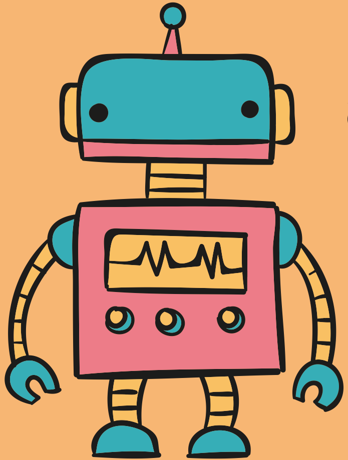
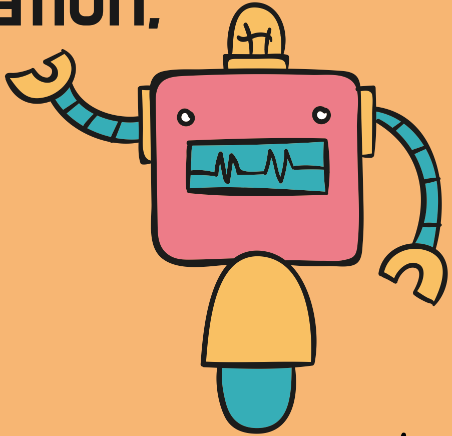
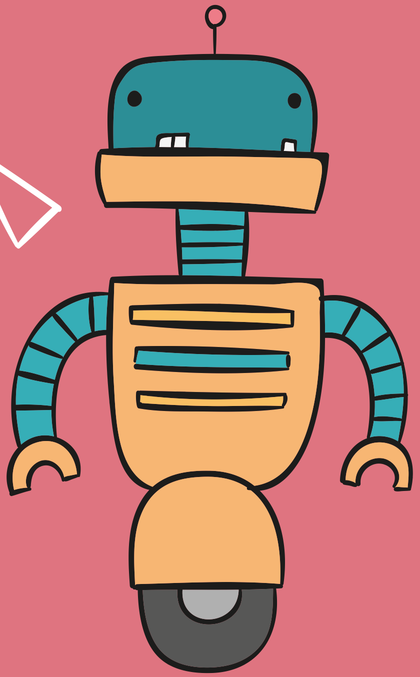
Adapting BERT for Diverse NLP Tasks: Author and Language Identification, News Categorization, and Sentiment Analysis



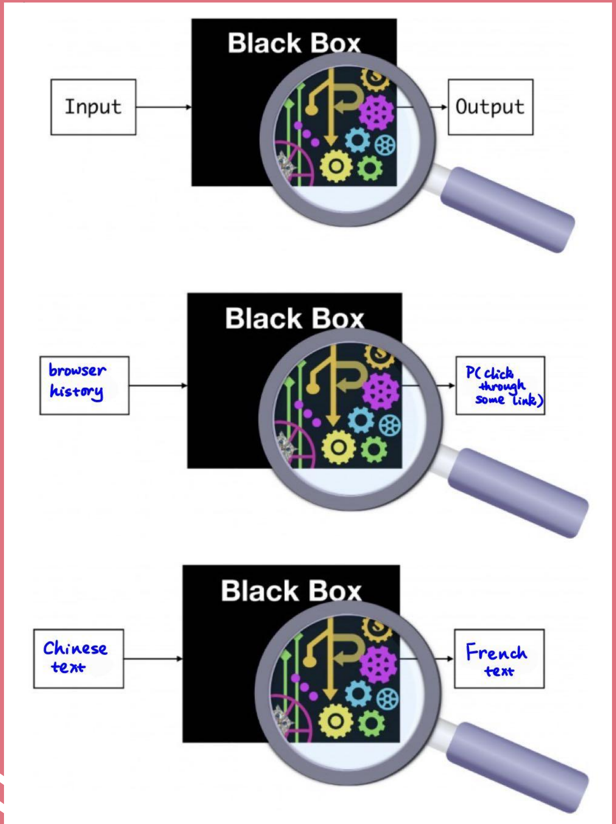


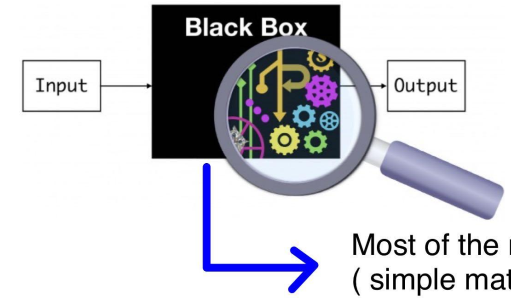
By Zixuan Cao, Kristina Kuznetsova, Sifan Zhu



01. Introduction

Meet ANNs (Artificial Neuron Networks)





Most of the math is basic arithmetic (simple math +, *, /, log in small pieces)

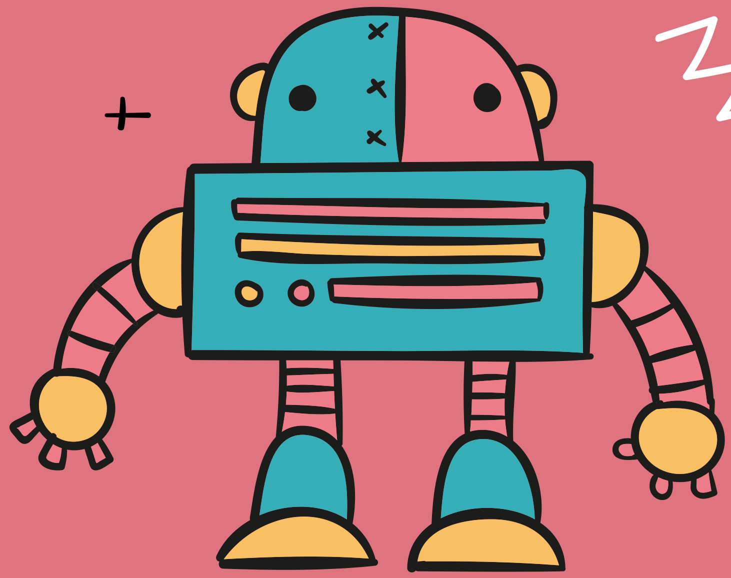
How about putting many, many, many small pieces together?

Picture by https://blog.ml.cmu.edu/2019/05/17/explaining-a-black-box-using-deep-variational-information-bottleneck-approach/

Why ANNs?

To distinguish between whether it's a cat or a dog in images, linear solution is not good enough. **(Image Classification)**

In fact, in the real world, most problems are complex and involve relationships or patterns that are inherently **non-linear** .



Deep Learning shines through!

- Amazing at solving **complicated problems** !
- No needs to know anything about the nature of the correct solution. It will **automatically** figure out for you.

Imagine: As a chef at a famous Michelin-star restaurant

To make the best sandwich with butter and jam in the world! 🥪

Solution by deep learning:
Keep making more sandwiches until the customer is happy. Use **negative feedback to adjust** your receipt; **stop adjusting** when you get **positive feedback** .

Picture by https://www.stockfood.com/images/11166997-Toast-with-butter-and-strawberry-jam

Bread

Jam

$$Y = a * (x0 + x1w1 + x2w2) + b$$

Sandwich

Activation Function (Add non-linearity)

Butter

Bias: default flavor & adjust the receipt

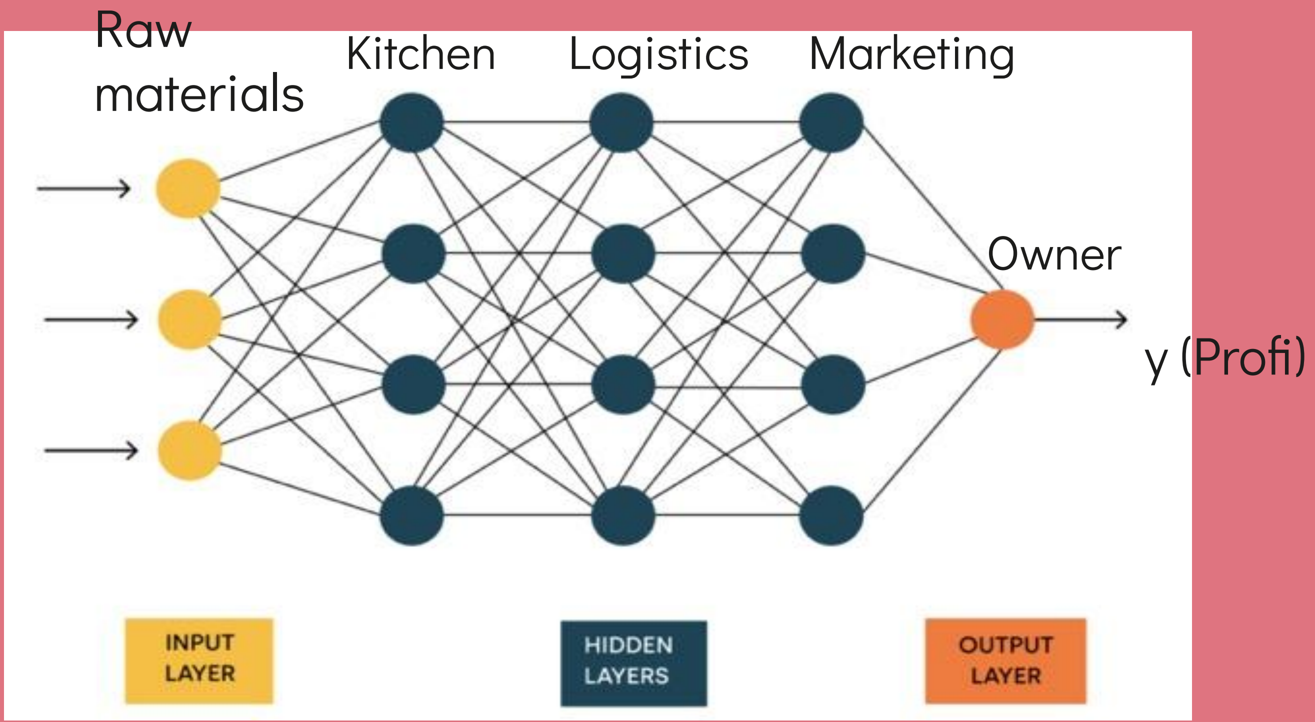
w1: represents the amount of the butter

w2: represents the amount of the jam

Forward propagation : Make the sandwich

Backwards propagation : Adjust the amounts based on negative feedback by customers

Run a business of making sandwich with butter and jam



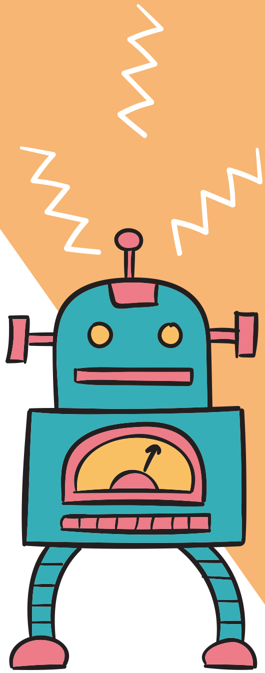
Key: We repeat the simple computational building block many, many, many times.

Picture by https://www.qtravel.ai/blog/what-are-neural-networks-and-what-are-their-applications/

In linguistics: Challenge of Context

"I can **bank** on you." → Trust

"I am going to the **bank**." → A physical location



01

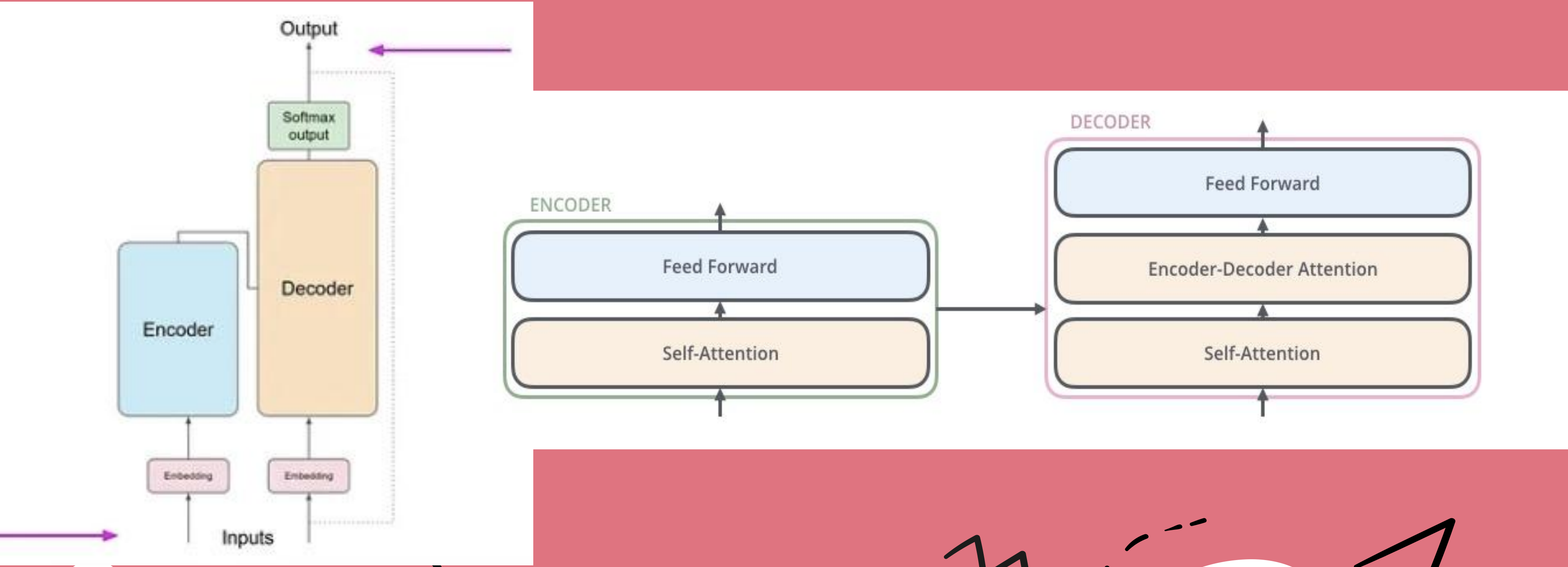
The role of "bank" in the sentence

02

The relationship of "bank" with other words nearby

Self-attention layers in Transformers excel!

Analyze the entire sentence focusing on relationships between words to understand context



Picture by https://python.plainenglish.io/a-basic-high-level-view-of-transformer-architecture-llms-from-35000-feet-c036dd2f7c25

Encoder? Decoder?

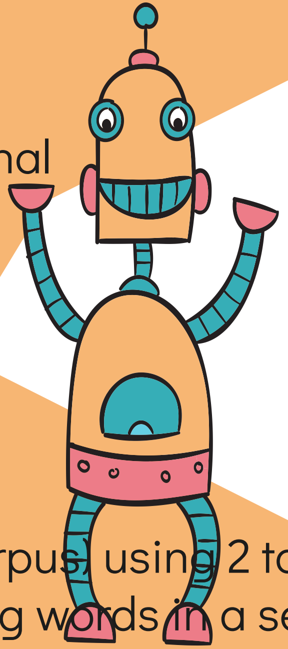
Used in tasks where the input and output are sequences of different lengths

- Encoder: Convert the input data into **a meaningful representation** so that can be easily understood and utilized by the decoder

- Decoder: **Generate the output sequence** from the representation provided by encoder

BERT (Bidirectional Encoder Representations From Transformers)

Focuses on the **encoder** . Learns deep contextual relationships within text. Maps every word in a sentence to a high-dimensional space, where similar meanings are close to each other.

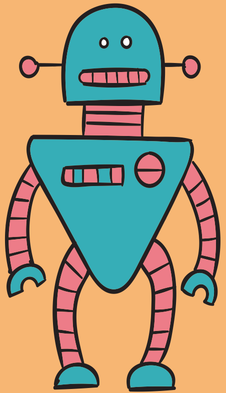


Bidirectional?
→ Processes the input from left-to-right & right-to-left.

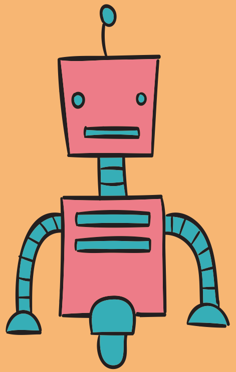
Pre-trained on massive text corpora (Wikipedia and BooksCorpus) using 2 tasks:

- Masked Language Modeling (MLM) → Predicting missing words in a sentence
- Next Sentence Prediction (NSP) → Understanding relationships between sentence pairs

Fine-tuned Bert models

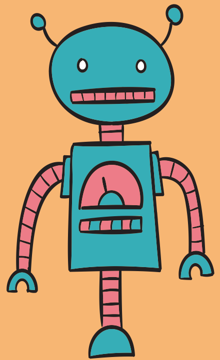


Sentiment Analysis



Letter Classification

(author + language)

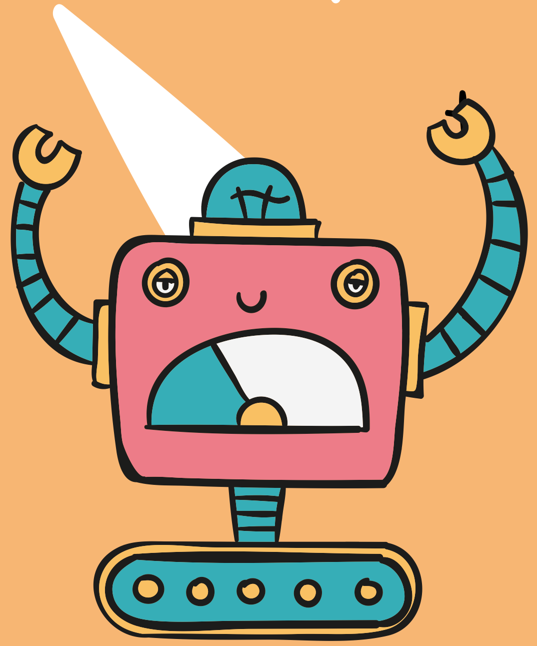


News Classification

02.

Sentiment analysis

fine-tuned on IMDB: Sentiment



Fine-Tuning

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.212500	0.309306	0.886000	0.888432	0.886000	0.885634
2	0.336400	0.346524	0.906000	0.906787	0.906000	0.905869
3	0.276700	0.343147	0.915000	0.914996	0.915000	0.914997

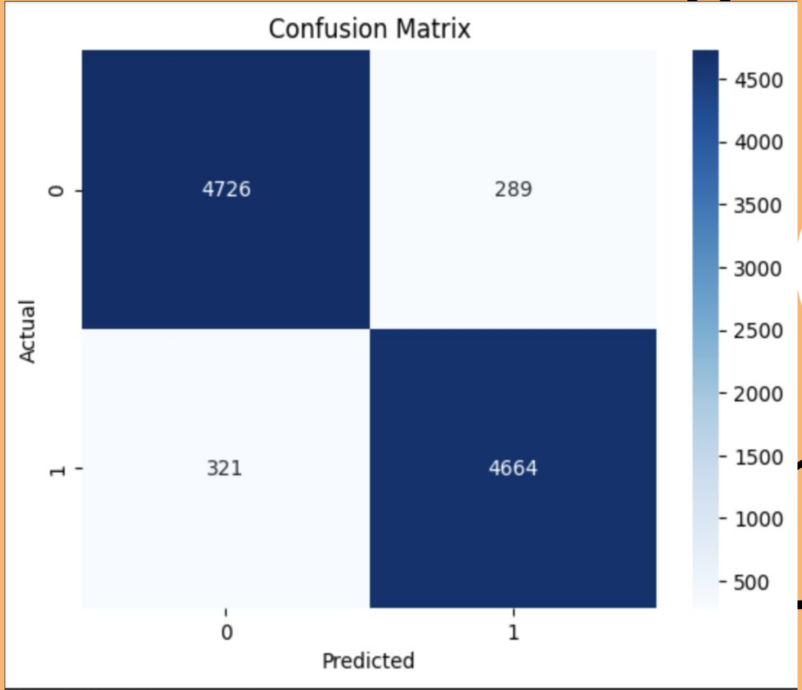
Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.209000	0.187946	0.928000	0.929497	0.928000	0.927944
2	0.136100	0.184067	0.936800	0.936815	0.936800	0.936799
3	0.047000	0.245845	0.939000	0.939017	0.939000	0.938999

Fine-Tuning

Error Analysis	Hyperparameter Tuning	Threshold Adjustment
• Ambiguous or mixed sentiment in the text. • Presence of sarcasm or idiomatic expressions that might confuse the model.	• Optimized the learning rate, batch size, and other hyperparameters to reduce misclassifications.	• Experimented with different thresholds

Confusion Matrix

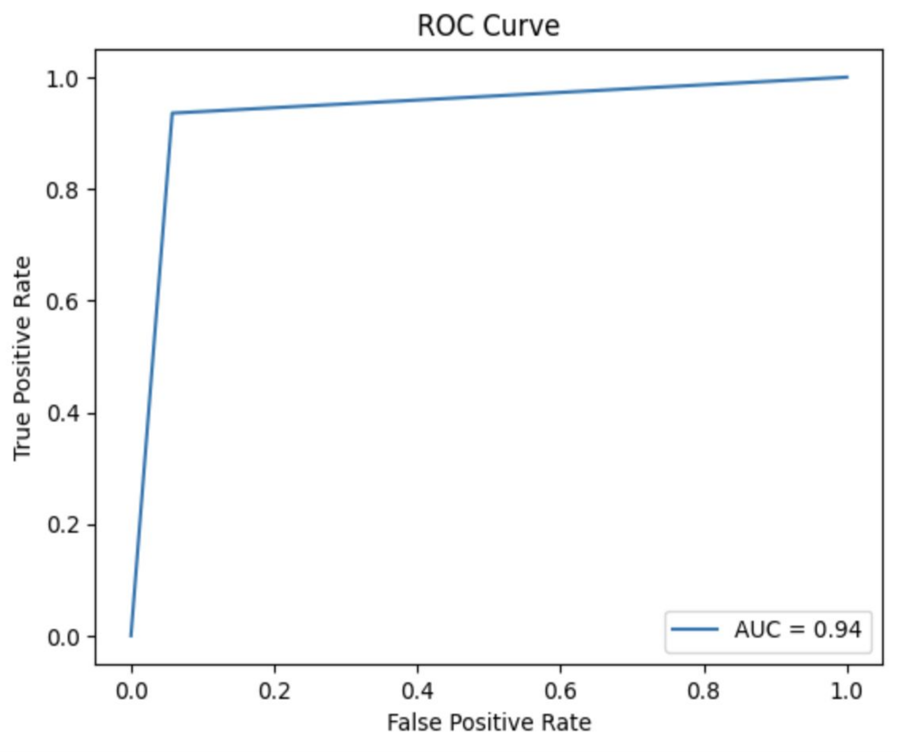
- **Balanced Performance:**
  Both positive and negative classes have identical precision, recall, and F1-scores.
- **Few Misclassifications:**
  The false positives (289) and false negatives (321) are relatively small compared to the correctly classified samples.
- **High Accuracy:**
  A 94% accuracy demonstrates that the fine-tuned BERT model is effective for the sentiment analysis task.



Classification Report:

	precision	recall	f1-score	support
positive	0.94	0.94	0.94	5015
negative	0.94	0.94	0.94	4985
accuracy			0.94	10000
macro avg	0.94	0.94	0.94	10000
weighted avg	0.94	0.94	0.94	10000

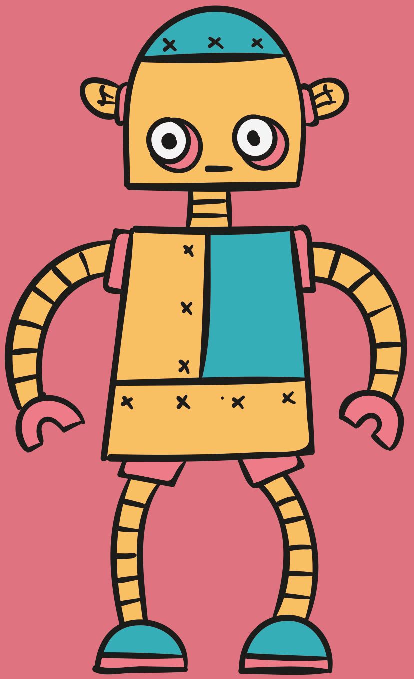
ROC Curve



A curve that approaches the top-left corner indicates better performance, as it signifies **high TPR** (correctly identifying positives) and **low FPR** (minimizing false alarms).

- **High Discrimination Power:**
  With an AUC of 0.94, our model performs exceptionally well at distinguishing between classes, even when accounting for different thresholds.
- **Balanced Trade-off:**
  The ROC curve suggests that the model maintains a good trade-off between recall and precision.

How well does this model perform outside its comfort zone?



Evaluation on Amazon Polarity Dataset

What It Tests:

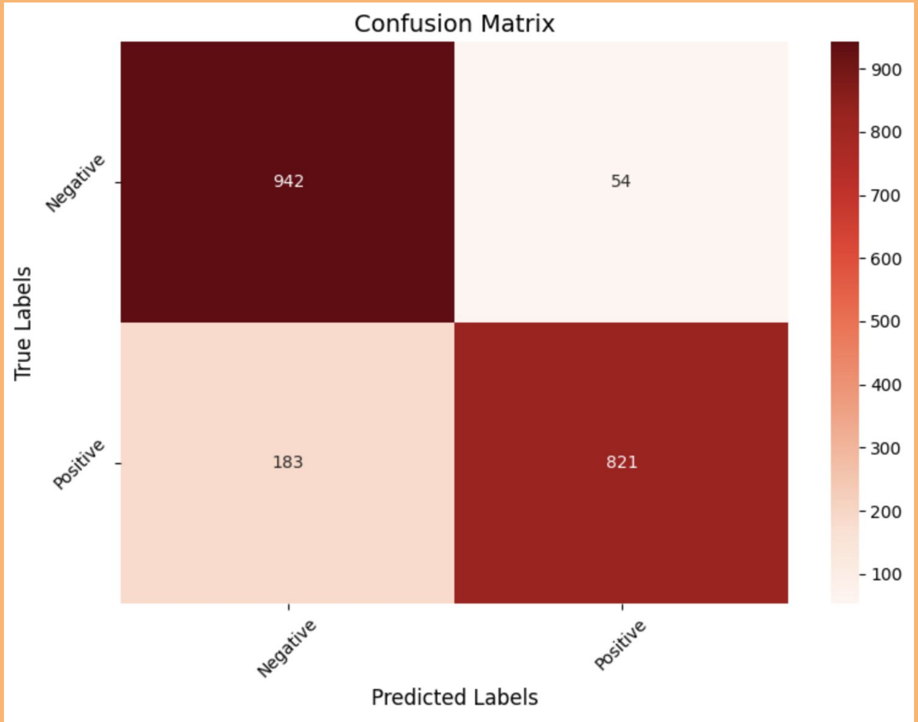
- ★ How well a model can handle new, unseen data from a **different domain** or dataset.
- ★ The ability to **transfer learned patterns** to other contexts.

Why It's Important:

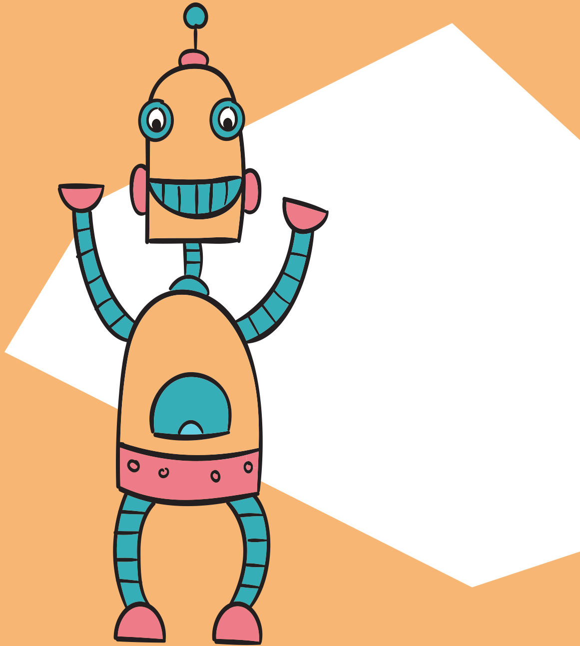
- ★ Ensures the model is **not overfitting** to the training dataset but instead learns generalizable

Goal:

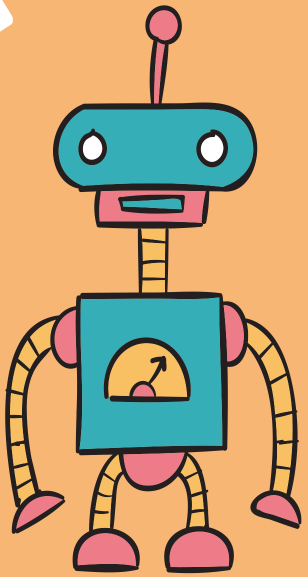
- ★ Evaluate if the model **maintains performance** (e.g., accuracy, precision, recall) when applied to different datasets.



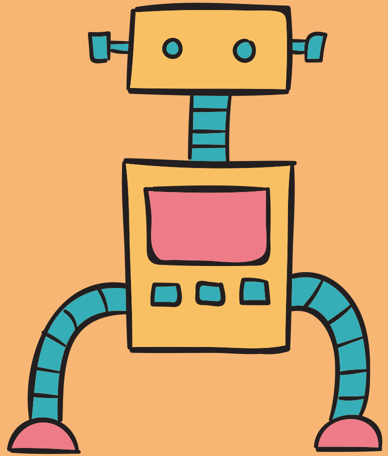
Amazing!



The model retains high performance on an OOD dataset, achieving **88% accuracy** , highlighting its robustness.



Ever wanted to know why a model made a certain prediction?



Human-readable explanations with LIME

Interpretability of Complex Models

LIME helps break down predictions of ML models into simpler, understand-able components.

Understanding Specific Predictions

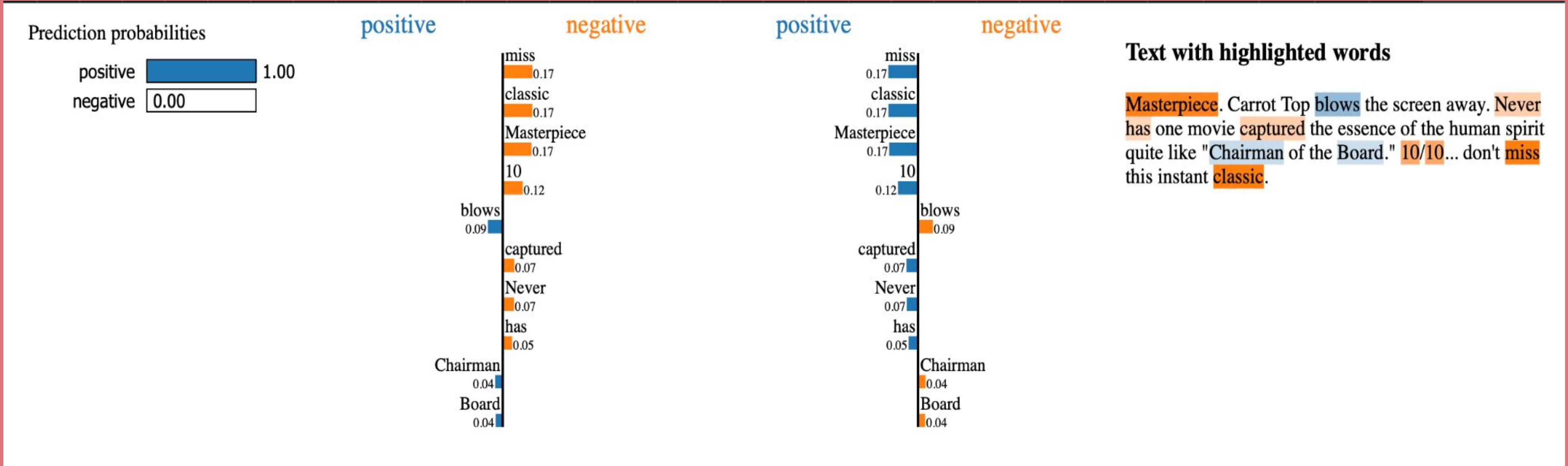
Unlike traditional methods that try to explain the global behavior of a model, LIME focuses on explaining **individual predictions**.

Model-Agnostic

LIME can be applied to **any** machine learning model, regardless of the underlying architecture.

Human-readable explanations with LIME

*LIME shows exactly which features influenced its decision.



Greater Interpretability + Context Ambiguity

Human-readable explanations with LIME

Contribution of Every Word :

The model treats every token in the input as relevant to the prediction.

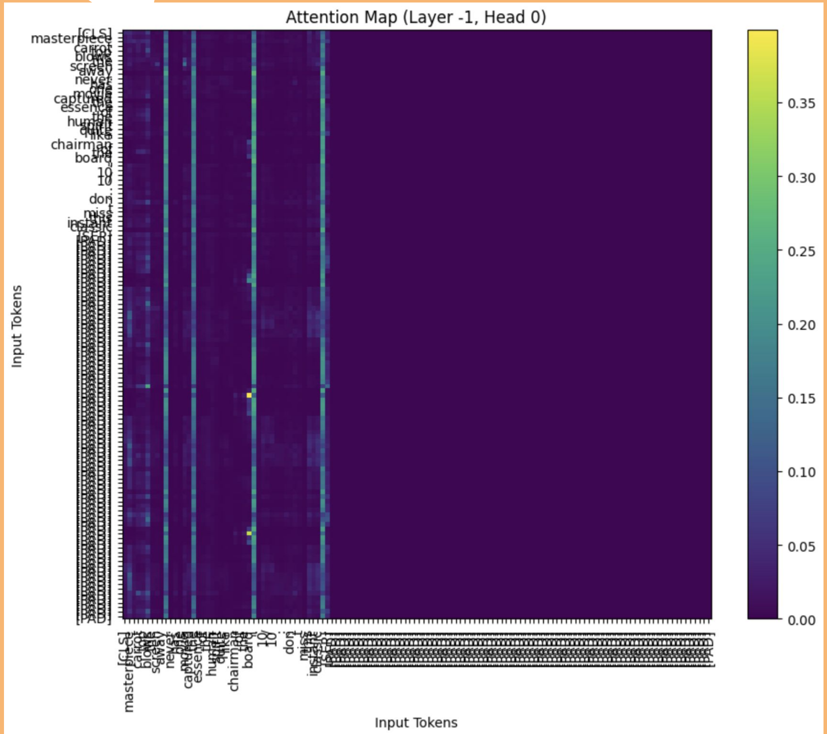
Balanced Contribution :

Some words have both **positive and negative contributions** This highlights how a word's meaning can change based on its context within the sentence.

Model's Context Sensitivity :

The model is considering their contextual interdependencies to derive sentiment.

Attention visualization

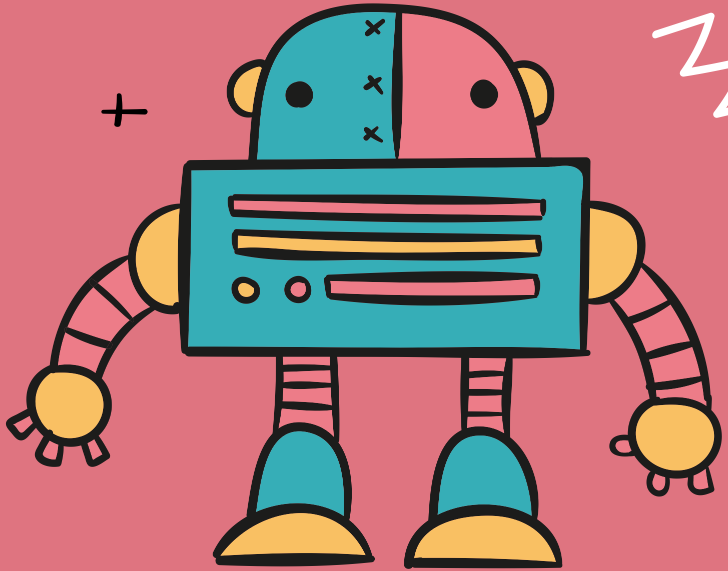


- **The attention mechanism** calculates the importance (or weight) of each token in relation to every other token in the sequence.
- This helps to:

  - Understand Token Contributions
  - By Explainability
  - Fine-Tuning Insight

03.

Author classification & Language classification



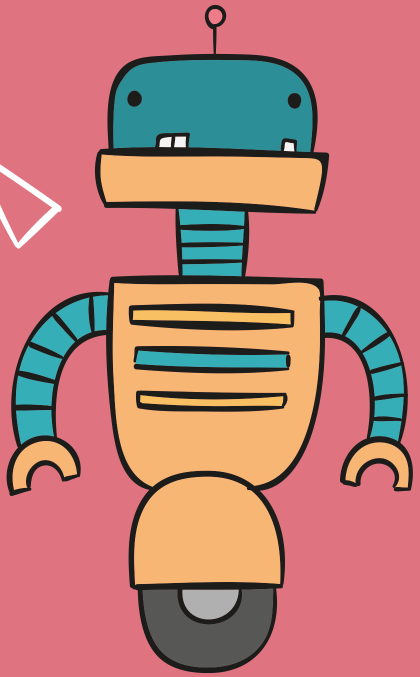
Process

Author classification:
Fine-tune the multilingual version of BERT (**mBERT**) on the dataset & Evaluate

Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	0.088600	0.153238	0.950420	0.887025	0.891602	0.887395
2	0.080800	0.124608	0.966400	0.927359	0.925595	0.929804
3	0.065900	0.134569	0.969678	0.934034	0.934995	0.935238

The model improves steadily across epochs.

Want to know more about the model? → **Probing**



Probing

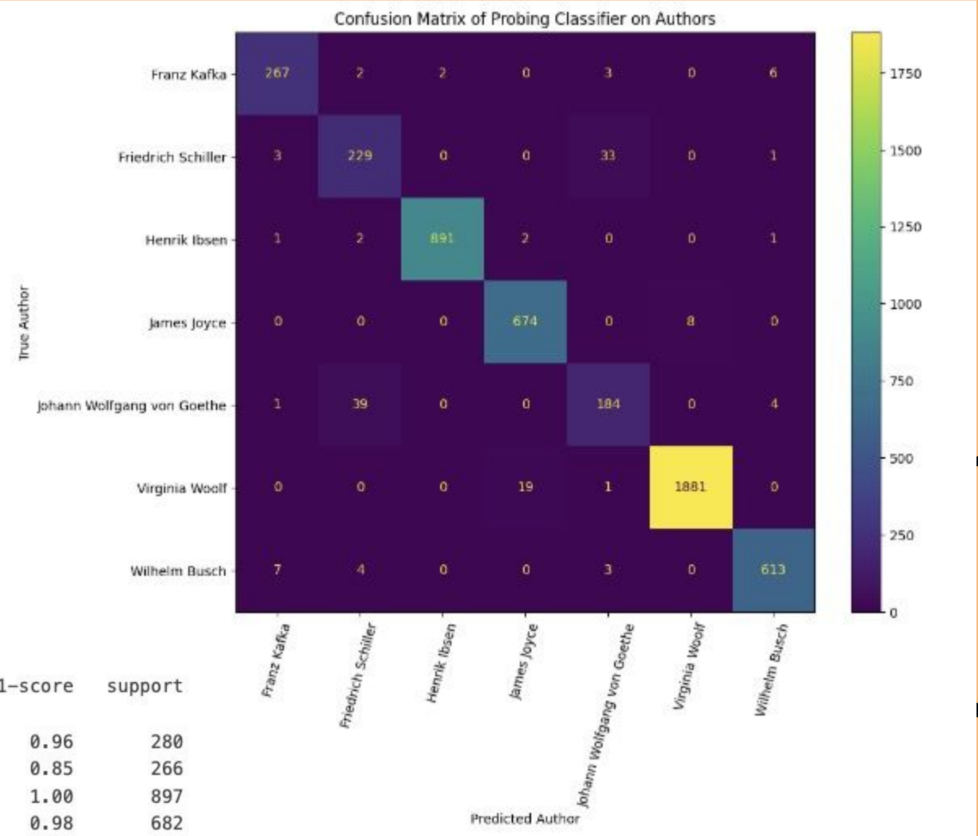
To analyze and interpret the representations in hidden states or embeddings encode about specific linguistic properties or features in the data.

In the author classification task: How well the representations from a specific BERT layer encode information on decision making between classes

3

Extract representations from a single layer → To analyze how much information is encoded in the last hidden layer (**classification task**)

Train a probing classifier to predict the corresponding labels



Probing Classifier Accuracy: 0.9709

	precision	recall	f1-score	support
Franz Kafka	0.96	0.95	0.96	280
Friedrich Schiller	0.83	0.86	0.85	266
Henrik Ibsen	1.00	0.99	1.00	897
James Joyce	0.97	0.99	0.98	682
Johann Wolfgang von Goethe	0.82	0.81	0.81	228
Virginia Woolf	1.00	0.99	0.99	1901
Wilhelm Busch	0.98	0.98	0.98	627
accuracy			0.97	4881
macro avg	0.94	0.94	0.94	4881
weighted avg	0.97	0.97	0.97	4881

LIME

To explain the predictions of our model

Attention Mechanism

A **heatmap visualization** of the attention weights in the last layer of a pre-trained BERT model. The heatmap shows how tokens in the input text attend to each other.

LIME

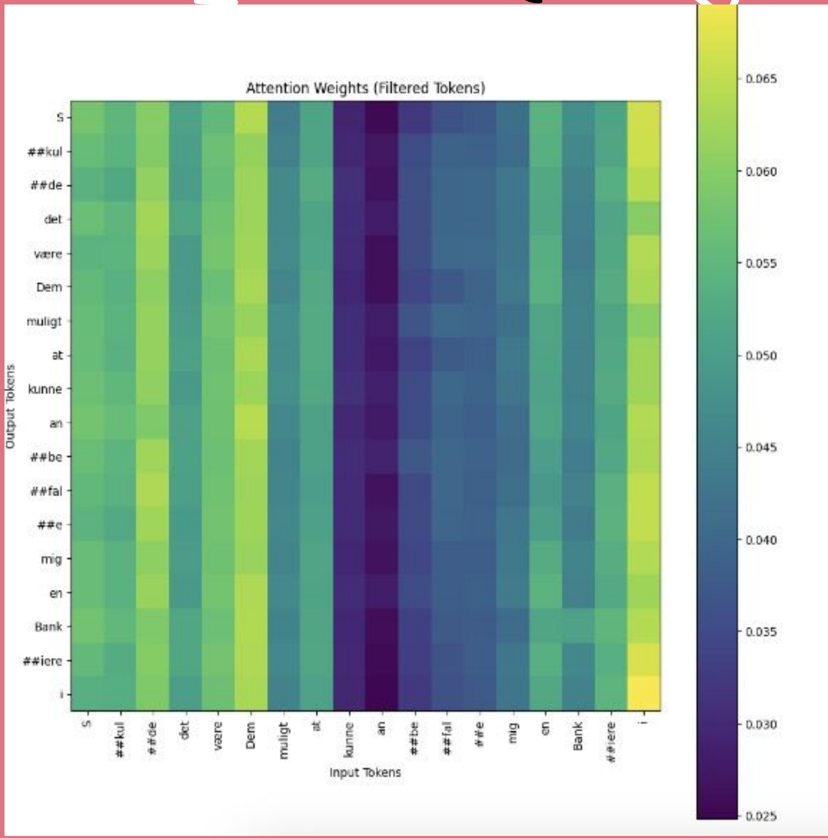
Prediction probabilities
Henrik Ibsen 1.00
James Joyce 0.00
Virginia Woolf 0.00
Johann Wolfga... 0.00
Other 0.00
NOT Franz Kafka
Franz Kafka
Text with highlighted words
Skulde det være Dem muligt at kunne anbefale mig en Bankiere i Salzburg, till hvem jeg kan sælge den, saa takker jeg meget. Er min Svigermoder i Kbhvn, saa bedes hun hilset; ligesaa alle Venner og Bekjendte. Det er utaaleligt i Længden at være udenfor al Communication med Hjemmet
NOT Friedrich Schiller
Friedrich Schiller
NOT Henrik Ibsen
Henrik Ibsen
NOT James Joyce
James Joyce
NOT Virginia Woolf
Virginia Woolf
NOT Wilhelm Busch
Wilhelm Busch

Attention Mechanism

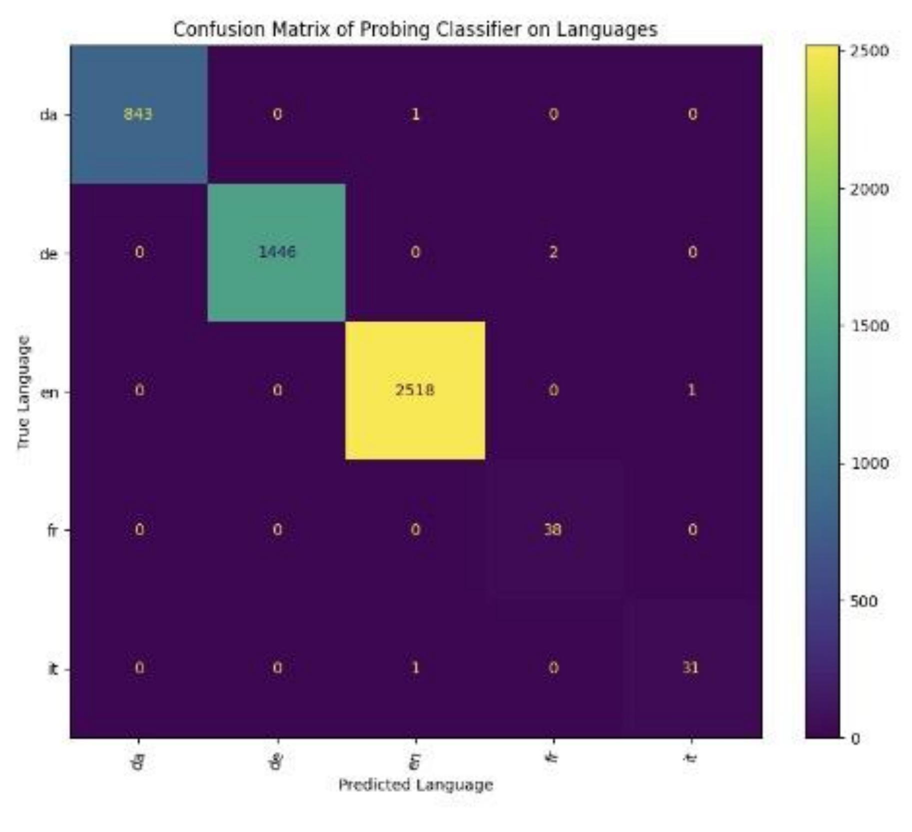
Many tokens (e.g., **##kul, muligt**) attend most strongly to themselves, as seen in the brighter diagonal.

Also, they exhibit more consistent attention to a range of other tokens → Importance in the broader sentence meaning

Some tokens (e.g., Bank) distribute attention more broadly to other tokens, as seen by higher intensity off-diagonal weights.



Problems in the language classification task



Probing Classifier Accuracy: 0.9990

	precision	recall	f1-score	support
da	1.00	1.00	1.00	844
de	1.00	1.00	1.00	1448
en	1.00	1.00	1.00	2519
fr	0.95	1.00	0.97	38
it	0.97	0.97	0.97	32
accuracy			1.00	4881
macro avg	0.98	0.99	0.99	4881
weighted avg	1.00	1.00	1.00	4881

7 classes in training dataset

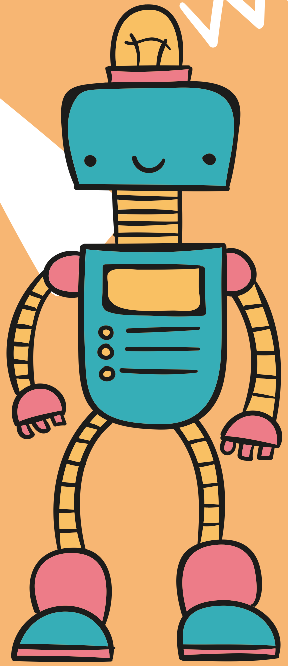
Only 5 in evaluation dataset

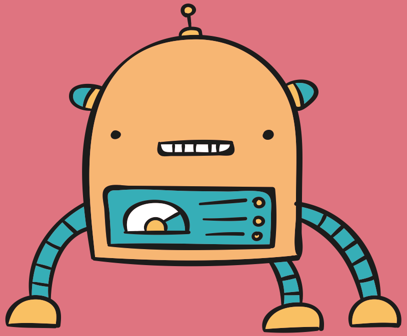
- Incomplete picture of the model's performance across all 7 classes
- Imbalanced dataset

04.

News Category Classification

by Zixuan



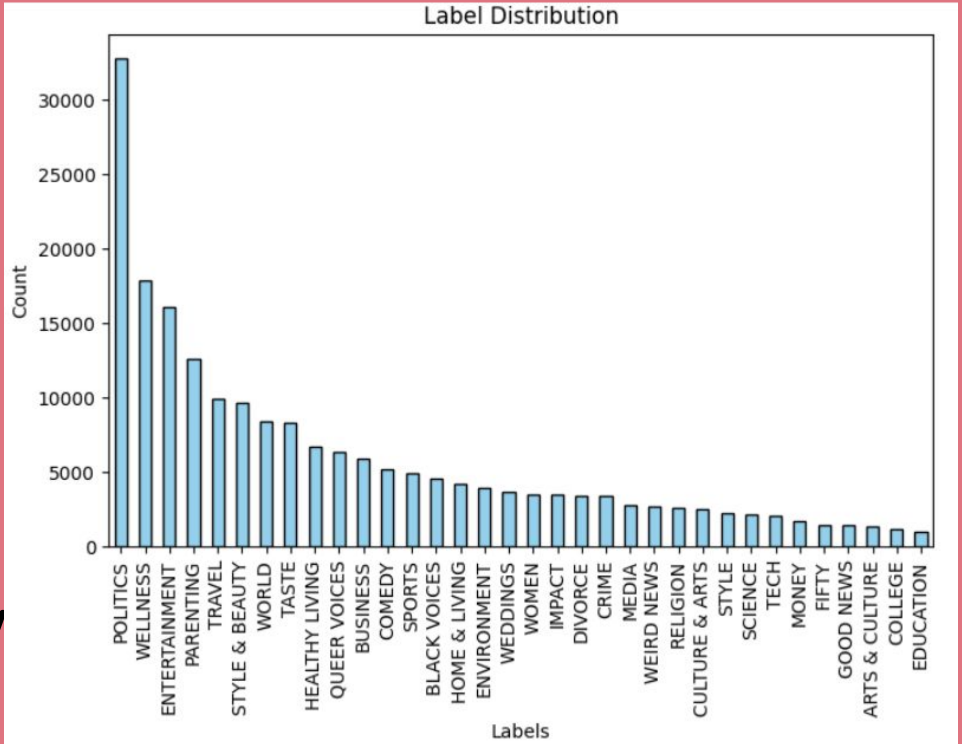


Know your Data

Is the dataset balanced?

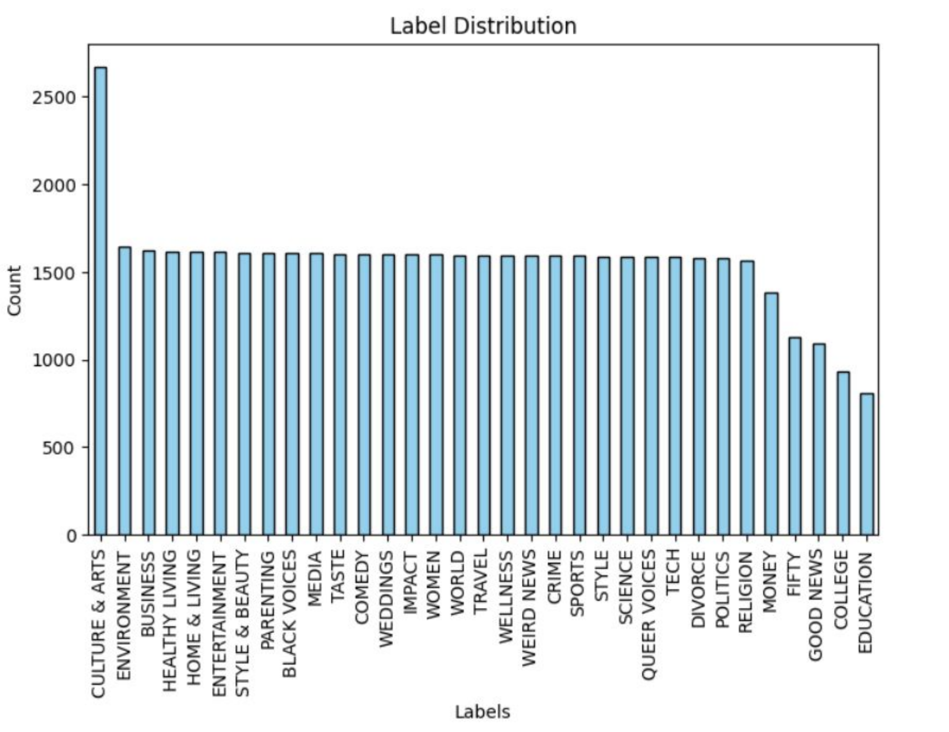
What are the features?

Unbalanced as a Whole



The whole dataset is unfortunately unbalanced.

Balanced in the Train Data



Generally, we can apply some sampling technique to balance the train data as possible, as done in the given one.

The Features to Consider

Headline
most informative

Description
additional information

Authors
contain self-introduction

Link
partly human-readable

Date
seemingly irrelevant

Some Authors with Introductions

The Pet Collective, Contributor, Your daily source for cute, funny, informative, and heartwarmi... **(ENVIRONMENT)**

Shahir ShahidSaless, Contributor. An Iranian-Canadian political analyst. He has extensively writ... **(POLITICS)**

Sister Jenna, Contributor, Award-winning Spiritual Mentor, Author, Host of the Popular Am... **(WELLNESS)**

Links can Contain Keywords

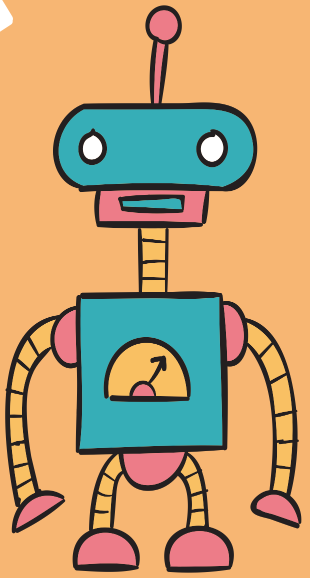
https://www.huffingtonpost.com/entry/***walking-for-stress*** _us_ 5b9d051ee4b03a1dcc83cce9 **(WELLNESS)**

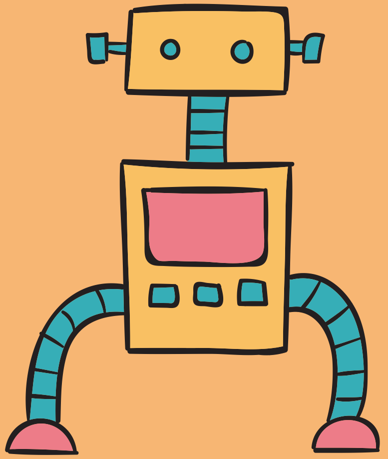
https://www.huffingtonpost.com/entry/***justin-trudeau-state-di nner-climate-change*** _us_56e1805be4b0b25c9180e224 **(POLITICS)**

Combine the Features, but

Which

How





Candidates of Selected Features

All

BERT is strong.

Why bother?

Clean Links

Let me make it more human-readable for you, if you are one...

No Date

"Not every day counts."

Candidates of Combining Methods

[SEP]

Sometimes the classic one is the best.

HTML-Style

<description>It provides additional signals. doesn’t it?</description>

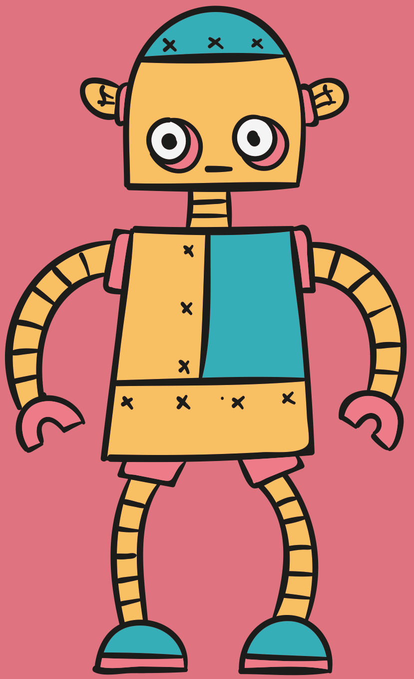
No Delimiter

Minimalist

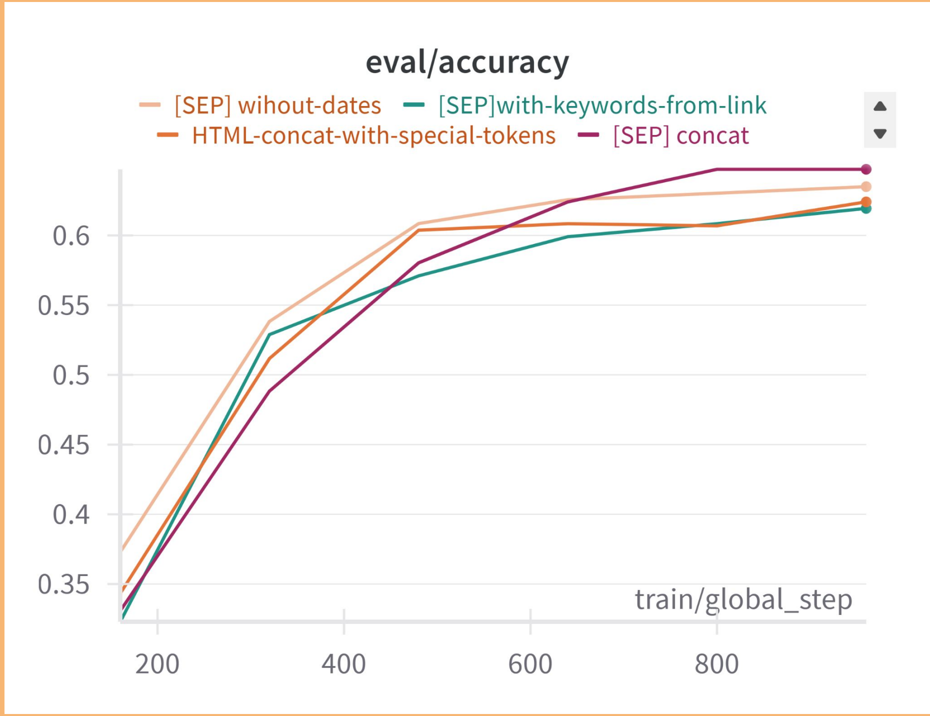
Wait, that was too much!

It is very expensive to train a full model for each possibility.

Let's experiment them on a small dataset.

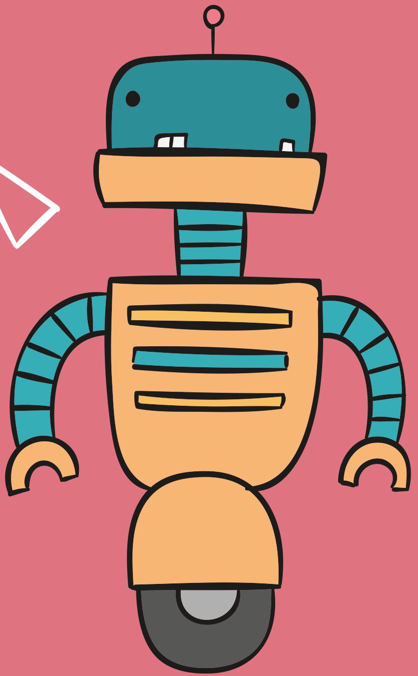


Some Insights from Experiments



In a word, the [SEP] with all features **simply wins** .

*Not all lines are displayed.



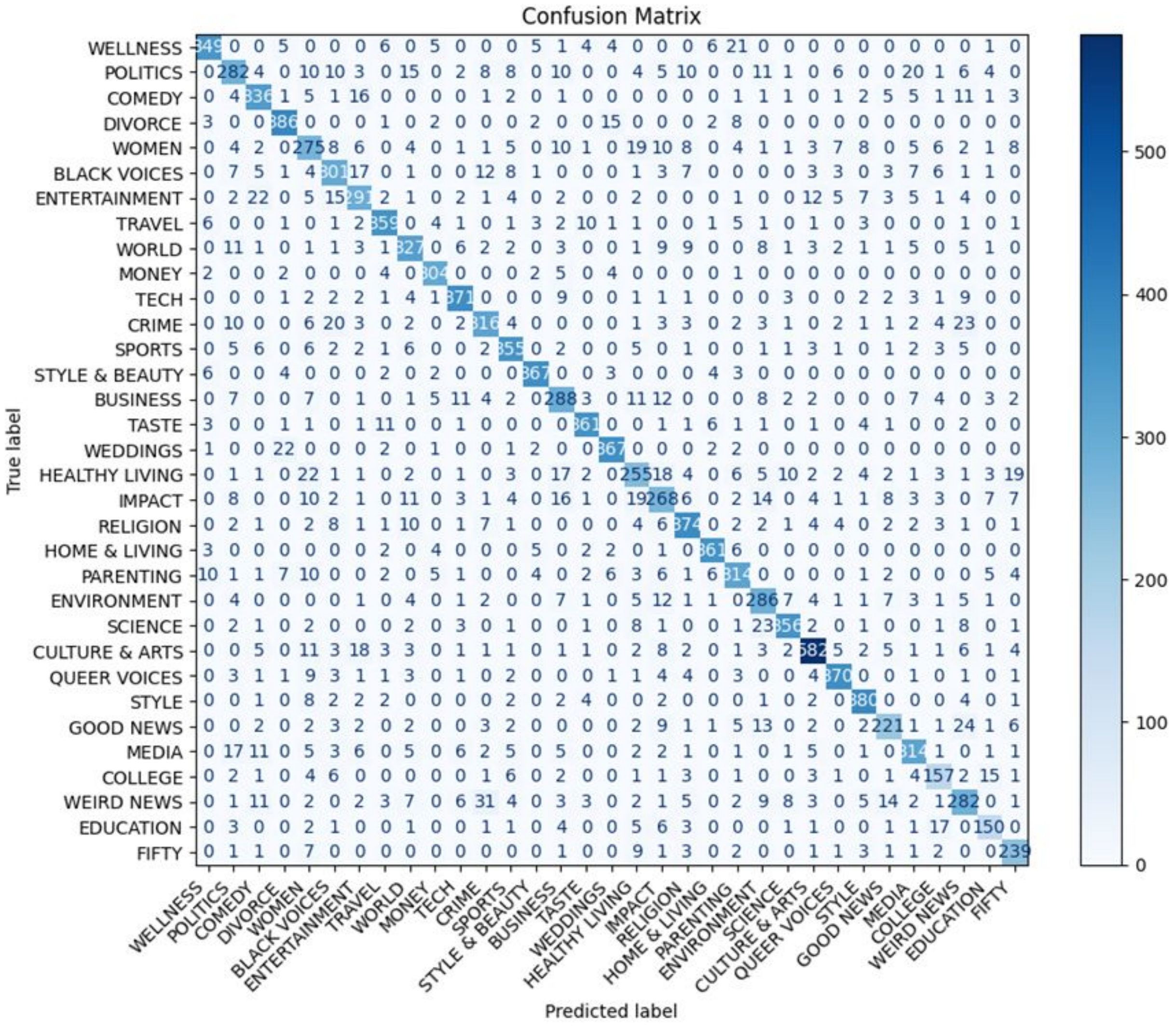
Time to scale up.

We will choose [SEP]-all and HTML-style-all to train a model on the full dataset.

Training on the Full Dataset

	Accuracy	F1-Score	Precision	Recall
[SEP]-All	0.827	0.823	0.823	0.824
HTML-Style -All	0.822	0.817	0.817	0.818

The performance of [SEP]-Full slightly exceeds the HTML-Style's, as expected on the smaller dataset. (3 epochs)

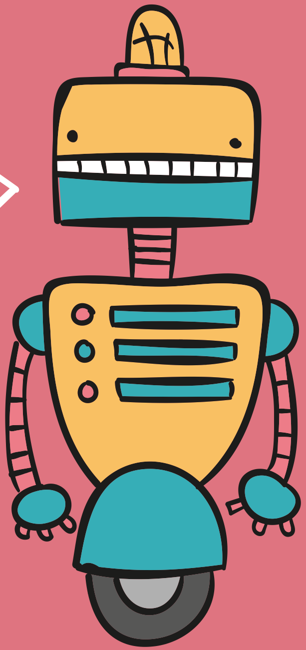


The most confused pairs:

DIVORCE vs.

WEDDINGS

*It's more like marriages.



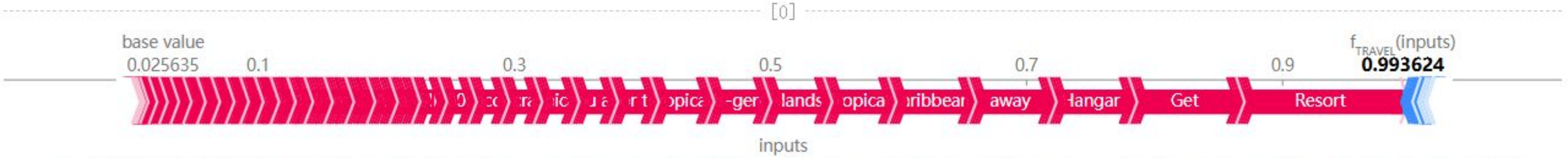
Where the model succeeds and fails, Why?

We will use the probing technique SHAP.

It will generate values indicating **how much a token contributes** to the decision.

Some Correct Predictions

Pred: TRAVEL

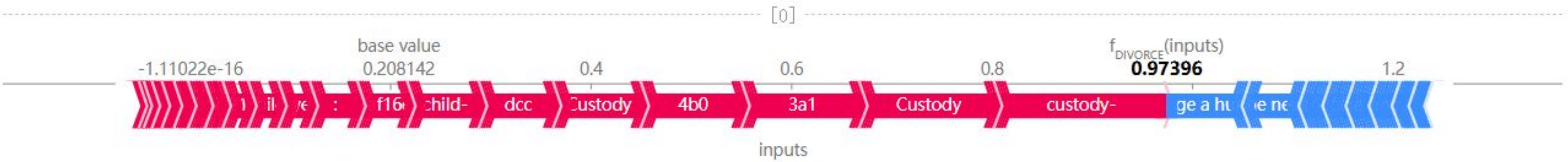


Tropical Islands Resort Is A Caribbean Getaway INSIDE A German Aircraft Hangar[SEP]Kate Auletta[SEP]https://www.huffingtonpost.com/entry/tropical-islands-germany_us_5b9cd620e4b03a1dcc824671[SEP]When life gives you an aircraft hangar, make an indoor tropical island. Oh, and let's not forget a Tropical Sea the size[SEP]2013-02-19 00:00:00

True: TRAVEL

Easy and correct.

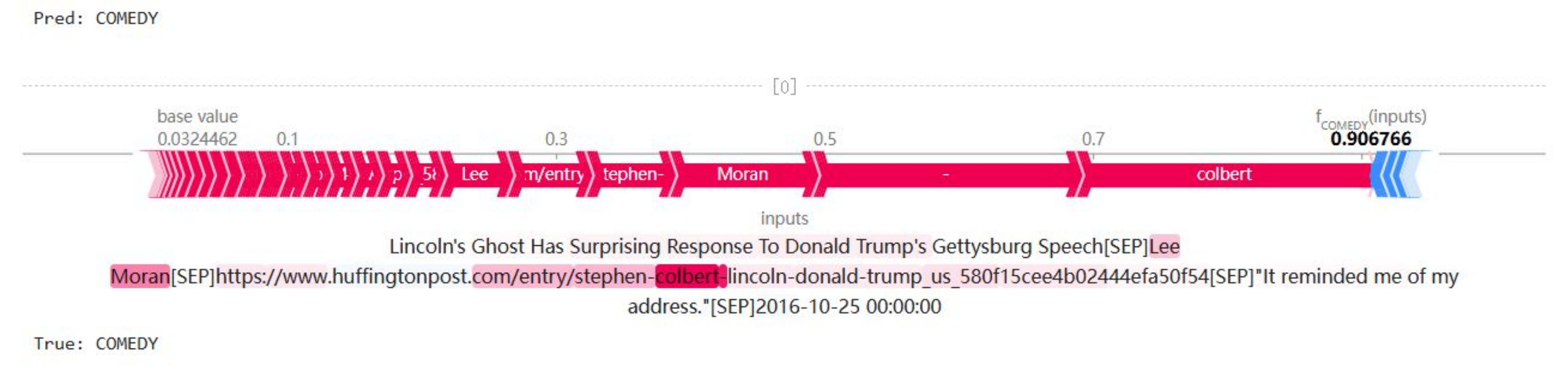
Pred: DIVORCE



Child Custody: Obese Father Plans Hunger Strike After Losing Custody[SEP][SEP]https://www.huffingtonpost.com/entry/child-custody-obese-fathe_us_5b9be1b5e4b03a1dcc7af16e[SEP]The 38-year-old, 360-pound man will appeal the case and stage a hunger strike at Ottawa's Parliament Hill Friday, the news[SEP]2012-06-21 00:00:00

True: DIVORCE

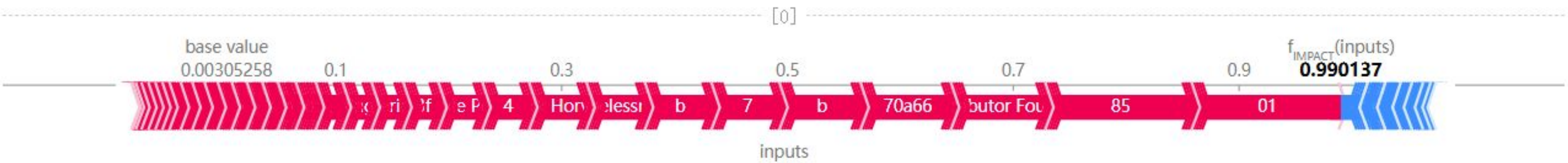
Sharp and correct.



It recognized the comedian.

We can discuss whether remembering the author name is a good thing though.

Pred: IMPACT



Check out the House That FreeFest Built[SEP]Mark Horvath, Contributor Founder, Invisible People[SEP]https://www.huffingtonpost.com/entry/check-out-the-house-that_us_5bb1b427e4b01857b70a663f[SEP]This Invisible People road trip is made possible by Sevenly and Virgin Mobile USA, who are partnering to end youth homelessness[SEP]2013-09-21 00:00:00

True: IMPACT

and cases where you just don't know why it's so confident and correct...

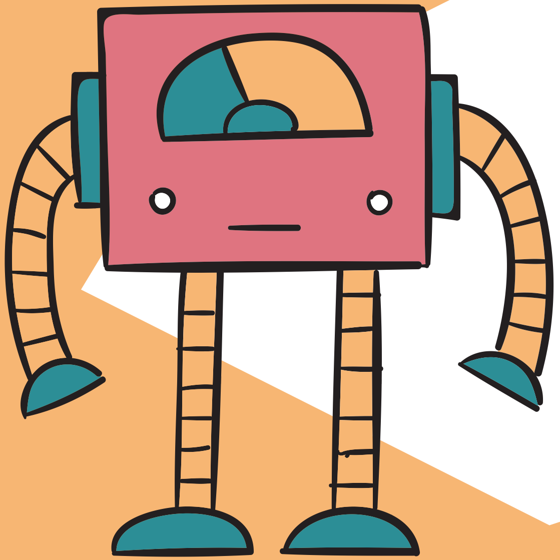
* remember we tried Regex and it was counter-effective.

Possible Reasons for Mistakes

Overlapping in categories

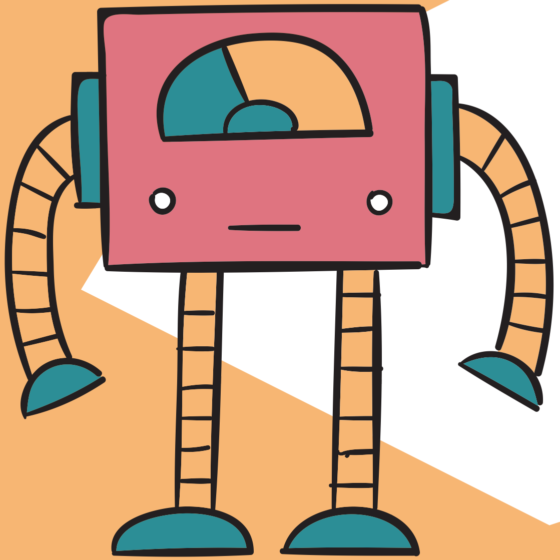
e.g. COMEDY vs. ENTERTAINMENT,
CULTURE & ARTS vs. TRAVEL

* In some cases, the information is not sufficient to disambiguate these confusion pairs.

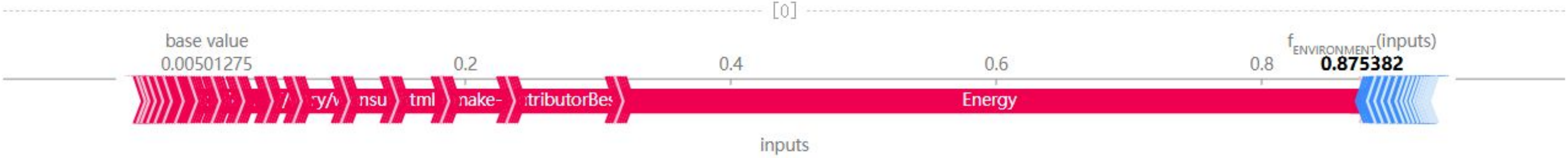


Possible Reasons for Mistakes

The model failed to grasp the focus of a text contextually, when there were multiple representative words.



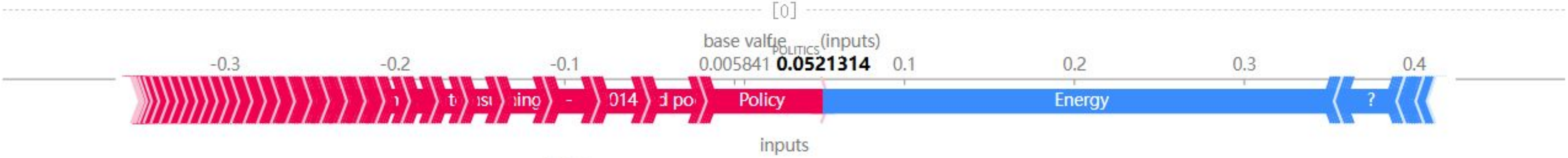
Pred: ENVIRONMENT



0.561

Will We Continue to Make the Same Mistakes on EnergyPolicy?[SEP] Cliff Schecter, ContributorBest-selling Author, Public Relations Consultant, Daily Beast ...[SEP]https://www.huffingtonpost.com/entry/will-we-continue-to-make-_b_5510953.html[SEP]Of course, in Washington, following the money is always a sound principle for explaining repetition of failed policy.[SEP]2014-06-20 00:00:00[SEP]

True: POLITICS



-0.304

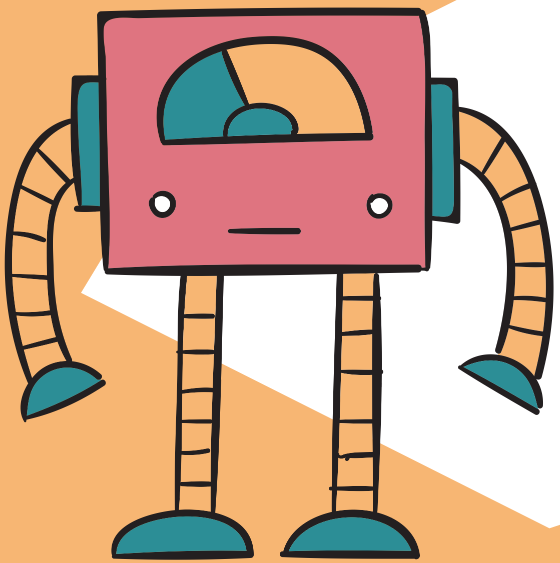
Will We Continue to Make the Same Mistakes on EnergyPolicy?[SEP] Cliff Schecter, ContributorBest-selling Author, Public Relations Consultant, Daily Beast ...[SEP]https://www.huffingtonpost.com/entry/will-we-continue-to-make-_b_5510953.html[SEP]Of course, in Washington, following the money is always a sound principle for explaining repetition of failed policy.[SEP]2014-06-20 00:00:00[SEP]

Possible Reasons for Mistakes

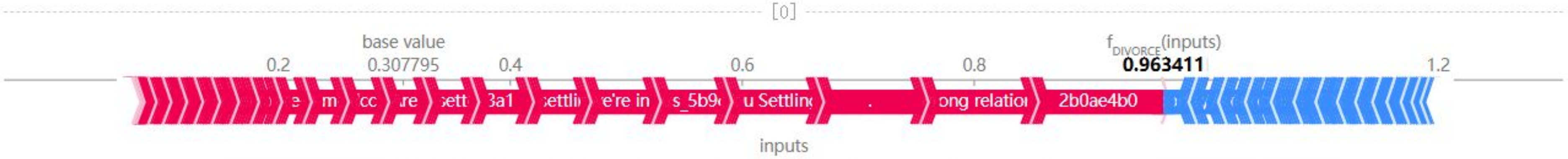
The model oversimplified some categories, instead of understanding like a human.

This happened a lot in the pair:

DIVORCE vs. WEDDINGS

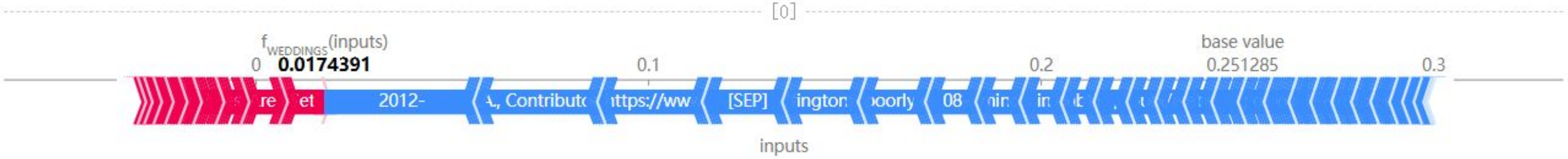


Pred: DIVORCE



Are You Settling?[SEP] Sheryl Paul, M.A., Contributor Anxiety expert[SEP]https://www.huffingtonpost.com/entry/are-you-settling_us_5b9c2b0ae4b03a1dcc7cca17[SEP]Because we're poorly educated about transitions in our culture, we mistake the fear for doubt and thus begins a scary domino effect of believing that we're in the wrong relationship. The message is: If you're doubting, you must be settling.[SEP]2012-08-23 00:00:00[SEP]

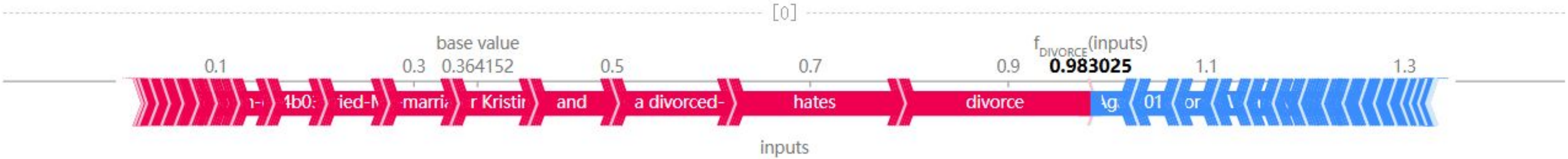
True: WEDDINGS



Are You Settling?[SEP] Sheryl Paul, M.A., Contributor Anxiety expert[SEP]https://www.huffingtonpost.com/entry/are-you-settling_us_5b9c2b0ae4b03a1dcc7cca17[SEP]Because we're poorly educated about transitions in our culture, we mistake the fear for doubt and thus begins a scary domino effect of believing that we're in the wrong relationship. The message is: If you're doubting, you must be settling.[SEP]2012-08-23 00:00:00[SEP]

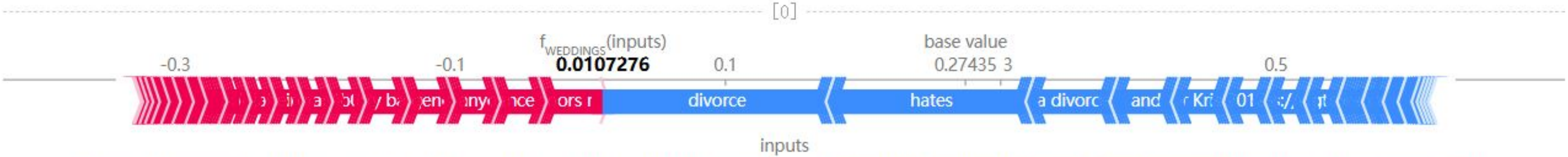
Negative words don't necessarily mean we want to divorce.

Pred: DIVORCE



Using Marriage As A Weapon Against Poverty Hurts Women[SEP] Kristin Tennant, Contributor Kristin Tennant is a divorced-Christian-liberal-remarried-Midw...[SEP]https://www.huffingtonpost.com/entry/using-marriage-as-a-weapo_us_5b9c483ee4b03a1dcc7d8cbb[SEP]My general stance is pretty basic: Don't take marriage lightly. It also seems like a fairly non-controversial stance -- one that anyone who truly honors marriage and hates divorce could get behind, whatever their political or religious leaning.[SEP]2012-09-14 00:00:00[SEP]

True: WEDDINGS

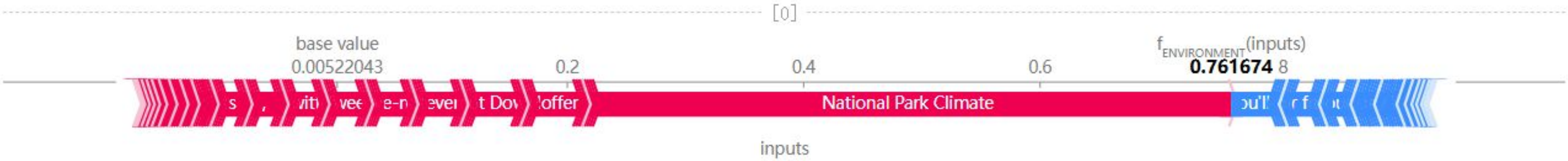


Using Marriage As A Weapon Against Poverty Hurts Women[SEP] Kristin Tennant, Contributor Kristin Tennant is a divorced-Christian-liberal-remarried-Midw...[SEP]https://www.huffingtonpost.com/entry/using-marriage-as-a-weapo_us_5b9c483ee4b03a1dcc7d8cbb[SEP]My general stance is pretty basic: Don't take marriage lightly. It also seems like a fairly non-controversial stance -- one that anyone who truly honors marriage and hates divorce could get behind, whatever their political or religious leaning.[SEP]2012-09-14 00:00:00[SEP]

hates + divorce = DIVORCE?

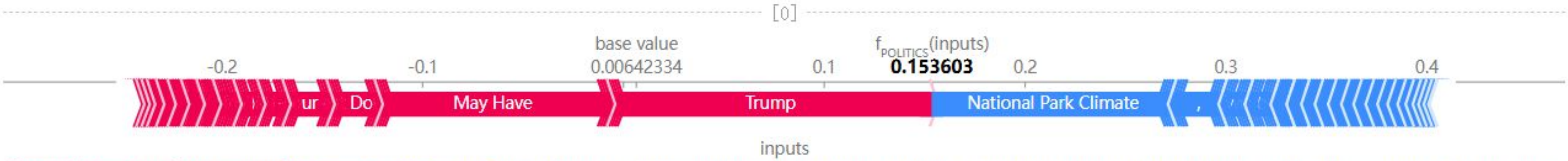
It behaved like a naive bag of words :c

```
Pred: ENVIRONMENT
```



Trump May Have Shut Down National Park Climate Tweets, But The Internet Won’t Let It Die[SEP]Steven Hoffer[SEP]https://www.huffingtonpost.com/entry/alternative-national-park-twitter_us_5888c087e4b0b481c76c2820[SEP]"You can take our official twitter, but you'll never take our free time!"[SEP]2017-01-25 00:00:00

```
True: POLITICS
```



Trump May Have Shut Down National Park Climate Tweets, But The Internet Won’t Let It Die[SEP]Steven Hoffer[SEP]https://www.huffingtonpost.com/entry/alternative-national-park-twitter_us_5888c087e4b0b481c76c2820[SEP]"You can take our official twitter, but you'll never take our free time!"[SEP]2017-01-25 00:00:00

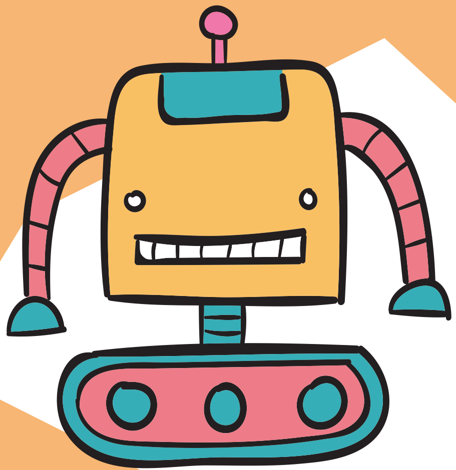
Similarly...

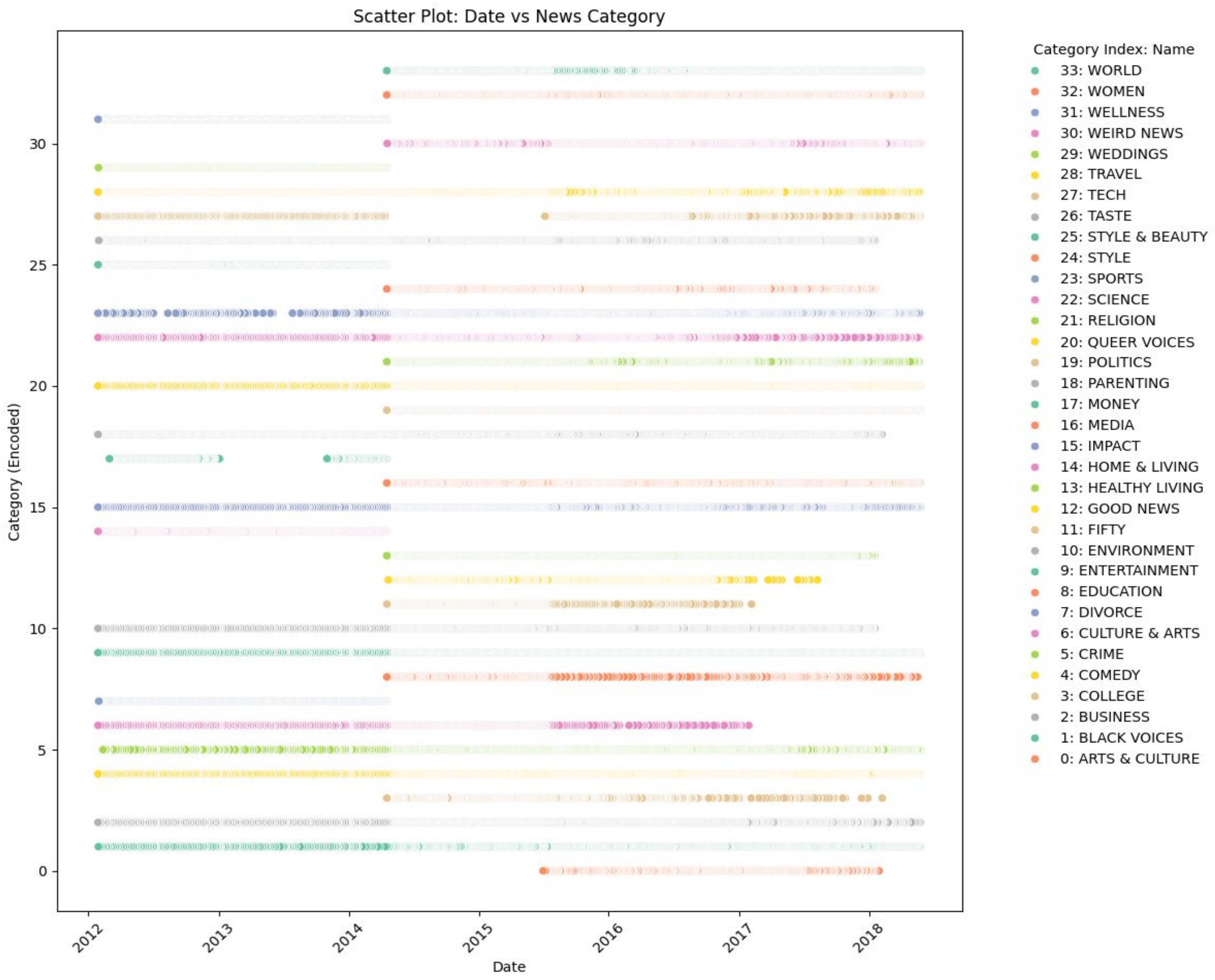
Just because we wrote a wedding there doesn't mean we are talking about it.

Still Something Interesting?

Yes, the DATES were sometimes contributing to the decision.

This means they are discriminative.





There are certain regularities* in the distribution of dates,

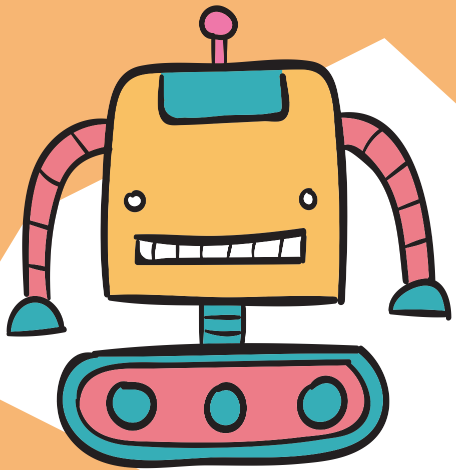
but these can be biases in the dataset.

* I realized it afterwards

Are we cheating or not?

The regularity of dates remain in both train_data and eval_data, but there is a high chance that it cannot be generalized.

In order to be fair, I trained another model without dates on the full dataset, expecting it to be much worse…



Well, it's not that bad!

	Accuracy	F1-Score	Precision	Recall
[SEP]-no -dates	0.818	0.814	0.814	0.815

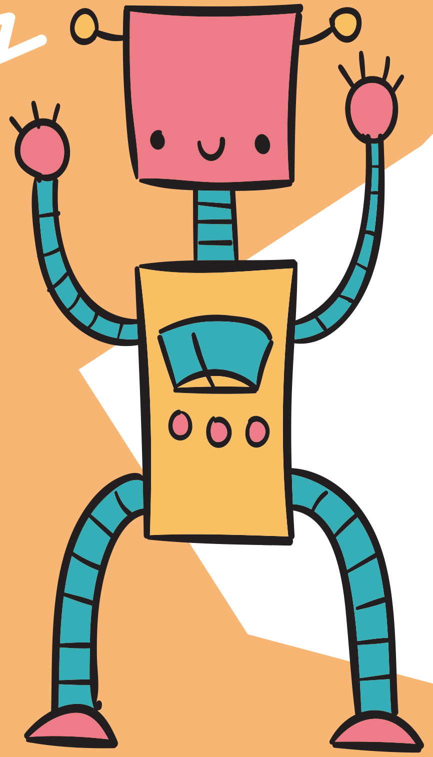
Considering this is only trained for 2 epochs, it's not worse than with dates anyway :D

It seems the model can still catch enough signals to make most decisions right.

What we know about BERT in this task

Generally speaking, BERT is: versatile and contextually powerful, but expensive to train.

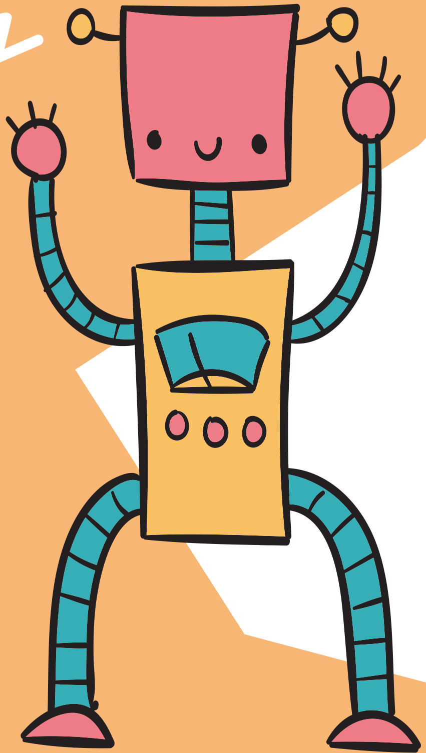
We basically confirmed these, but in this task, it did not behave very contextually in more difficult tasks either.



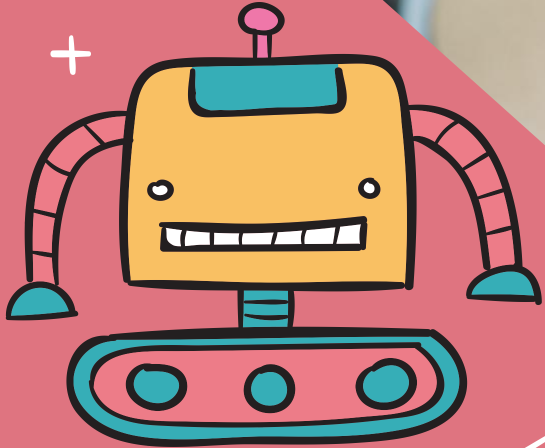
What to do next?

Is it possible that we just haven't unleashed the full power of BERT?

Try to cherry-pick those difficult samples and specifically finetune BERT.
Also try to do it without links given that sometimes the model relies on the identifier.

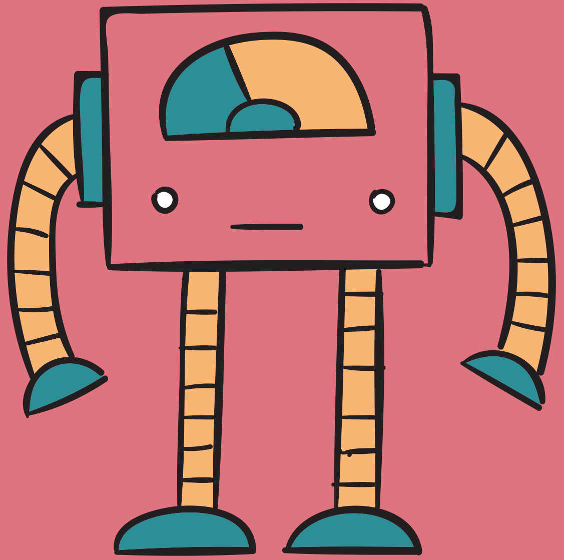


Project Report: Time, Cost, and Outcome



2.6 kg CO_2

The total carbon footprint generated throughout the project's lifecycle.



Environmental impact

Transportation

- Taking a short **domestic flight** of about **15–20 minutes**. (Flights emit around 133 g CO_2 per passenger per kilometer.)
- Driving a typical gasoline-powered car for about **10 kilometers (6 miles)**.

Household Activitieskness

- Using a **microwave oven** continuously for **3 hours**.
- Leaving an **LED light bulb (10W)** on for **23 days** straight.

Waste Disposal

- Producing and discarding about **8 plastic bottles**.

Digital Activities

- Streaming **10 hours** of high-definition video

Project Resource Summary

~120h

Approximately **120 hours** were spent on parameter tuning and testing different configurations to enhance the model's performance.

Time

$22

The **total cost** of GPU time on Google Colab was 22$

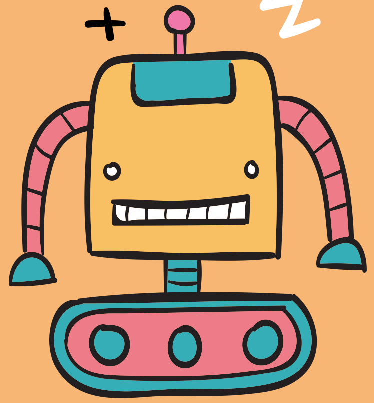
Computational Costs

2563.1MB

The data expanded significantly, increasing by a factor of 22.93, from **111.8MB to 2563.1MB** .

Data growth

Key Takeaways



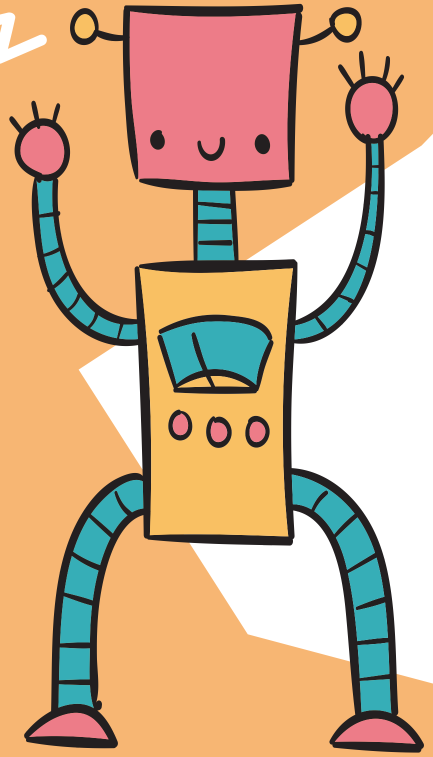
1. **Explored BERT's versatility across three tasks:**
   - Sentiment Analysis.
   - Author and Language Detection.
   - News Categorization.
2. **Technical Insights:**
   - Identified strengths and limitations of BERT across tasks.
   - Applied advanced evaluation techniques:

     Confusion Matrices – for performance analysis.

     SHAP & LIME – for interpretability.

     Attention Visualization & Probing – to understand decision-making.
3. **Long-term Vision**
   - Highlighted the importance of responsible and sustainable AI development.

Thanks!



Questions?