SAPIENZA
UNIVERSITÀ DI ROMA

SARCASM DETECTION ON REDDIT

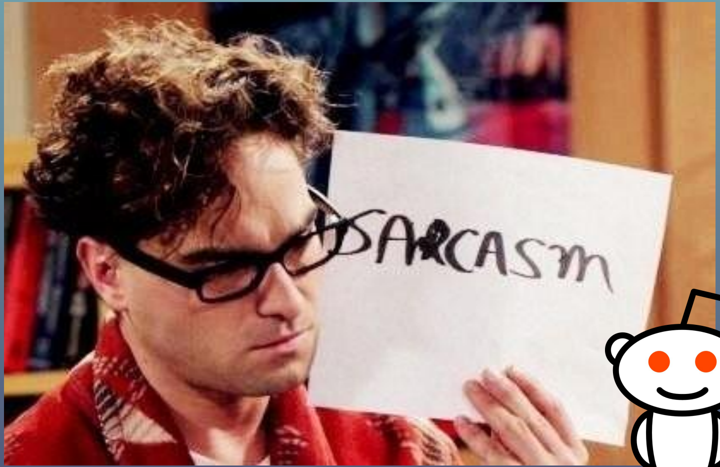
Big Data Computing - A.Y. 2022/2023
MSc in Computer Science
Faculty of Information Engineering, Informatics and Statistics
Sapienza University of Rome

Ballanti Chiara
ballanti.1844613@studenti.uniroma1.it

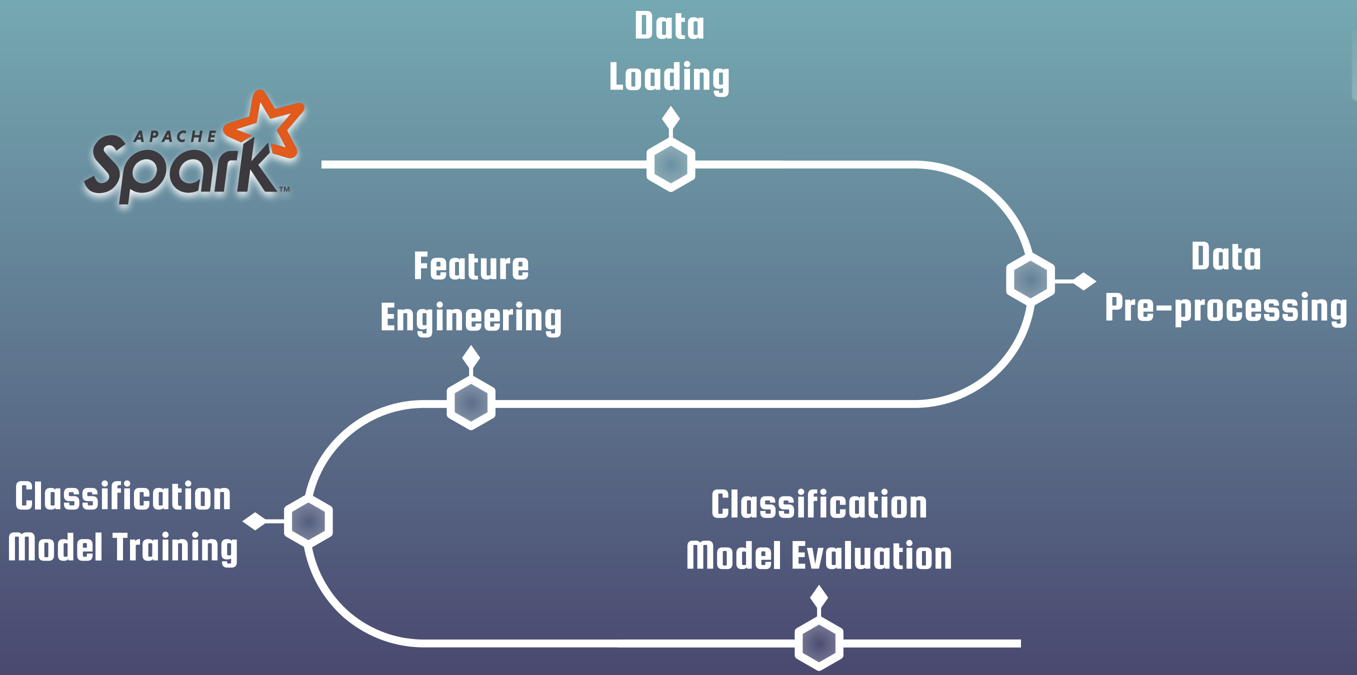
Bevilacqua Paolo Pio
bevilacqua.2002288@studenti.uniroma1.it

INTRODUCTION

The **goal** of this project is to address a binary classification problem performing sarcasm detection on Reddit comment in real-time.



PROJECT PIPELINE



Data Loading

Data Pre-processing

Feature Engineering

Classification Model Training

Classification Model Evaluation

01

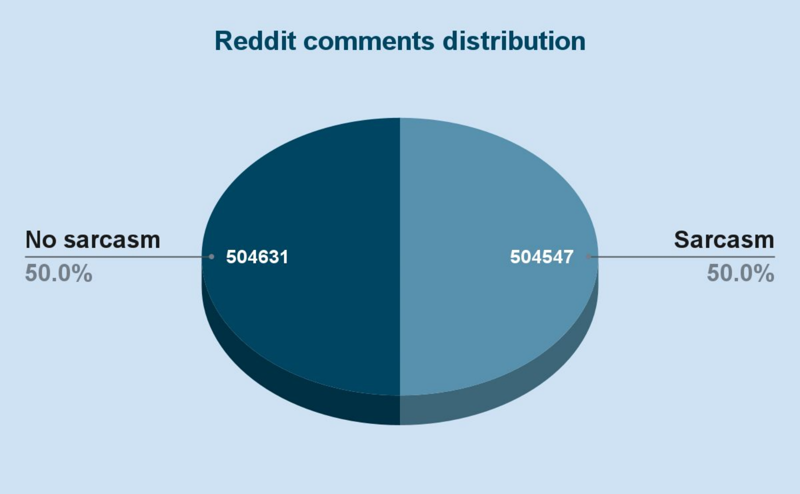
DATASET

Reddit comments tagged with sarcasm flag
downloaded from Kaggle

DATASET

1.01 M*

DATASET SIZE



DATASET COLUMNS

- **label**
- **comment**
- author
- subreddit
- score
- ups
- downs
- date
- created_utc
- parent_comment

(*) Due to the limited resources available to us and the high computational time required, we were forced to work with only part of the original dataset. Therefore, we randomly selected only **10%** of the comments in the dataset.

02

DATA PRE-PROCESSING
AND
FEATURE ENGINEERING

FEATURE ENGINEERING

TF-IDF + PCA

500 features

Word2Vec

300 features

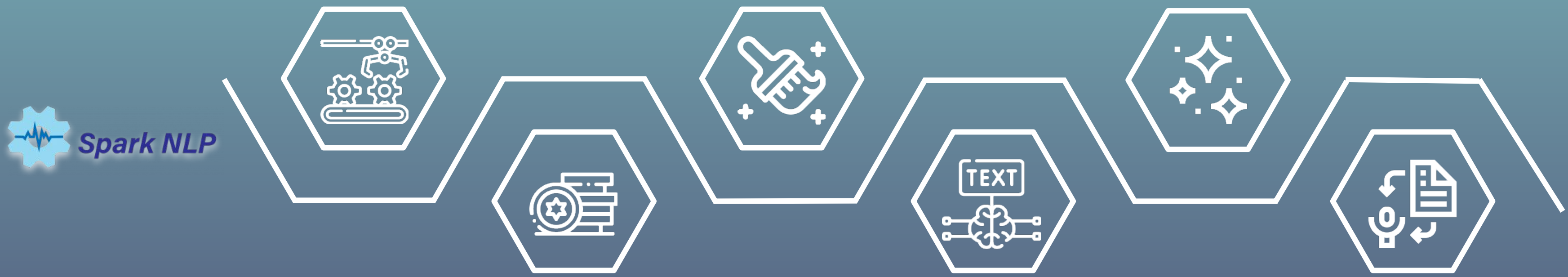
BERT

768 features

BERT Sentence

768 features

DATA PREPARATION



Spark NLP

DOCUMENT ASSEMBLER

TOKENIZER

NORMALIZER

LEMMATIZER

STOP WORDS CLEANER

FINISHER

We removed rows with same comments but different labels to avoid ambiguity.
Then we dropped duplicated comments.

TF-IDF + PCA

Term Frequency – Inverse Document Frequency (TF-IDF)

A statistical metric employed to assess the significance of a word within a document in a corpus.

10_000 extracted features

Principal Component Analysis (PCA)

A dimensionality reduction technique to project the vector of extracted features into a low-dimensional space.

500 extracted features

Word2Vec

Neural-network-based model to map each word to a vector of numbers.

Trained by us

- 300 extracted features for each token
- Average pooling

Pre-trained (word2vec_gigaword_wiki_300)

- 300 extracted features for each token
- Average pooling

Pre-trained using Gigaword 5th Edition and English Wikipedia Dump of February 2017

BERT

A deep learning model that employs a transformer-based architecture for generating contextualized word representations.

BERT Embedding

It provides **word-level** embedding using the BERT architecture.
It takes as input a sequence of tokens and generate contextualized embeddings for each token in the sequence.

BERT Sentence Embedding

It provides **sentence-level** embedding using the BERT architecture.
It takes as input a sequence of sentences and outputs a single embedding vector representing the entire sentence.

BERT Embedding

Smaller BERT Embeddings (L-2_H-768_A-12)

- 2 transformer layers
- 768 extracted features for each token
- 12 attention heads
- Case insensitive
- Average pooling

BERT base uncased Embedding (L-12_H-768_A-12)

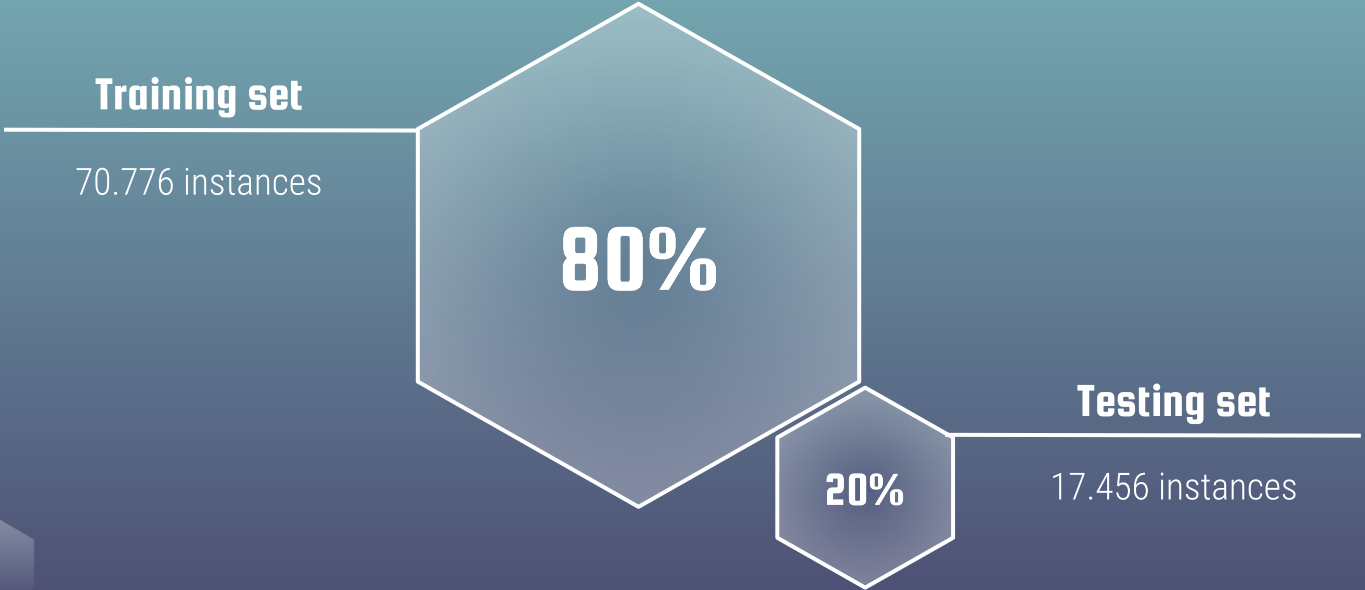
- 12 transformer layers
- 768 extracted features for each token
- 12 attention heads
- Case insensitive
- Average pooling

BERT Sentence Embedding

BERT Sentence base uncased Embedding (L-12_H-768_A-12)

- 12 transformer layers
- 768 extracted features for each token
- 12 attention heads
- Case insensitive

TRAINING SETUP



Training set

70.776 instances

80%

Testing set

17.456 instances

20%

03

CLASSIFICATION MODELS

CLASSIFICATION MODELS

Logistic regression

Random forest

Linear SVC

Multilayer perceptron

04

EVALUATION

EVALUATION RESULTS

	Logistic regression					
	TF-IDF	Word2Vec	Word2Vec pre-trained	Small BERT	BERT Base	BERT Sentence Base
F1-score	0.58	0.64	0.61	0.56	0.58	0.69
AUROC	0.64	0.67	0.64	0.59	0.61	0.76
AUPR	0.65	0.68	0.65	0.59	0.59	0.77
MCC	0.18	0.25	0.21	0.13	0.16	0.37
Accuracy	0.59	0.63	0.60	0.57	0.58	0.70

EVALUATION RESULTS

	Random forest					
	TF-IDF	Word2Vec	Word2Vec pre-trained	Small BERT	BERT Base	BERT Sentence Base
F1-score	0.56	0.63	0.58	0.54	0.59	0.65
AUROC	0.61	0.66	0.62	0.58	0.59	0.71
AUPR	0.63	0.67	0.63	0.58	0.58	0.72
MCC	0.16	0.23	0.16	0.11	0.14	0.30
Accuracy	0.58	0.62	0.59	0.56	0.57	0.66

EVALUATION RESULTS

	Linear SVC					
	TF-IDF	Word2Vec	Word2Vec pre-trained	Small BERT	BERT Base	BERT Sentence Base
F1-score	0.58	0.63	0.59	0.56	0.58	0.69
AUROC	0.63	0.66	0.64	0.59	0.62	0.76
AUPR	0.64	0.66	0.64	0.59	0.60	0.76
MCC	0.18	0.24	0.21	0.14	0.17	0.36
Accuracy	0.59	0.62	0.60	0.57	0.59	0.69

EVALUATION RESULTS

	Multilayer perceptron					
	TF-IDF	Word2Vec	Word2Vec pre-trained	Small BERT	BERT Base	BERT Sentence Base
F1-score	0.60	0.64	0.60	0.56	0.61	**0.72**
AUROC	0.63	0.66	0.64	0.59	0.66	**0.79**
AUPR	0.65	0.67	0.65	0.59	0.67	**0.81**
MCC	0.20	0.26	0.21	0.14	0.23	**0.42**
Accuracy	0.60	0.63	0.60	0.57	0.61	**0.72**

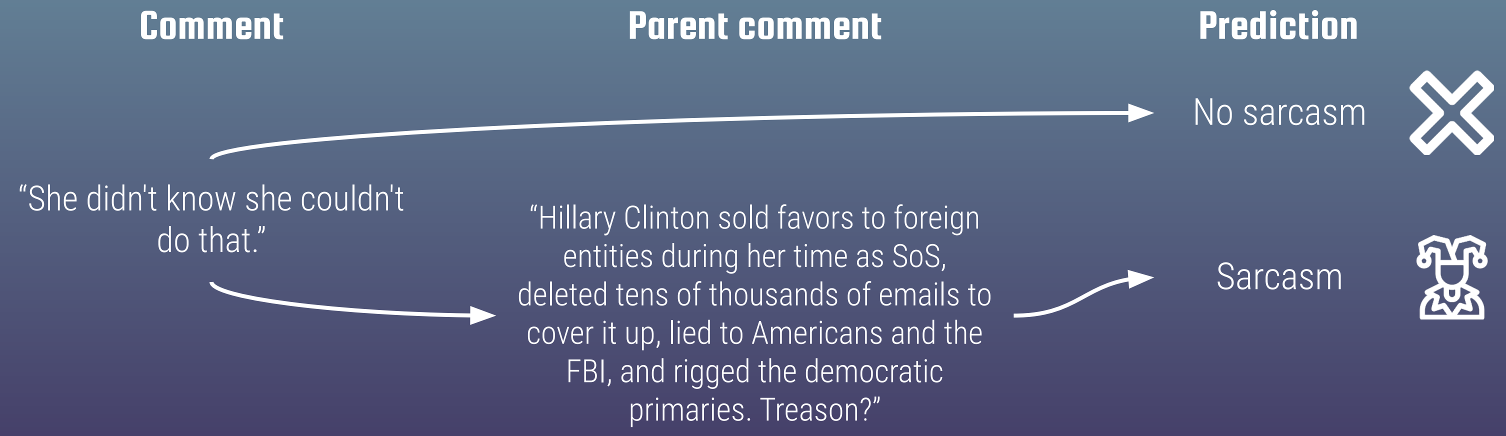
05

COMMENTS ANALYSIS AND SECOND DATASET

COMMENTS ANALYSIS

We found that the dataset is full of comments that can be used either in a sarcastic way or not.
The classification of these comments strongly depends on the context provided by the parent comment

Let's make an example:



Comment

Parent comment

Prediction

No sarcasm ✗

"She didn't know she couldn't do that."

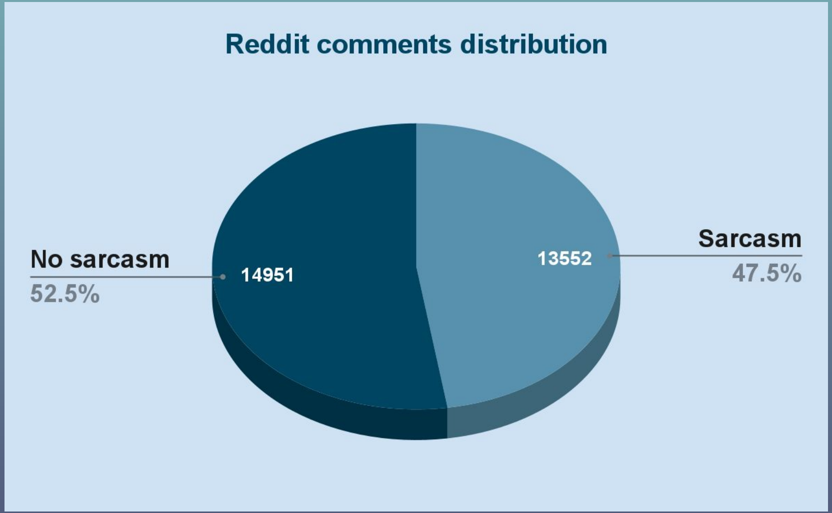
"Hillary Clinton sold favors to foreign entities during her time as SoS, deleted tens of thousands of emails to cover it up, lied to Americans and the FBI, and rigged the democratic primaries. Treason?"

Sarcasm

SECOND DATASET

29K

DATASET SIZE



DATASET COLUMNS

- label
- comment

The dataset was slightly unbalanced; we applied **downsampling** so that the two classes had the same number of samples (13.552 comments).

06

EVALUATION SECOND DATASET

EVALUATION RESULTS

	Logistic regression		
	Word2Vec	BERT Base	BERT Sentence Base
F1-score	0.70	0.74	0.87
AUROC	0.77	0.79	0.94
AUPR	0.77	0.81	0.94
MCC	0.40	0.43	0.74
Accuracy	0.70	0.72	0.87

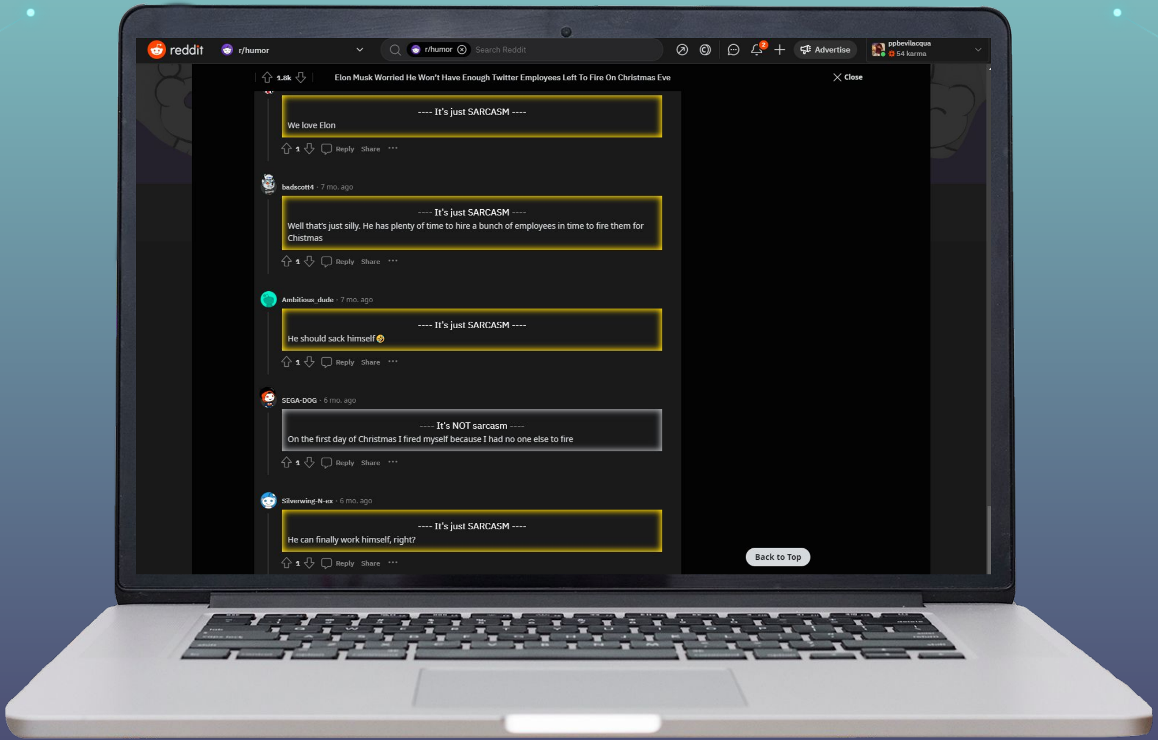
EVALUATION RESULTS

	Multilayer perceptron		
	Word2Vec	BERT Base	BERT Sentence Base
F1-score	0.69	0.75	**0.88**
AUROC	0.76	0.82	**0.95**
AUPR	0.75	0.84	**0.96**
MCC	0.37	0.48	**0.76**
Accuracy	0.69	0.74	**0.89**

07

DEMO

DEMO



Elon Musk Worried He Won't Have Enough Twitter Employees Left To Fire On Christmas Eve

---- It's just SARCASM ----
We love Elon

1 Reply Share ···

badscott4 · 7 mo. ago

---- It's just SARCASM ----
Well that's just silly. He has plenty of time to hire a bunch of employees in time to fire them for Chistmas

1 Reply Share ···

Ambitious_dude · 7 mo. ago

---- It's just SARCASM ----
He should sack himself 🤣

1 Reply Share ···

SEGA-DOG · 6 mo. ago

---- It's NOT sarcasm ----
On the first day of Christmas I fired myself because I had no one else to fire

1 Reply Share ···

Silverwing-N-ex · 6 mo. ago

---- It's just SARCASM ----
He can finally work himself, right?

1 Reply Share ···

Reddit home page. Each comment under a post will be analyzed and highlighted if it is recognized as *sarcastic*.

08

FUTURE WORK

FUTURE WORK

Experiment with more data in the initial dataset.

Experiment with different classification models and increase their complexity.

Expand the study by investigating sarcasm between a comment and its responses.

Experiment with different feature engineering techniques and extract a larger vector.

THANKS FOR THE ATTENTION!