlayCanvas	PlayCanvas is an open source 3D game engine/interactive 3D application engine alongside a proprietary cloud-hosted creation platform that allows for simultaneous editing from multiple computers via a browser-based interface . It runs in modern browsers that support WebGL , including Mozilla Firefox and Google Chrome . The engine is capable of rigid-body physics simulation , handling three-dimensional audio and 3D animations . PlayCanvas has gained the support of ARM , Activision and Mozilla . The PlayCanvas engine was open-sourced on June 4 , 2014 .
MojoPac	MojoPac was an application virtualization product from RingCube Technologies . MojoPac turns any USB 2.0 storage device into a portable computing environment . The term `` MojoPac '' is used by the company to refer to the software application , the virtualized environment running inside this software , and the USB storage device that contains the software and relevant applications . MojoPac supports popular applications such as Firefox and Microsoft Office , and it is also high performance enough to run popular PC Games such as World of Warcraft , Minecraft and Half-Life 2 . The RingCube website is currently forwarded to Citrix , which has apparently purchased the company and discontinued MojoPac .
Firefox_Portable	PlayCanvas is an open source 3D game engine/interactive 3D application engine alongside a proprietary cloud-hosted creation platform that allows for simultaneous editing from multiple computers via a browser-based interface . It runs in modern browsers that support WebGL , including Mozilla Firefox and Google Chrome . The engine is capable of rigid-body physics simulation , handling three-dimensional audio and 3D animations . PlayCanvas has gained the support of ARM , Activision and Mozilla . The PlayCanvas engine was open-sourced on June 4 , 2014 .

2

Bi-Encoder



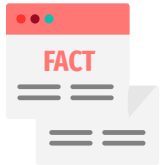
all-MiniLM-L12-v2 this is a **sentence-transformers model**: It maps sentences & paragraphs to a 384 dimensional dense vector space and can be used for tasks like clustering or semantic search. It was pretrained on microsoft/MiniLM-L12-H384-uncased model (above 1 billion sentences).



Source: 2020, Wang et al. - https://arxiv.org/abs/2012.15828

2

Bi-Encoder



Firefox is a computer game

	qid	docid	docno	score	query	text	rank
39	1	1666452	Firefox_(video_game)	0.687985	Firefox is a computer game	Firefox is a single player arcade laserdisc ga...	0
8	1	1643760	Firefox_for_Android	0.599837	Firefox is a computer game	Firefox for Android (codenamed Fennec) is th...	1
19	1	1687831	Firefox_3.6	0.596709	Firefox is a computer game	Mozilla Firefox 3.6 is a version of the Firefo...	2
15	1	1657754	Firefox_OS	0.584208	Firefox is a computer game	Firefox OS (project name : Boot to Gecko , al...	3
5	1	1661512	Firefox_Portable	0.577272	Firefox is a computer game	Mozilla Firefox , Portable Edition (formerly ...	4

3

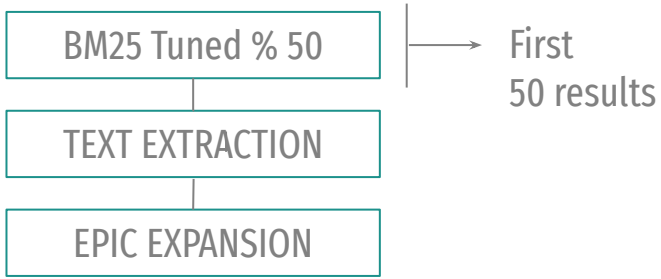
EPIC

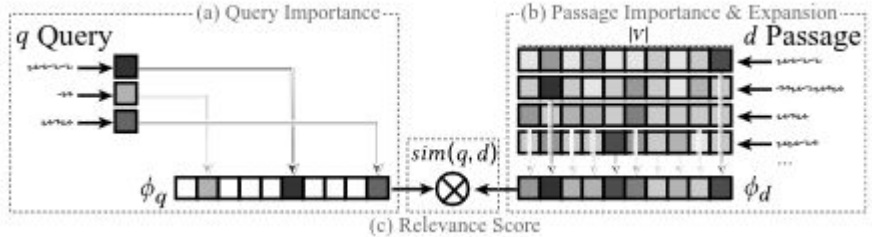
EPIC (Expansion via Prediction of Importance with Contextualization) passage retrieval faces difficulties due to **limited contextual information**. Traditional approaches struggle with relevance identification

Step:

1. Explicit modeling of **term importance** using contextualized language model.
2. Passage **expansion** by propagating importance to similar terms.
3. Grounding representations in the lexicon for interpretability.

Long Computation Time: around 4 hours with Kaggle GPU P100

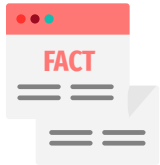




Source: 2020, MacAnvey, Nardini, Perego, Tonellotto e Goharian
https://arxiv.org/abs/2004.14245

3

EPIC



Firefox is a computer game

	qid	docid	docno	rank	score	query	text
0	1	4212581	River_Trail_(JavaScript_engine)	0	41.371147	Firefox is a computer game	River Trail (also known as Parallel JavaScrip...
1	1	1696167	FF2	1	49.716286	Firefox is a computer game	FF2 may refer to : Final Fantasy II , a 1988 ...
2	1	1637797	FF3	2	48.159203	Firefox is a computer game	FF3 may refer to : Mozilla Firefox 3 Fatal F...
3	1	3816930	PlayCanvas	3	43.518269	Firefox is a computer game	PlayCanvas is an open source 3D game engine/in...
4	1	3455306	MojoPac	4	38.385441	Firefox is a computer game	MojoPac was an application virtualization prod...

Comparison Basic & Neural Models

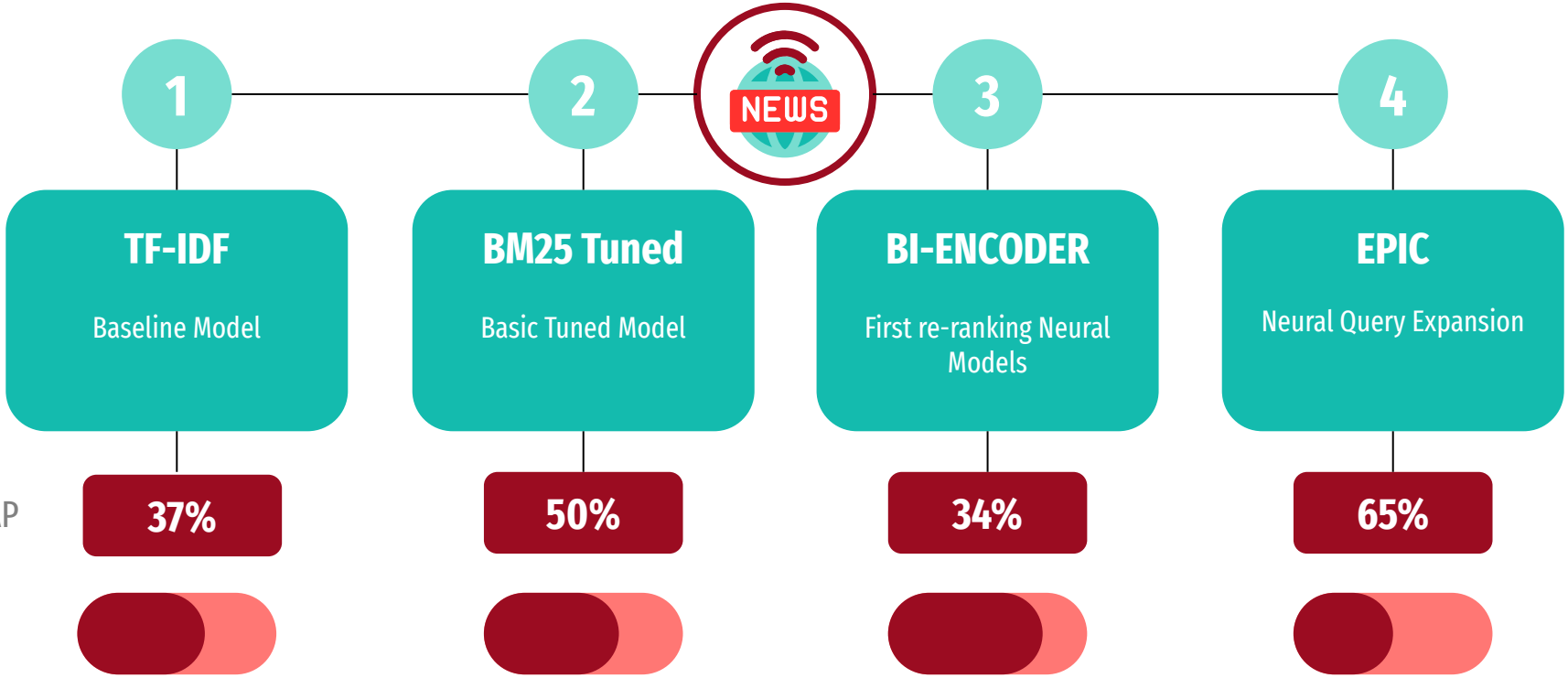
		P@5	P@10	R@5	R@10	MAP	NDCG	MRR
SIMPLE MODELS	TF-IDF	0.11	0.06	0.50	0.60	0.37	0.47	0.39
	BM-25	0.11	0.07	0.49	0.60	0.37	0.46	0.39
	TUNED BM25	0.14	0.08	0.63	0.71	0.50	0.60	0.53
	TUEND BM25 + BO1	0.14	0.07	0.64	0.72	0.50	0.60	0.52
NEURAL MODELS	KNRM	0.006	0.006	0.03	0.06	0.04	0.17	0.04
	BI-ENCODER	0.09	0.05	0.42	0.48	0.34	0.44	0.36
	EPIC	0.17	0.09	0.76	0.80	0.65	0.69	0.68

6

Analysis of Result

Final Analysis

For each of the previous analysis, the best retrieval model appears to be the expansion using **EPIC**, retrieving the most relevant information



Conclusion

Combat the Spread of Misinformation

Social media and online platforms

Deep Learning Approaches

Great Potential

Preprocessing Impact

Importance of preprocessing

Evaluation in Real Use

Importance of user feedback

Feature Development

Limited Resources and larger models

Extra ...
An ML project

Fact Retrieval vs Fact Detection

Two **distinct tasks** within the field of information retrieval and natural language processing, each with its own objectives and methodologies. Fact detection involves **assessing the truthfulness of specific claims** or statements, fact retrieval focuses on **retrieving relevant factual information** from vast collections of documents or knowledge bases.

FAKE (NOT TRUE)

Previous Work using NLP Pipeline and BERT (Accuracy 96%)

	qid	docid	docno	rank	score	query
0	1	1661512	Firefox_Portable	0	33.148275	Firefox is a computer game
1	1	1643760	Firefox_for_Android	1	31.226661	Firefox is a computer game
2	1	1693421	Firefox_Sync	2	29.974616	Firefox is a computer game
3	1	1705595	FireTune	3	28.465207	Firefox is a computer game
4	1	4212581	River_Trail_(JavaScript_engine)	4	27.062045	Firefox is a computer game
5	1	1687831	Firefox_3.6	5	26.267953	Firefox is a computer game

Feature Development

Does combining both approaches lead to an improvement in fact-checking effectiveness?

Project Purpose

The objective of this project is to compare and evaluate the performance of different Deep Learning and NLP techniques on various datasets and determine which technique is the most effective across different datasets for fake news classification

Analysis Fake News Language

Fake News use a specific language?

Preprocessing Impact

Does NLP pipeline improve the performance?

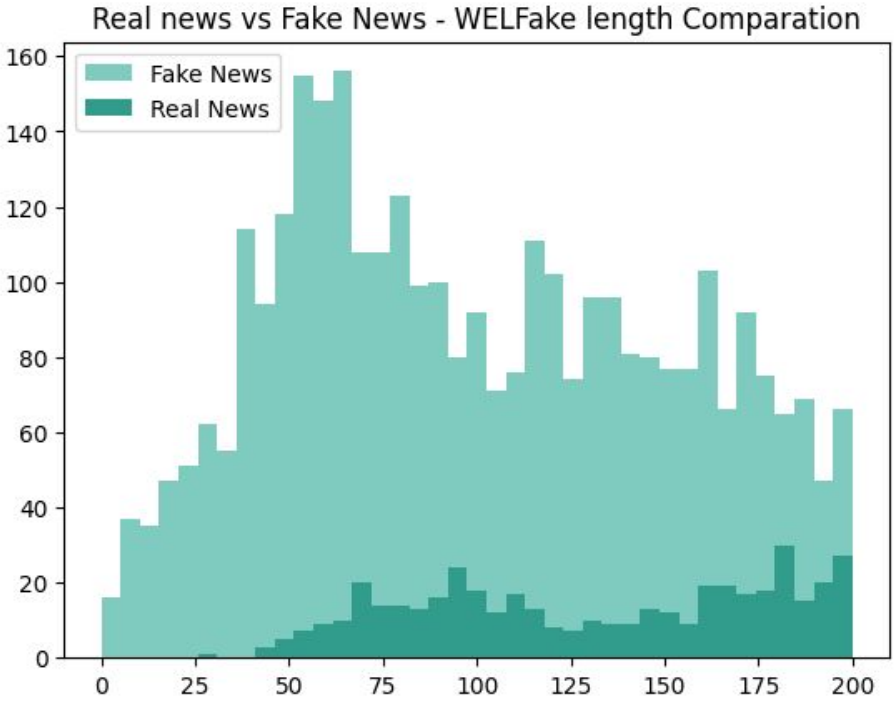
Compare different Classifier

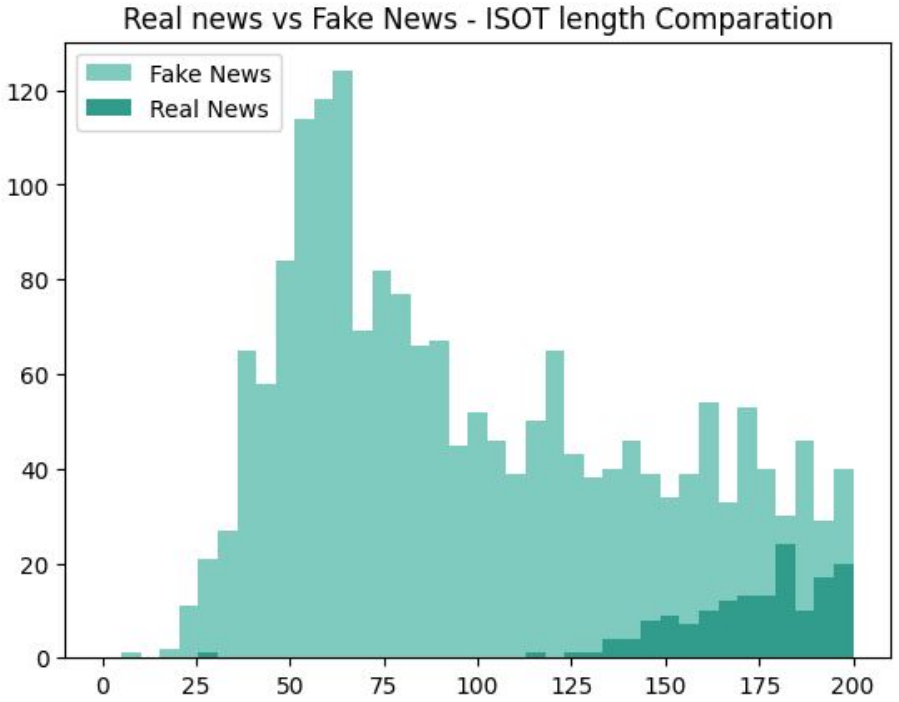
Which model achieves the best performance?"

Augmentation Impact

Augmentation Techniques Improve the performance?

Fake vs Real Length Comparation





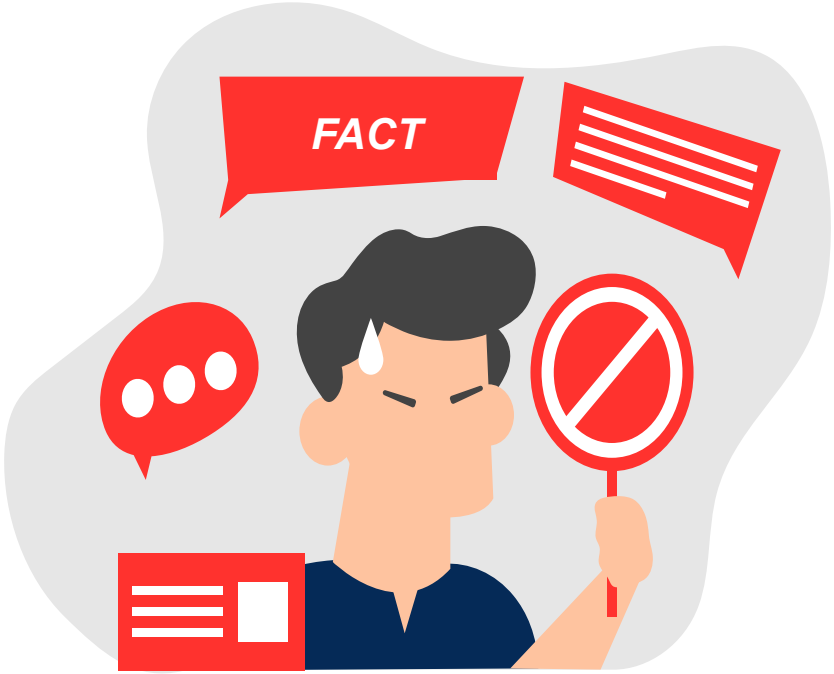
Model Comparison

Why CNN are better than LSTM in classification tasks?
AJIK ET AL., 2023 - Fake news detection using optimized cnn and lstm techniques

		Accuracy	Precision	Recall	F1-Score	Training Time	Number Parameters
WELFAKE	GLOVE + CNN	0.91	0.92	0.91	0.91	24.5s	1.036.489
	GLOVE +LSTM	0.75	0.81	0.75	0.75	18m 32s	1.017.257
	GLOVE +BI-LSTM	0.94	0.94	0.94	0.94	35m 44s	1.482.201
	BERT	0.97	0.97	0.90	0.97	36m 28s	28.764.162
ISOT	GLOVE +CNN	1.00	1.00	1.00	1.00	13.7s	1.036.489
	GLOVE + LSTM	0.95	0.95	0.95	0.95	12m 27s	1.017.257
	GLOVE +BI-LSTM	1.00	1.00	1.00	1.00	23m 26s	1.482.201
	BERT	1.00	1.00	1.00	1.00	21m 50s	28.764.162

Knowledge Distillation - WELFAKE & ISOT

		Accuracy	Precision	Recall	F1-Score	Training Time	Number Parameters
WELFAKE	BERT FINE TUNED	0.97	0.97	0.97	0.91	36m 28s	28.764.162
	STUDENT NETWORK - CNN	0.90	0.91	0.90	0.90	41.9s	7.361
	DISTILLED STUDENT NETWORK	0.90	0.91	0.90	0.90	18m 10s	7.361
ISOT	BERT FINE TUNED	1.00	1.00	1.00	1.00	21m 50s	28.764.162
	STUDENT NETWORK - CNN	1.00	1.00	1.00	1.00	21.4s	7.361
	DISTILLED STUDENT NETWORK	1.00	1.00	1.00	1.00	11m 24s	7.361



Thank you for your attention

Fact Finder
Information Retrieval
Academic Year 2023-2024

Nicolò Urbani 856213
Mattia Piazzalunga 851931