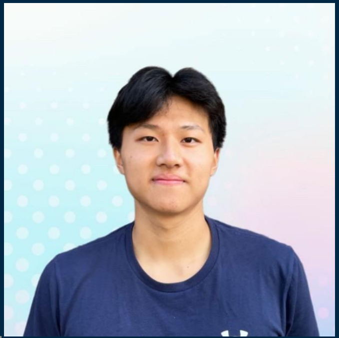
HUMOR & SARCASM DETECTION

Presented by
The Anti-Sarcasm Sarcastic Club

MEET THE CLUB

CHRISTOPHER “Boat Boy” MOPPEL
Grinnell College, ‘23



JOHN “LeetCode” BILLOS
Wake Forest, ‘24

ALBERT “Not Geoffrey” JING
Carleton College, ‘25

GEOFFREY “Not Albert” JING
Carleton College, ‘25

YUSUF “Edgy Jokes Trained the Model” ISMAIL
Carleton College, ‘24

PRESENTATION OUTLINE

01

DESCRIPTION & MOTIVATION

02

PREVIOUS RELATED WORKS

03

DATA CLEANING & PREPROCESSING

PRESENTATION OUTLINE (cont.)

04

MODEL ANALYSIS

05

LIVE DEMO

06

NEXT STEPS

DESCRIPTION & MOTIVATION

01

HUMOR & SARCASM DETECTION?


SARCASM
DETECTOR

OUR PROCESS

30K News Headlines labeled as Sarcastic/Humorous or not

DATA

MODELING

Build suite of different models: Neural Networks, kNN, Naive Bayes, XGBoost, Random Forest, etc.

Evaluate performance of all models on cleaned dataset and determine highest performing model

EVALUATION

TUNING

Fine-tune hyperparameters of the winning model

OUR MOTIVATION

CHALLENGE

Humor & Sarcasm Detection is still an active field of research

OPEN-ENDED

Always room for improvement (ex. text-scraping)

TWIN ID FAILED

Can we design a facial-rec model to distinguish between Geoffrey & Albert?

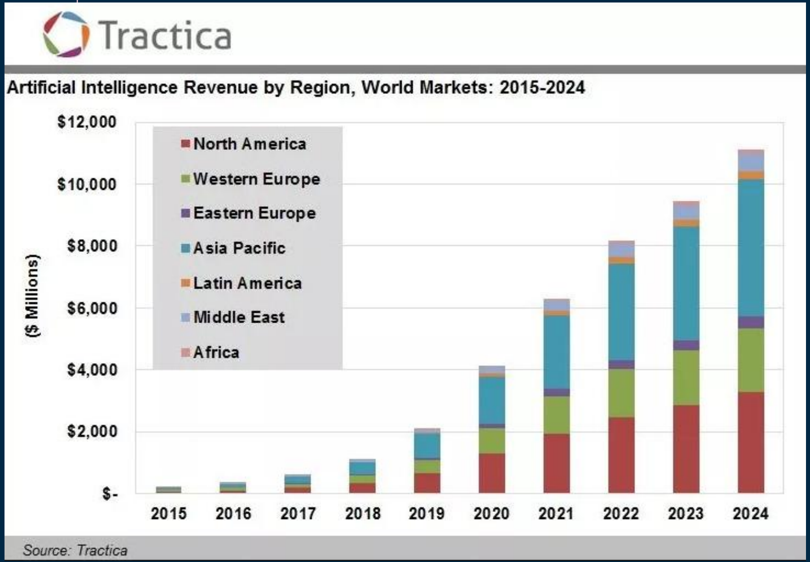
LEARNING

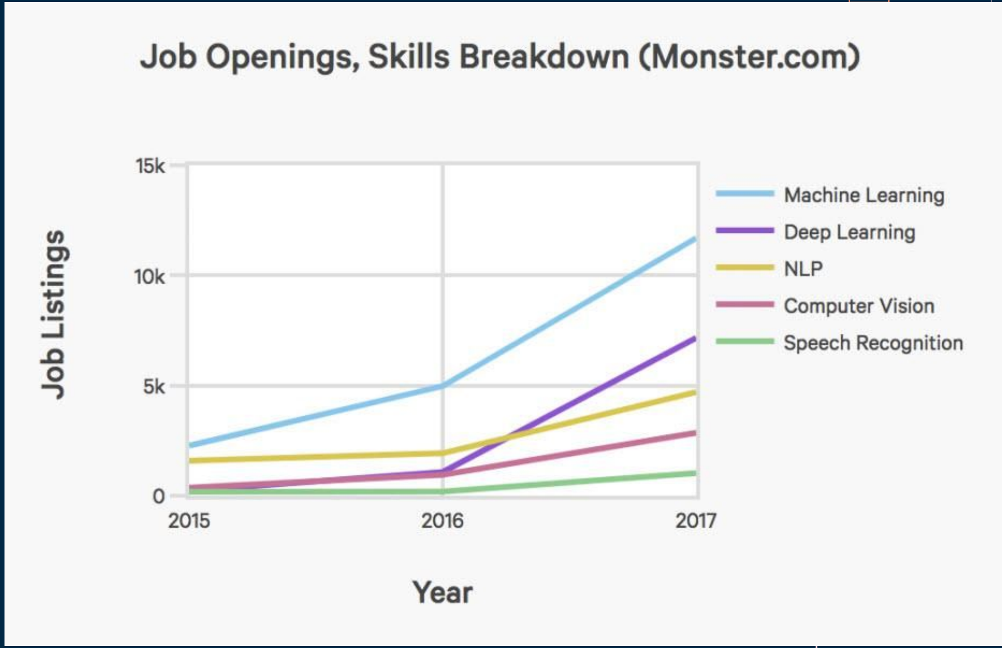
New modelling and data-transformation techniques

PREVIOUS RELATED WORKS

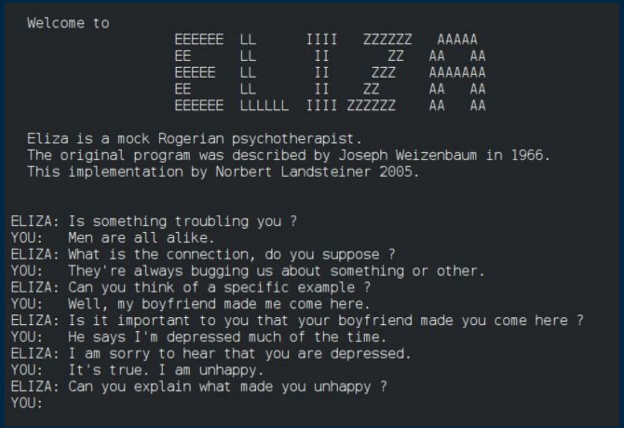
02

AI Market Trends



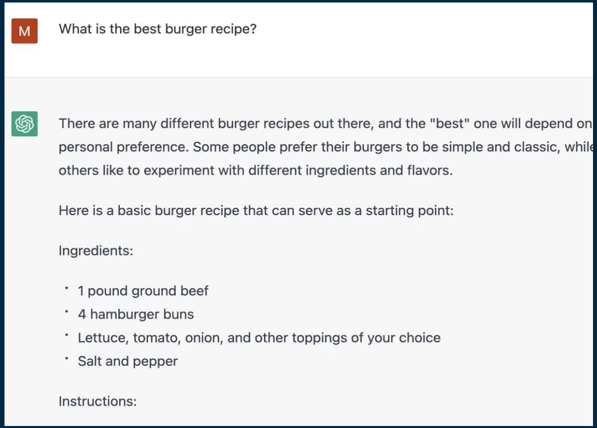


Communication with Machines



50s-70s

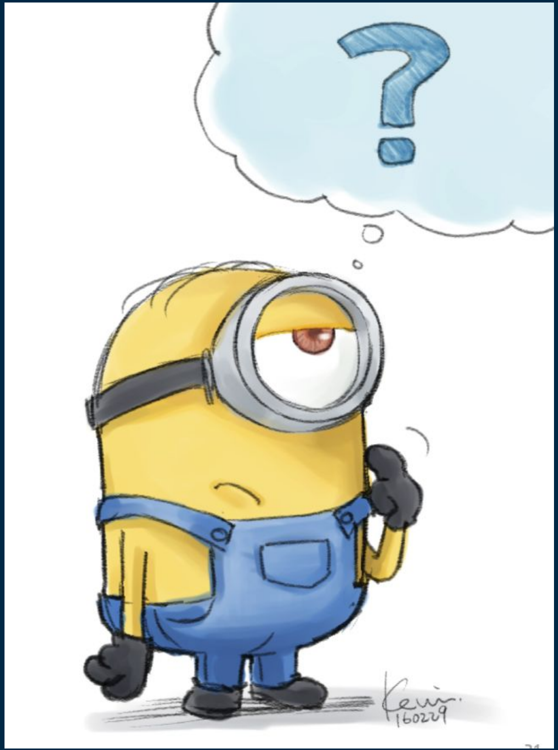
90s



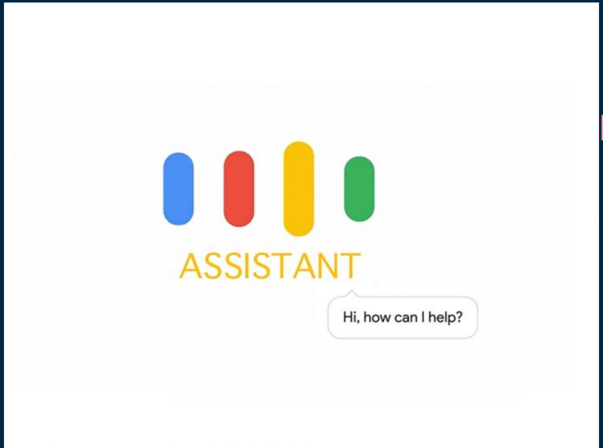
Today

WHY IS NLP HARD?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressity
6. Unknown representations



NLP TODAY


ASSISTANT
Hi, how can I help?


Daily News
Headlines
8 years from 2008-2016
reddit
Dow Jones
Industrial Average(DJIA)
Same day close value


OpenAI

Sarcasm Detection: A Comparative Study

July 2021

Authors:

Hamed Yaghoobian

Hamid Arabnia
University of Georgia

Khaled Rasheed
University of Georgia

Machine Learning-Based Model for Sentiment and Sarcasm Detection

Hamada Nayel, Eslam Amer, Aya Allam, Hanya Abdallah

ChatGPT: Optimizing Language Models for Dialogue

Sarcasm Analysis and Mood Retention Using NLP Techniques

IGI Global

January 2022 · International Journal of Information Retrieval Research 12(1):23

DOI:10.4018/IJIRR.289952

Authors:

Srijita Majumdar
University of Hyderabad

Debabrata Datta
St. Xavier's College, Kolkata

Arpan Deyasi
RCC Institute of Information Technology

Soumen Mukherjee
RCC Institute of Information Technology

On the naturalness of fuzzer-generated code

Authors: Rajeswari Hita Kambhamettu, John Billos, Tomi Oluwaseun-Apo, Benjamin Gafford, Rohan Padhye, Vincent J. Hellendoorn Authors Info & Claims

On the naturalness of software

Authors: Abram Hindle, Earl T. Barr, Mark Gabel, Zhendong Su, Premkumar Devanbu

Claims

TECHNOLOGIES

NAME	DESCRIPTION	EXAMPLES
NATURAL LANGUAGE PROCESSING	Subfield of linguistics, Computer Science and AI concerned with the interactions between computers and human language.	Speech Recognition, etc.
NEURAL NETWORKS	Artificial network of artificial neurons (perceptrons) used for solving AI problems.	Convolutional (CNN), Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), etc.
DATA ENCODING	Process of transforming data into a digestible format for regression (NN) algorithms.	N-grams, TF-IDF, etc.
ADVANCED ML ALGORITHMS	Advanced regression techniques tailored towards NLP problems.	XGBoost, BERT, Word2Vec, etc.

DATA CLEANING & PREPROCESSING

03

WHAT IS HUMOR ANYWAY?

HUMOR EXPLAINED

Humor: the capacity to express or perceive what's funny, is both a source of entertainment and a means of coping with difficult or awkward situations and stressful events.

Sarcasm: a type of phenomenon with specific perlocutionary effects on the hearer, such as to break their pattern of expectation.

Correct understanding of humor/sarcasm requires an understanding of the utterance, the conversational context, and, frequently some real-world facts.

DATA ENCODING TECHNIQUES

Bag of Words

Count Vectorization

N-Grams

TF-IDF

Word2Vec

BERT

DATASET

News Headlines Dataset For Sarcasm Detection

High quality dataset for the task of Sarcasm and Fake News Detection

30,000

Observations

(Individual Headlines)

3

Datapoints/Observation

1. Headline (string)
2. Link to the Article (hyperlink)
3. Humorous/Sarcastic? (binary)

2

News Sources

1. The Onion
2. HuffPost

SAMPLE OBSERVATION (RAW)

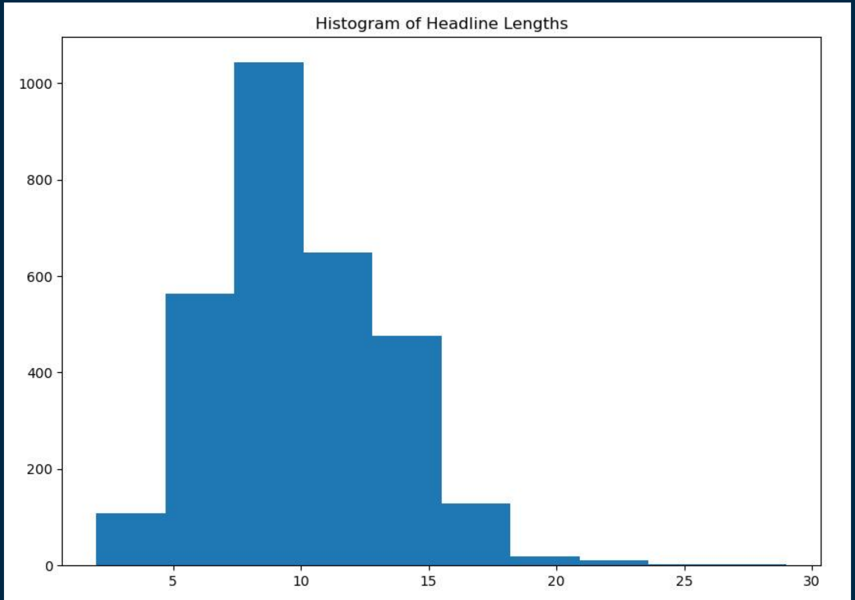
Sarcastic/Humorous:

```
is_sarcastic                                                     1
headline        mother comes pretty close to using word 'strea...
article_link    https://www.theonion.com/mother-comes-pretty-c...
Name: 4, dtype: object
```

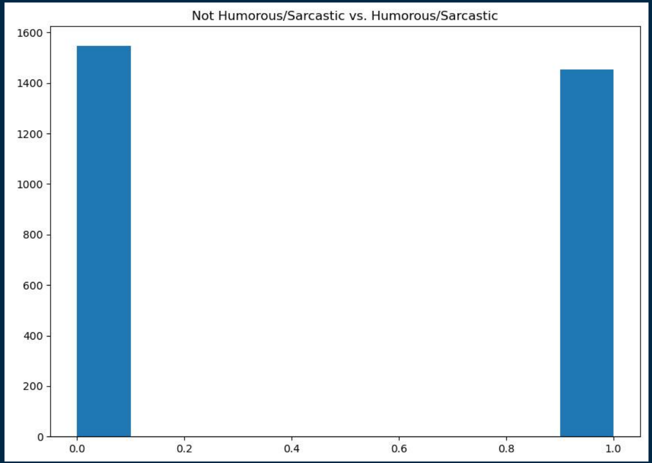
Not Sarcastic/Humorous:

```
is_sarcastic                                                     0
headline        eat your veggies: 9 deliciously different recipes
article_link    https://www.huffingtonpost.com/entry/eat-your-...
Name: 2, dtype: object
```

DATASET SUMMARY STATISTICS



Headline Lengths



Sarcastic/Humorous vs. Not Sarcastic/Humorous

DATASET WORD CLOUDS

Sarcastic/Humorous Headlines

Not Sarcastic/Humorous Headlines

MODEL ANALYSIS

04

MODELS

K Nearest Neighbors (kNN)

Naive Bayes

XG Boost

Bidirectional Encoder Representations from Transformers (BERT)

Convolutional Neural Network (CNN)

Decision Trees

Random Forests

Long Short Term Memory Network (LSTM)

Gated Recurrent Unit Network (GRU)

Logistic Regression

MODEL PERFORMANCES

	Sarcastic/Humorous Headlines			NOT Sarcastic/Humorous Headlines		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Naive Bayes	*0.70*	*0.18*	*0.29*	*0.56*	*0.93*	*0.70*
kNN	*0.53*	*0.42*	*0.47*	*0.56*	*0.67*	*0.61*
XGBoost	*0.63*	*0.42*	*0.50*	*0.59*	*0.78*	*0.67*
Logistic Reg.	*0.79*	*0.41*	*0.54*	*0.63*	*0.90*	*0.74*
Random Forest	*0.62*	*0.55*	*0.59*	*0.63*	*0.70*	*0.67*
CNN	*0.78*	*0.84*	*0.81*	*0.84*	*0.78*	*0.81*
GRU	*0.84*	*0.81*	*0.83*	*0.83*	*0.86*	*0.84*
LSTM	*0.82*	*0.84*	*0.83*	*0.85*	*0.82*	*0.84*

MODEL PERFORMANCES

	Sarcastic/Humorous Headlines			NOT Sarcastic/Humorous Headlines		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Naive Bayes	0.70	0.18	0.29	0.56	0.93	0.70
kNN	0.53	0.42	0.47	0.56	0.67	0.61
XGBoost	0.63	0.42	0.50	0.59	0.78	0.67
Logistic Reg.	0.79	0.41	0.54	0.63	0.90	0.74
Random Forest	0.62	0.55	0.59	0.63	0.70	0.67
CNN	0.78	0.84	0.81	0.84	0.78	0.81
GRU	0.84	0.81	0.83	0.83	0.86	0.84
LSTM	0.82	0.84	0.83	0.85	0.82	0.84

MODEL PERFORMANCES

	Sarcastic/Humorous Headlines			NOT Sarcastic/Humorous Headlines		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Naive Bayes	*0.70*	*0.18*	*0.29*	*0.56*	*0.93*	*0.70*
kNN	*0.53*	*0.42*	*0.47*	*0.56*	*0.67*	*0.61*
XGBoost	*0.63*	*0.42*	*0.50*	*0.59*	*0.78*	*0.67*
Logistic Reg.	*0.79*	*0.41*	*0.54*	*0.63*	*0.90*	*0.74*
Random Forest	*0.62*	*0.55*	*0.59*	*0.63*	*0.70*	*0.67*
CNN	*0.78*	*0.84*	*0.81*	*0.84*	*0.78*	*0.81*
GRU	*0.84*	*0.81*	*0.83*	*0.83*	*0.86*	*0.84*
LSTM	*0.82*	*0.84*	*0.83*	*0.85*	*0.82*	*0.84*

MODEL PERFORMANCES

	Sarcastic/Humorous Headlines			NOT Sarcastic/Humorous Headlines		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Naive Bayes	0.70	0.18	0.29	0.56	0.93	0.70
kNN	0.53	0.42	0.47	0.56	0.67	0.61
XGBoost	0.63	0.42	0.50	0.59	0.78	0.67
Logistic Reg.	0.79	0.41	0.54	0.63	0.90	0.74
Random Forest	0.62	0.55	0.59	0.63	0.70	0.67
CNN	0.78	0.84	0.81	0.84	0.78	0.81
GRU	0.84	0.81	0.83	0.83	0.86	0.84
LSTM	0.82	0.84	0.83	0.85	0.82	0.84

MODEL PERFORMANCES

	Sarcastic/Humorous Headlines			NOT Sarcastic/Humorous Headlines		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Naive Bayes	0.70	0.18	0.29	0.56	0.93	0.70
kNN	0.53	0.42	0.47	0.56	0.67	0.61
XGBoost	0.63	0.42	0.50	0.59	0.78	0.67
Logistic Reg.	0.79	0.41	0.54	0.63	0.90	0.74
Random Forest	0.62	0.55	0.59	0.63	0.70	0.67
CNN	0.78	0.84	0.81	0.84	0.78	0.81
GRU	0.84	0.81	0.83	0.83	0.86	0.84
LSTM	0.82	0.84	0.83	0.85	0.82	0.84

MODEL PERFORMANCES

	Sarcastic/Humorous Headlines			NOT Sarcastic/Humorous Headlines		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Naive Bayes	0.70	0.18	0.29	0.56	0.93	0.70
kNN	0.53	0.42	0.47	0.56	0.67	0.61
XGBoost	0.63	0.42	0.50	0.59	0.78	0.67
Logistic Reg.	0.79	0.41	0.54	0.63	0.90	0.74
Random Forest	0.62	0.55	0.59	0.63	0.70	0.67
CNN	0.78	0.84	0.81	0.84	0.78	0.81
GRU	0.84	0.81	0.83	0.83	0.86	0.84
LSTM	0.82	0.84	0.83	0.85	0.82	0.84

MODEL PERFORMANCES

	Sarcastic/Humorous Headlines			NOT Sarcastic/Humorous Headlines		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Naive Bayes	0.70	0.18	0.29	0.56	0.93	0.70
kNN	0.53	0.42	0.47	0.56	0.67	0.61
XGBoost	0.63	0.42	0.50	0.59	0.78	0.67
Logistic Reg.	0.79	0.41	0.54	0.63	0.90	0.74
Random Forest	0.62	0.55	0.59	0.63	0.70	0.67
CNN	0.78	0.84	0.81	0.84	0.78	0.81
GRU	0.84	0.81	0.83	0.83	0.86	0.84
LSTM	0.82	0.84	0.83	0.85	0.82	0.84

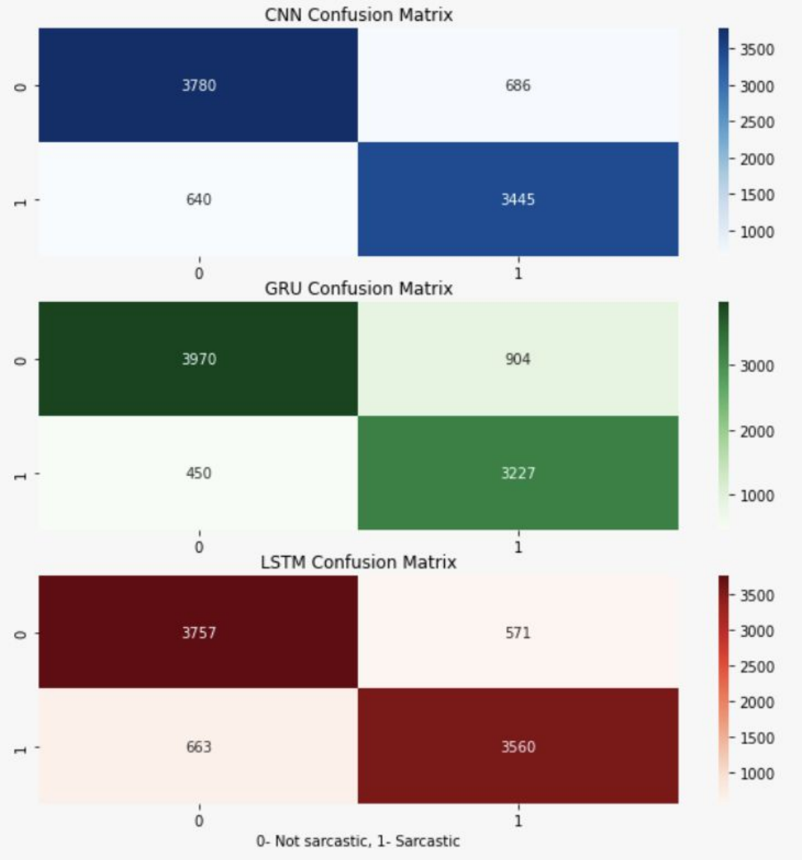
MODEL PERFORMANCES

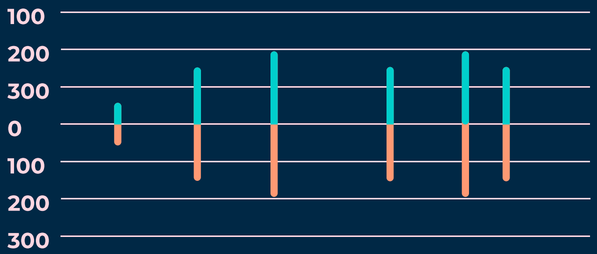
	Sarcastic/Humorous Headlines			NOT Sarcastic/Humorous Headlines		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Naive Bayes	0.70	0.18	0.29	0.56	0.93	0.70
kNN	0.53	0.42	0.47	0.56	0.67	0.61
XGBoost	0.63	0.42	0.50	0.59	0.78	0.67
Logistic Reg.	0.79	0.41	0.54	0.63	0.90	0.74
Random Forest	0.62	0.55	0.59	0.63	0.70	0.67
CNN	0.78	0.84	0.81	0.84	0.78	0.81
GRU	0.84	0.81	0.83	0.83	0.86	0.84
LSTM	0.82	0.84	0.83	0.85	0.82	0.84

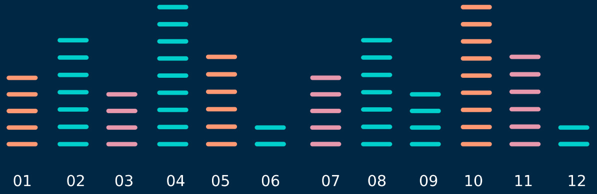
MODEL PERFORMANCES

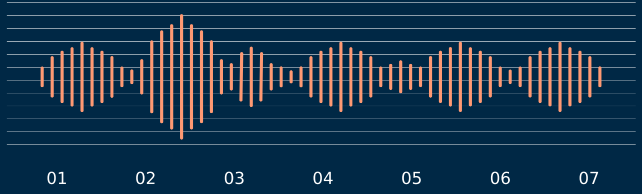
	Sarcastic/Humorous Headlines			NOT Sarcastic/Humorous Headlines		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Naive Bayes	0.70	0.18	0.29	0.56	0.93	0.70
kNN	0.53	0.42	0.47	0.56	0.67	0.61
XGBoost	0.63	0.42	0.50	0.59	0.78	0.67
Logistic Reg.	0.79	0.41	0.54	0.63	0.90	0.74
Random Forest	0.62	0.55	0.59	0.63	0.70	0.67
CNN	0.78	0.84	0.81	0.84	0.78	0.81
GRU	0.84	0.81	0.83	0.83	0.86	0.84
LSTM	0.82	0.84	0.83	0.85	0.82	0.84

NEURAL NETWORKS









BERT

BERT

Bidirectional Encoder Representations from Transformers

Example

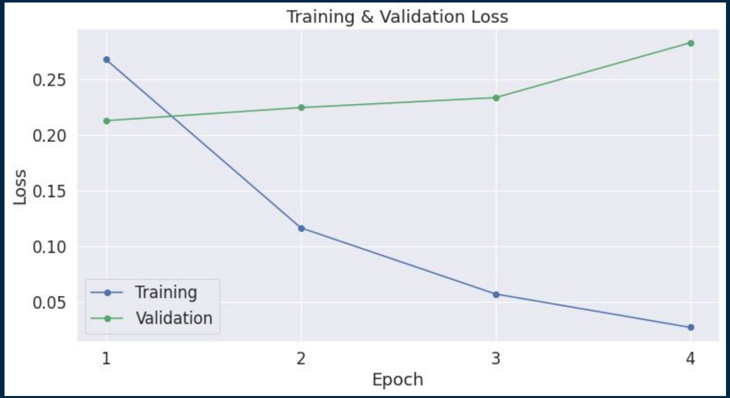
After stealing money from the bank vault, the bank robber was seen fishing on the Mississippi river bank.

```
First 5 vector values for each instance of "bank".

bank vault    tensor([ 3.3596, -2.9805, -1.5421,  0.7065,  2.0031])
bank robber   tensor([ 2.7359, -2.5577, -1.3094,  0.6797,  1.6633])
river bank    tensor([ 1.5266, -0.8895, -0.5152, -0.9298,  2.8334])
```

BERT

94% accuracy

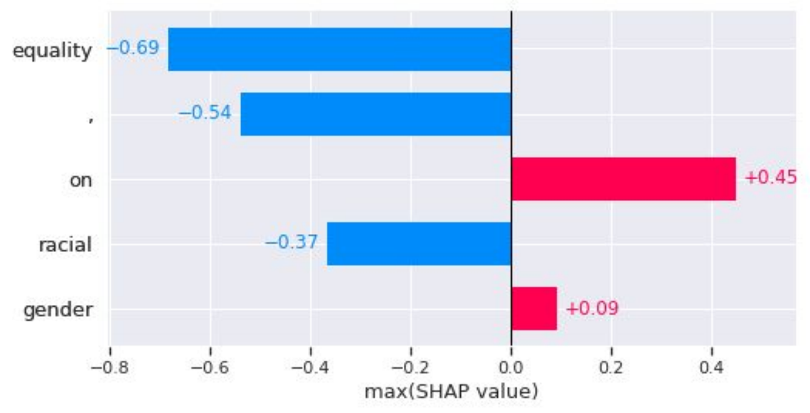


epoch					
1	0.27	0.21	0.92	0:04:33	0:00:10
2	0.12	0.22	0.93	0:04:33	0:00:10
3	0.06	0.23	0.94	0:04:33	0:00:11
4	0.03	0.28	0.94	0:04:33	0:00:11

SHAP explainer

Some words have greater effects than others

Open this link



BERT

Advantages:

- Very accurate
- Pre-trained
- Versatile, there is multiple versions with multiple datasets that they were trained on
- Easy to fine tune
- Developed by google and has lots of support
- Can be equipped with explainers very easily
- Contextual understanding

Disadvantages:

- Takes very long time to fine tune
- Pytorch is more complicated than tensorflow but explainers are easily adapted into it afterwards
- Much more code than neural networks

LIVE DEMO

05

NEXT STEPS

06

If We Had More Time…

1. Repeat previous analysis with TF-IDF Data Encoding
2. Refine Neural Network Models (CNN, GRU, LSTM)
3. Enhance Neural Network models through web-scraping article content using each observations' hyperlink
4. Investigate Word2Vec Encoding
5. Enhance the BERT encoder

THANK YOU

Questions?

WARNING: If you don’t want a sarcastic answer, don’t ask a stupid question…

OUR COMPANY

Mercury is the closest planet to the Sun and the smallest one in the Solar System—it's only a bit larger than the Moon. The planet's name has nothing to do with the liquid metal

UNDERSTANDING THE PROBLEM

MARS

Despite being red, Mars is a cold place. It's full of iron oxide dust, which gives the planet its reddish cast

VENUS

Venus has a beautiful name and is the second planet from the Sun. It's terribly hot, even hotter than Mercury

MAIN COMPETITORS

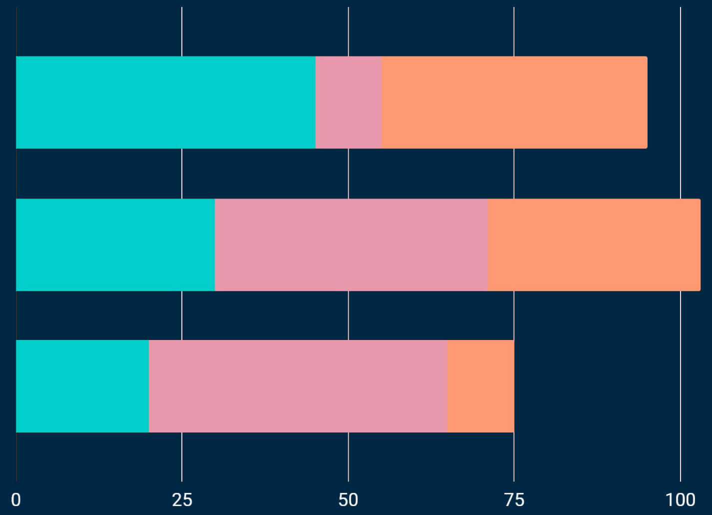
NEPTUNE
It’s the farthest planet from the Sun

MARS
Despite being red, Mars is a cold place

VENUS
Venus is the second planet from the Sun

SATURN
It’s composed mostly of hydrogen and helium

MARKET RESEARCH



NEPTUNE

It’s the farthest planet from the Sun

MERCURY

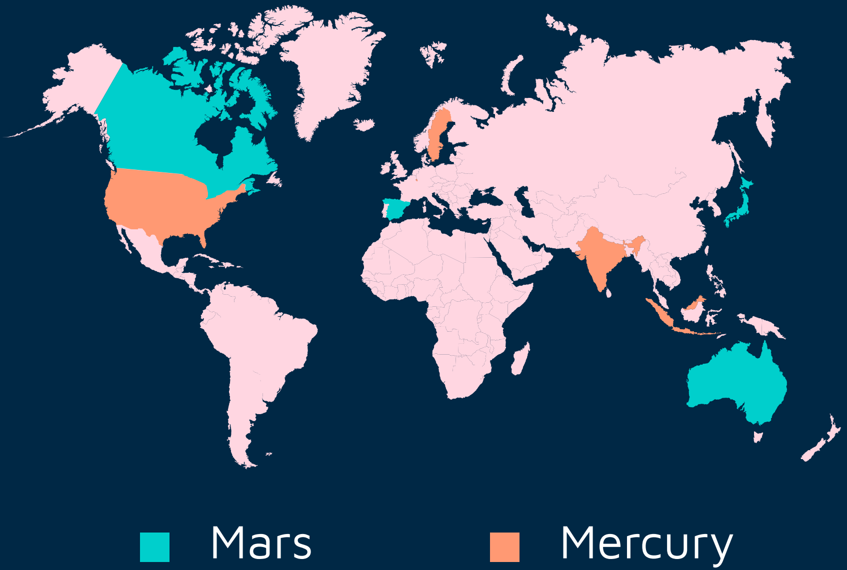
Mercury is the closest planet to the Sun

SATURN

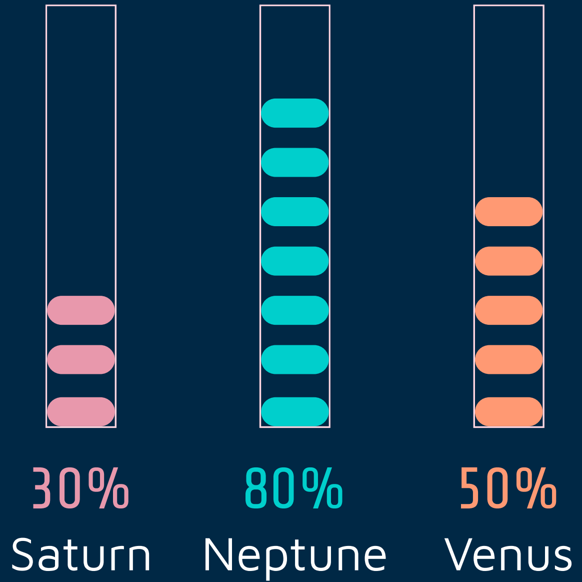
Saturn is composed of hydrogen and helium

ANALYSIS

OUTREACH



TOP RATED VALUES



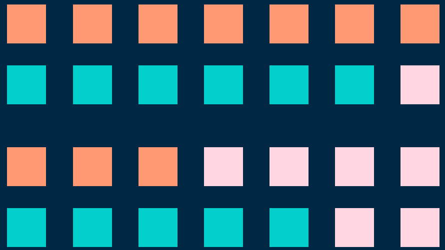
TARGET

GENDER

75% Female

60% Male

AGE

20-39

80%

40-60
50%

500,000+

Employees

A Picture Is Worth a Thousand Words

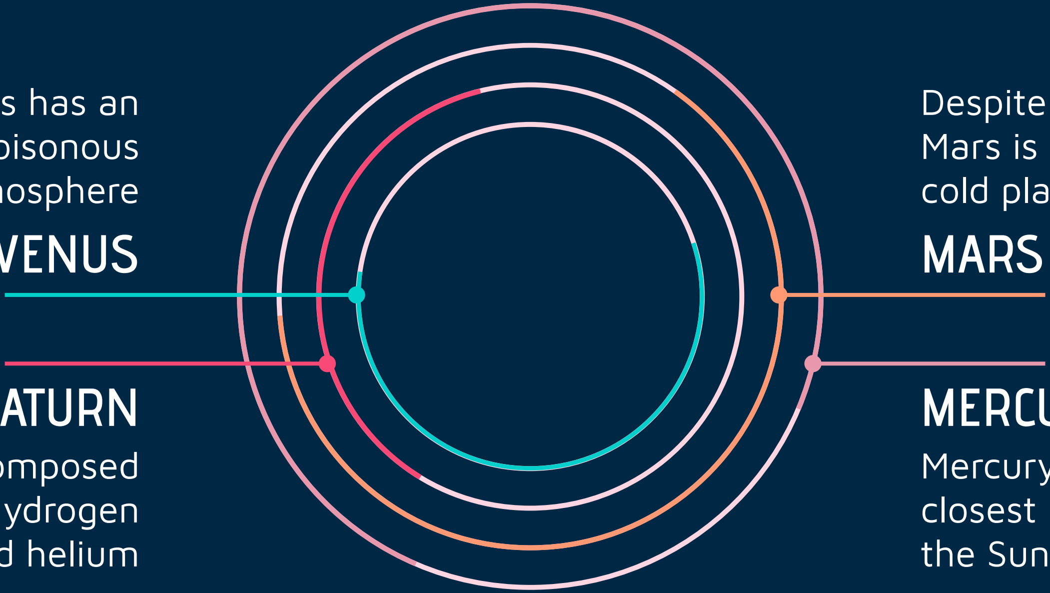
“This is a quote. Words full of wisdom that someone important said and can make the reader get inspired.”

—SOMEONE FAMOUS

OUR PARTNERS

Venus has an extremely poisonous atmosphere

VENUS



SATURN

Saturn is composed mostly of hydrogen and helium

Despite being red, Mars is actually a cold place

MARS

MERCURY

Mercury is the closest planet to the Sun

TESTIMONIALS

"Mercury is the closest planet to the Sun and the smallest of them all"

—RYAN DIXON

"Venus has a beautiful name and is the second planet from the Sun"

—ALIYA FARLEY

"Jupiter is a gas giant and the biggest planet in the Solar System"

—HENRY McKANE

"Saturn is composed mostly of hydrogen and helium"

—BILLY BROOKS

"The Sun is the star at the center of the Solar System"

—LUCY JADE

"Neptune is the fourth-largest planet in the Solar System"

—ROSE CLARK

4,498,300,000

Big numbers catch your audience’s attention

UPCOMING GOALS

JUPITER

JUNE 2

It’s the biggest planet in the Solar System

SATURN

OCTOBER 14

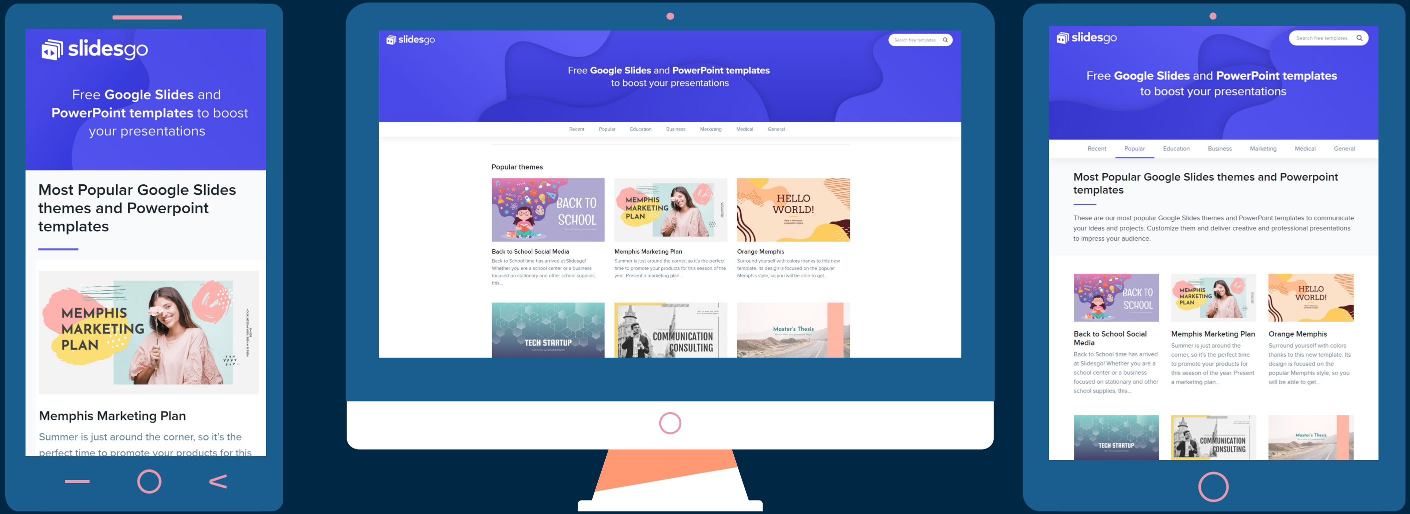
Saturn is composed mostly of hydrogen and helium

NEPTUNE

JANUARY 23

Neptune is the farthest planet from the Sun

SNEAK PEEK



You can replace the images on these screens with your own work

Do you have any questions?

youremail@freepik.com
+91 620 421 83
yourcompany.com

THANKS

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik
Please keep this slide for attribution

ALTERNATIVE RESOURCES