Social Media Sentiment Analysis

Yue Li, Aihan Liu, Shuting Cai

Instructor: Amir Jafari

NLP Fall 2022

Objectives

- Identify the sentiment for **Windows** and **Mac OS** subreddits

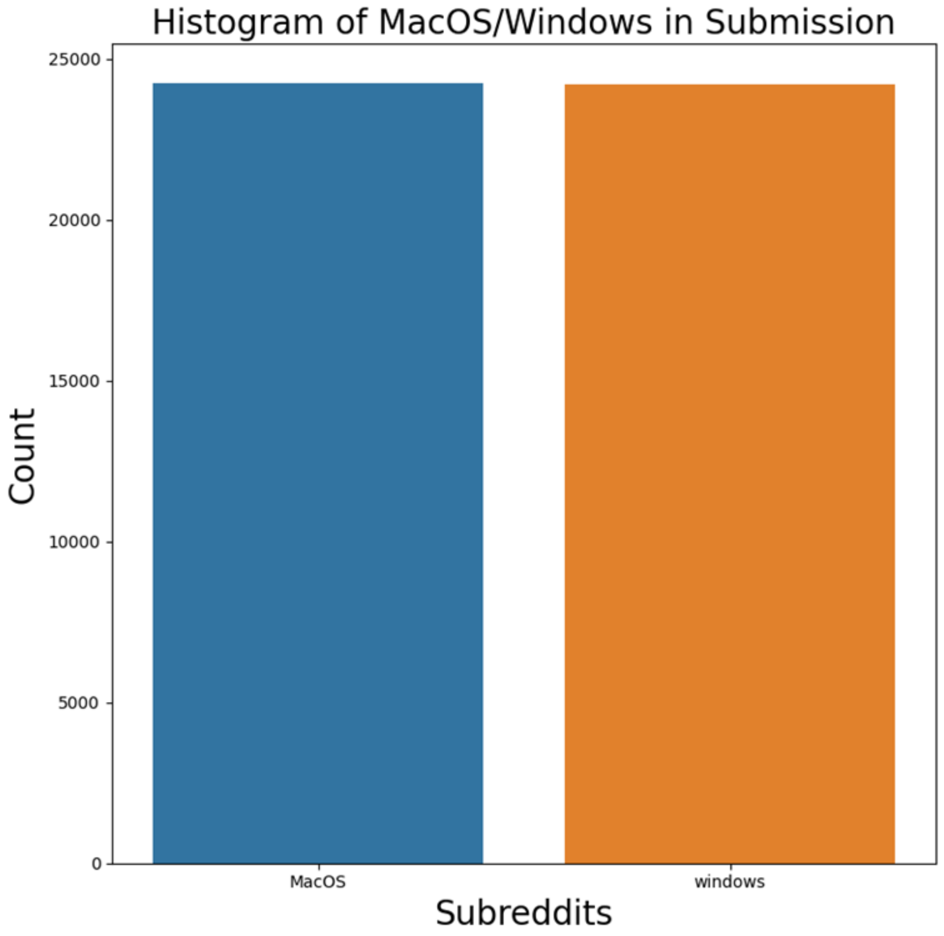
  To help Microsoft and Apple get public opinion to improve their operating system or to gain useful and related information

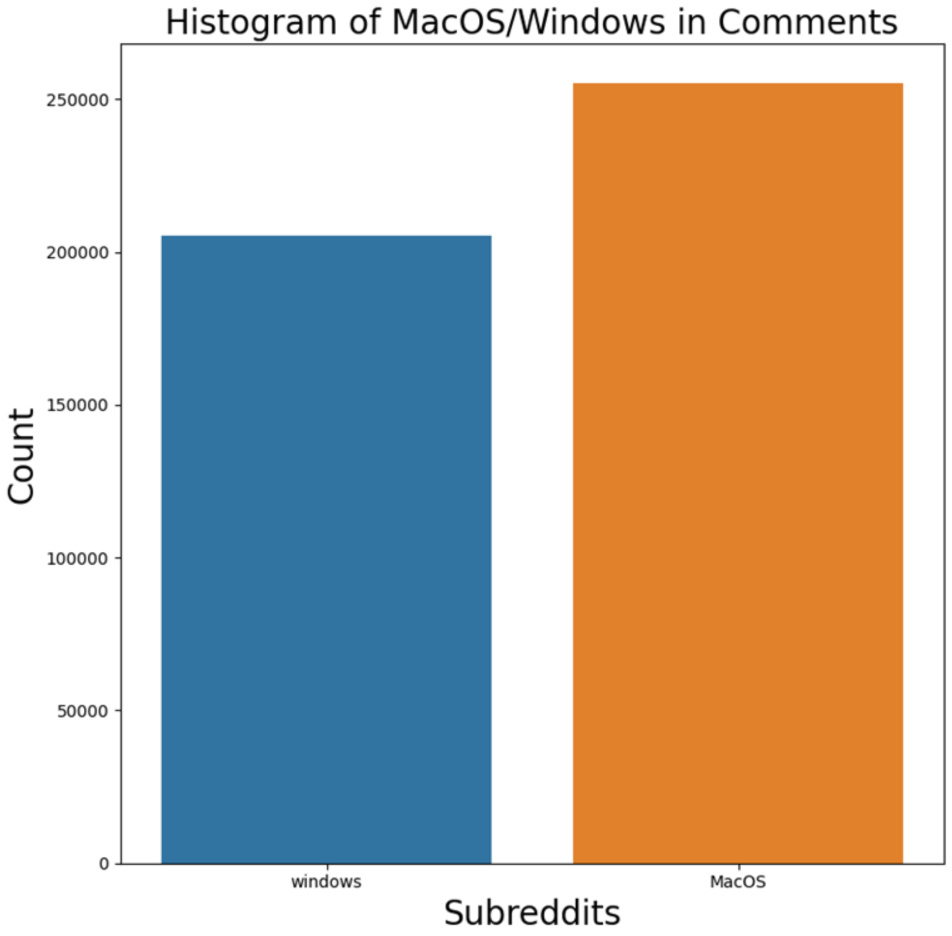
- Sentiment Analysis: **RoBERTa-base** model trained on ~58M tweets & **Distilling BERT** base trained on IMDB
- Interpretability or Explainability to determine the **importance** of word by word or sequence by sequence on the predicted sentiments

Description of Dataset

Subreddits: Windows & MacOS

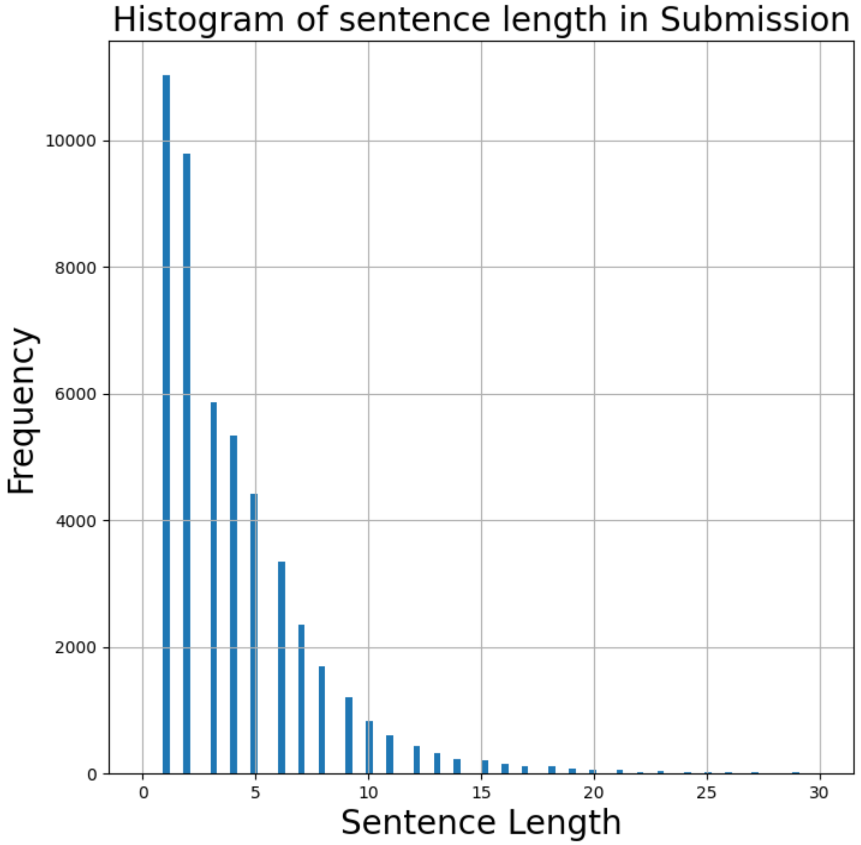
- Data Source: **Reddit** dataset collected from Pushshift API
- Dataset: Submission & Comments on **Windows** & **MacOS ~100k observations**
- Class of emotion: **Positive, Negative, Neutral**

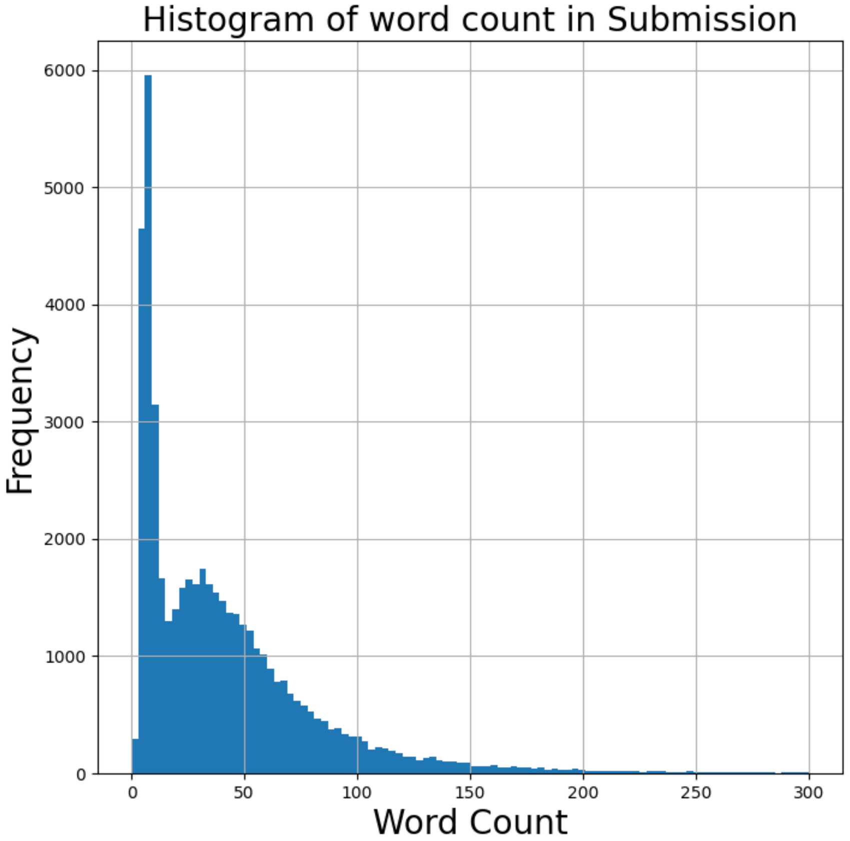


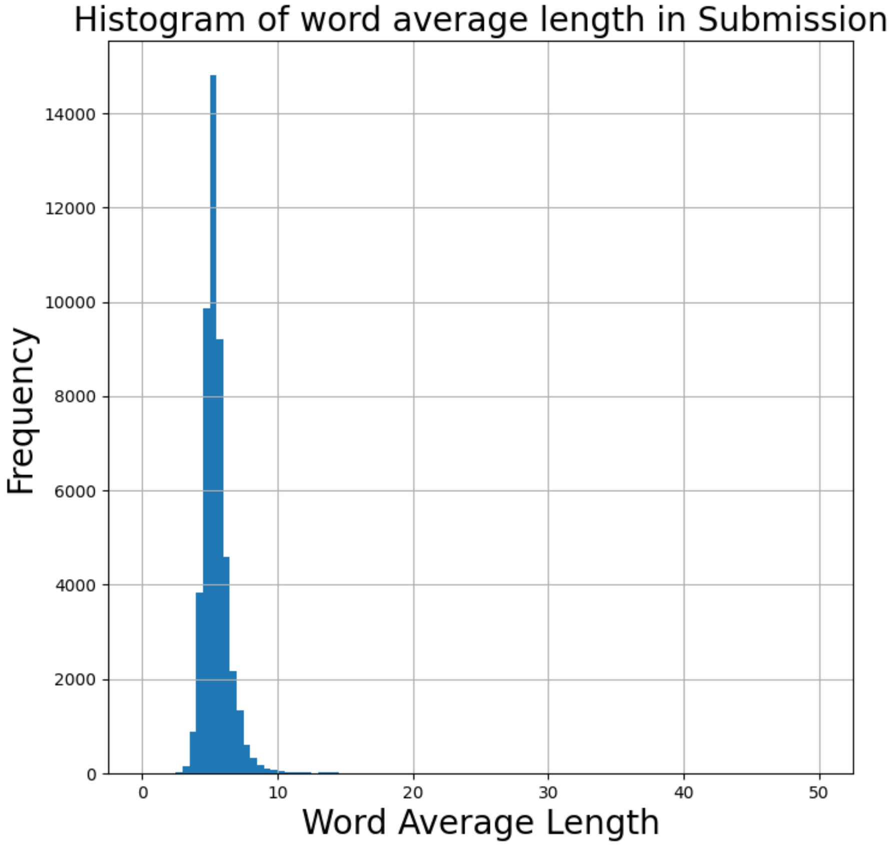


Description of Dataset

Sentences & Words


Histogram of sentence length in Submission


Histogram of word count in Submission

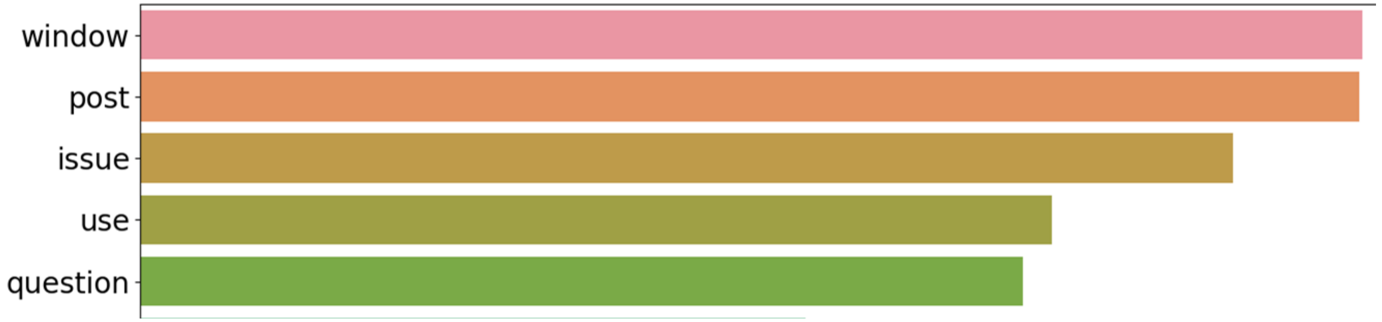

Histogram of word average length in Submission

- Most submission within 10 sentences and 50 words
- Average of 5 words in each sentence

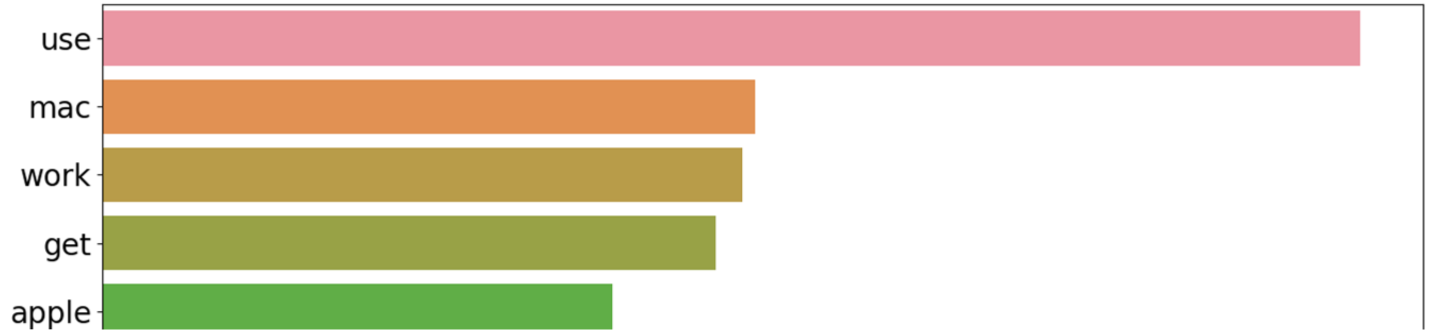
Description of Dataset

Most Common Word & Word Cloud

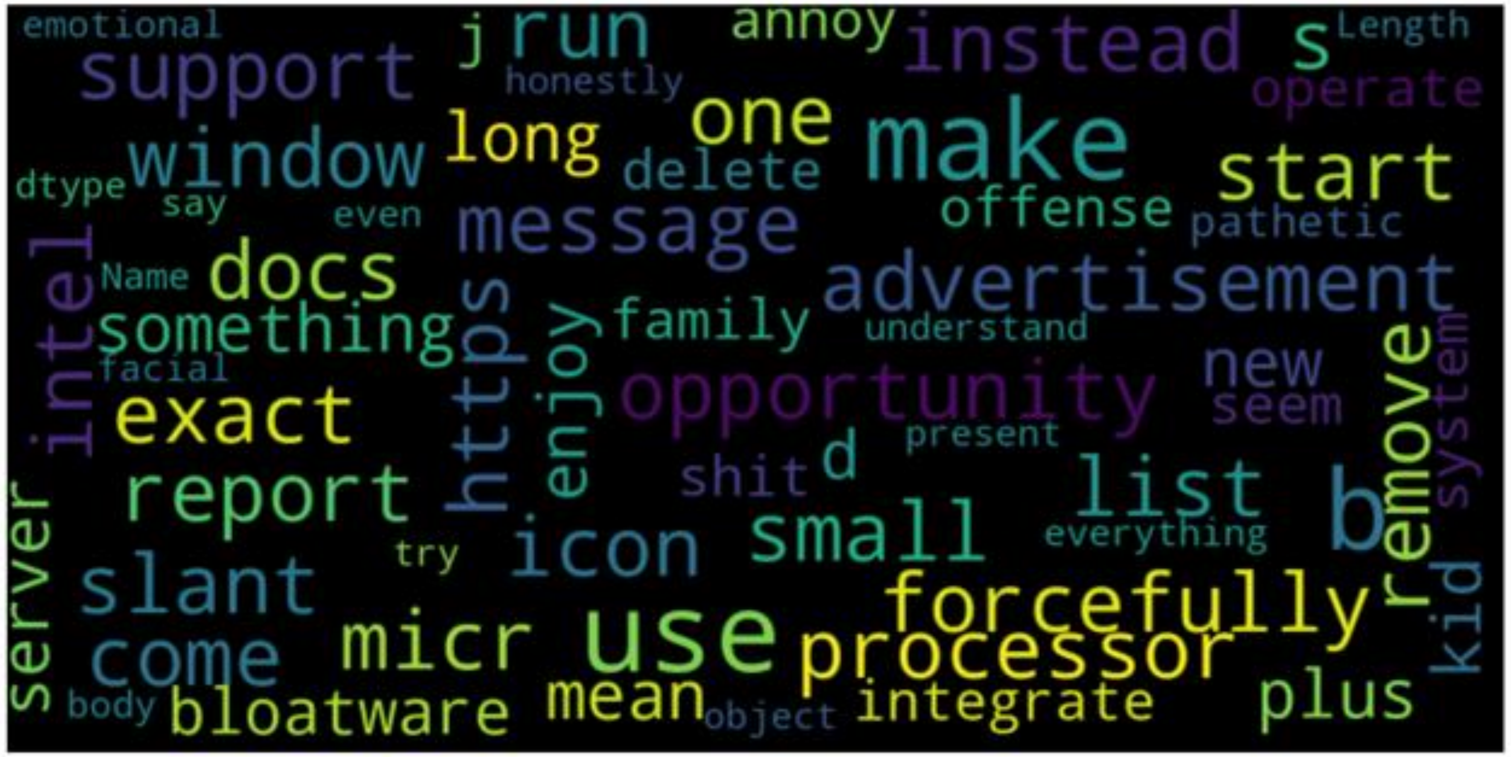
The most occurrences of word in Comments related to Windows



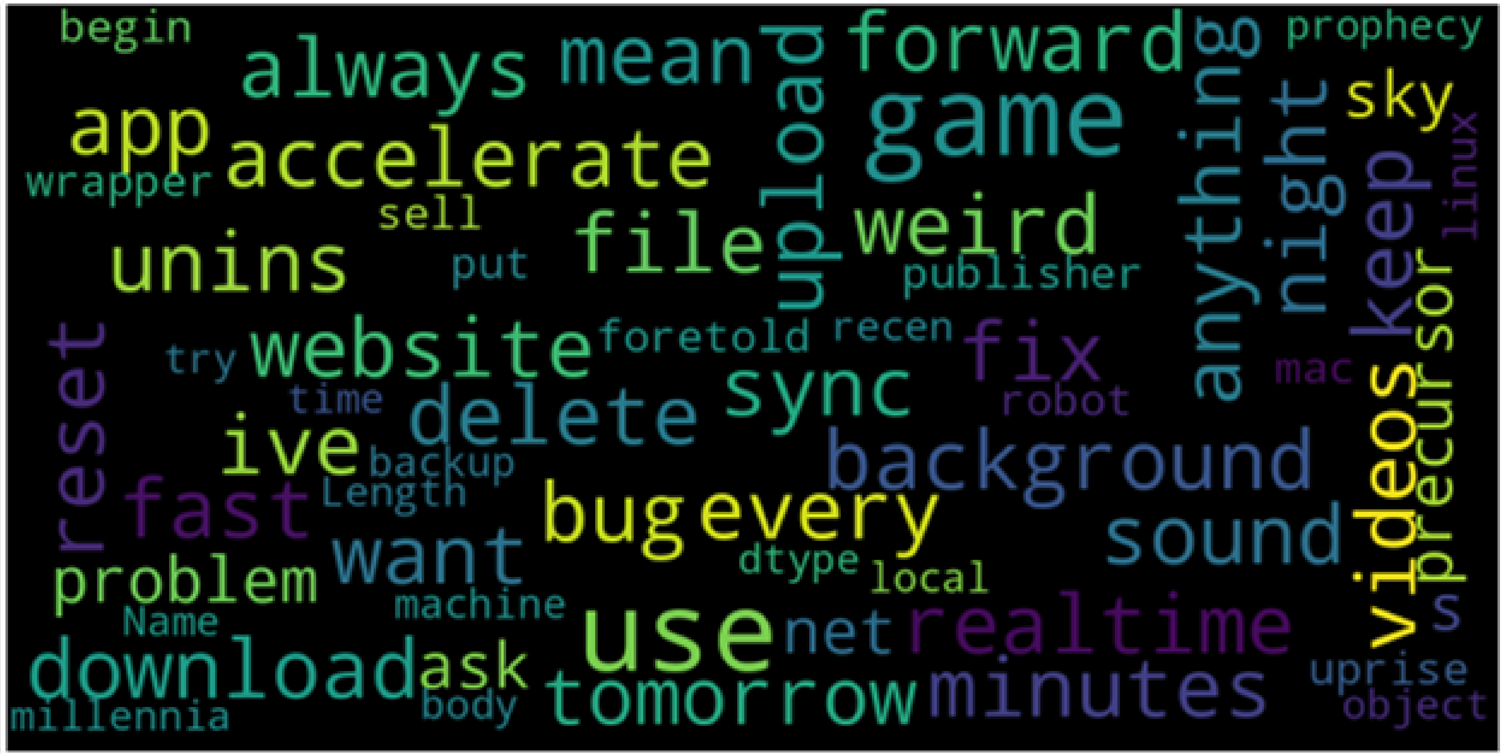
The most occurrences of word in Comments related to MacOS



Word Cloud of Windows in Comments



Word Cloud of MacOS in Comments



Attribution

- The problem of **attributing** the prediction of a deep network to its input features

Example:

- Attribute a linear model’s prediction to its features
- Attribute an object recognition network’s prediction to its pixels
- Attribute a text sentiment network’s prediction to individual works

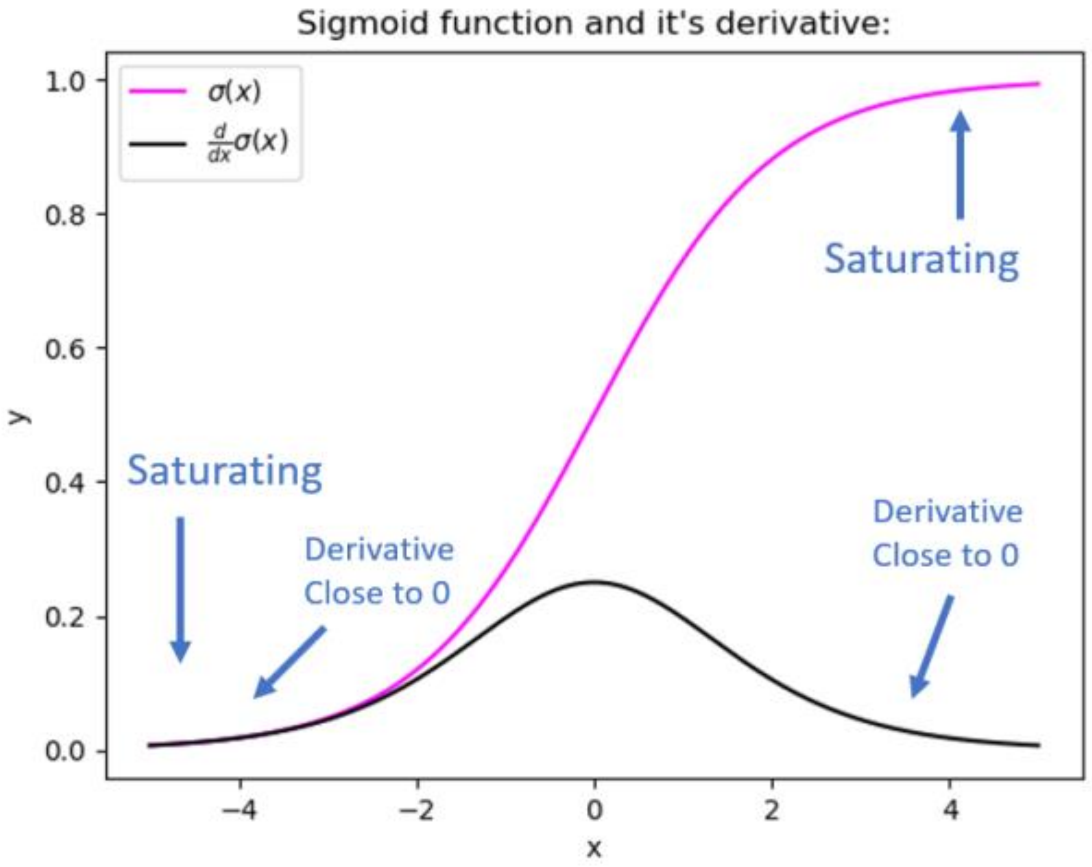
A **reductive formulation** of ‘why this prediction’ but surprisingly useful.

Gradient

- Gradients (of the output with respect to the input) can be approximated as the coefficient of the input feature for a deep network.

- Therefore the product of the gradient and feature values is a reasonable starting point for an attribution method.

- **Gradient Saturation:**



Interpretability or Explainability

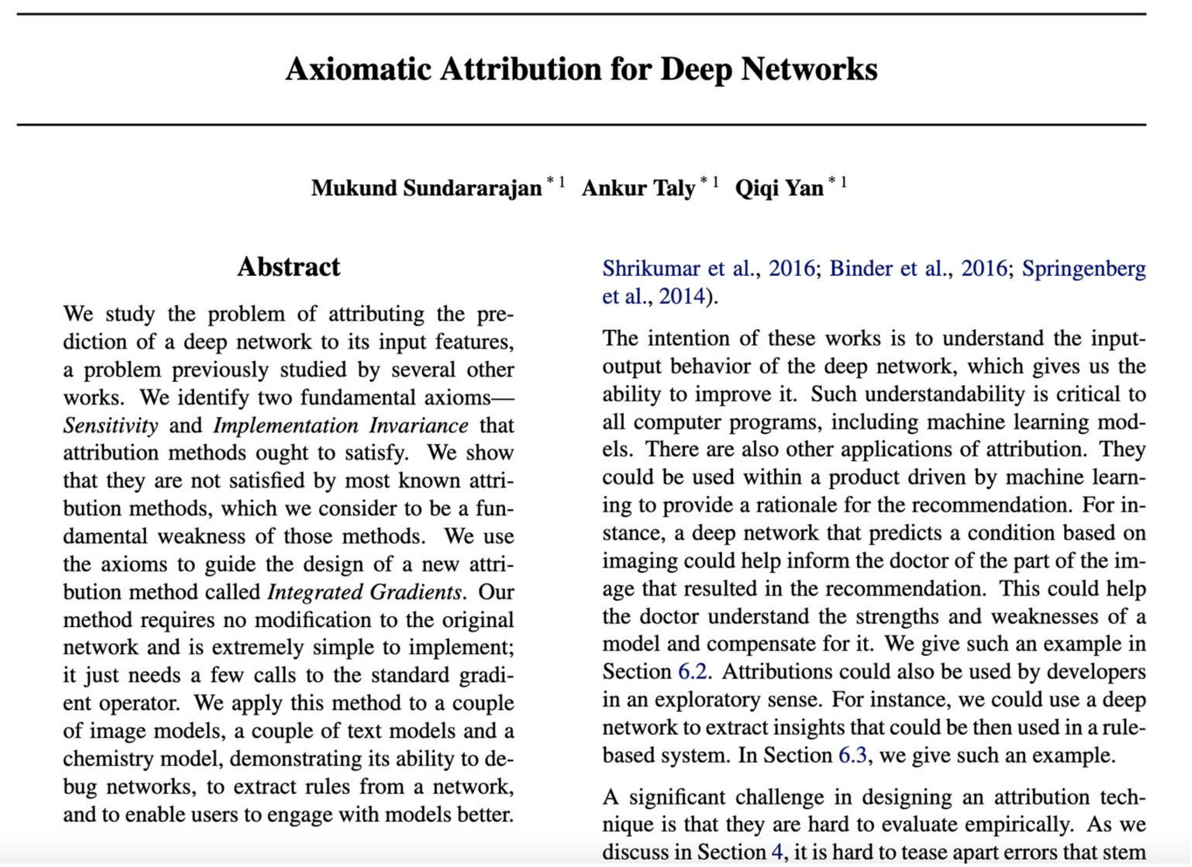
Integrated Gradient

- **Two fundamental axioms:** Sensitivity and Implementation Invariance

- **Sensitivity:** If for every input and baseline that differ in one feature but have different predictions,
then the differing feature should be given a non-zero attribution.

- **Implementation Invariance :**
Two networks are functionally equivalent if their outputs are equal for all inputs, despite having very different implementations.

$$\frac{\partial f}{\partial g} = \frac{\partial f}{\partial h} \cdot \frac{\partial h}{\partial g}$$



Axiomatic Attribution for Deep Networks

Mukund Sundararajan[*1] Ankur Taly[*1] Qiqi Yan[*1]

Abstract

We study the problem of attributing the prediction of a deep network to its input features, a problem previously studied by several other works. We identify two fundamental axioms—*Sensitivity* and *Implementation Invariance* that attribution methods ought to satisfy. We show that they are not satisfied by most known attribution methods, which we consider to be a fundamental weakness of those methods. We use the axioms to guide the design of a new attribution method called *Integrated Gradients*. Our method requires no modification to the original network and is extremely simple to implement; it just needs a few calls to the standard gradient operator. We apply this method to a couple of image models, a couple of text models and a chemistry model, demonstrating its ability to debug networks, to extract rules from a network, and to enable users to engage with models better.

Shrikumar et al., 2016; Binder et al., 2016; Springenberg et al., 2014).

The intention of these works is to understand the input-output behavior of the deep network, which gives us the ability to improve it. Such understandability is critical to all computer programs, including machine learning models. There are also other applications of attribution. They could be used within a product driven by machine learning to provide a rationale for the recommendation. For instance, a deep network that predicts a condition based on imaging could help inform the doctor of the part of the image that resulted in the recommendation. This could help the doctor understand the strengths and weaknesses of a model and compensate for it. We give such an example in Section 6.2. Attributions could also be used by developers in an exploratory sense. For instance, we could use a deep network to extract insights that could be then used in a rule-based system. In Section 6.3, we give such an example.

A significant challenge in designing an attribution technique is that they are hard to evaluate empirically. As we discuss in Section 4, it is hard to tease apart errors that stem

Interpretability or Explainability

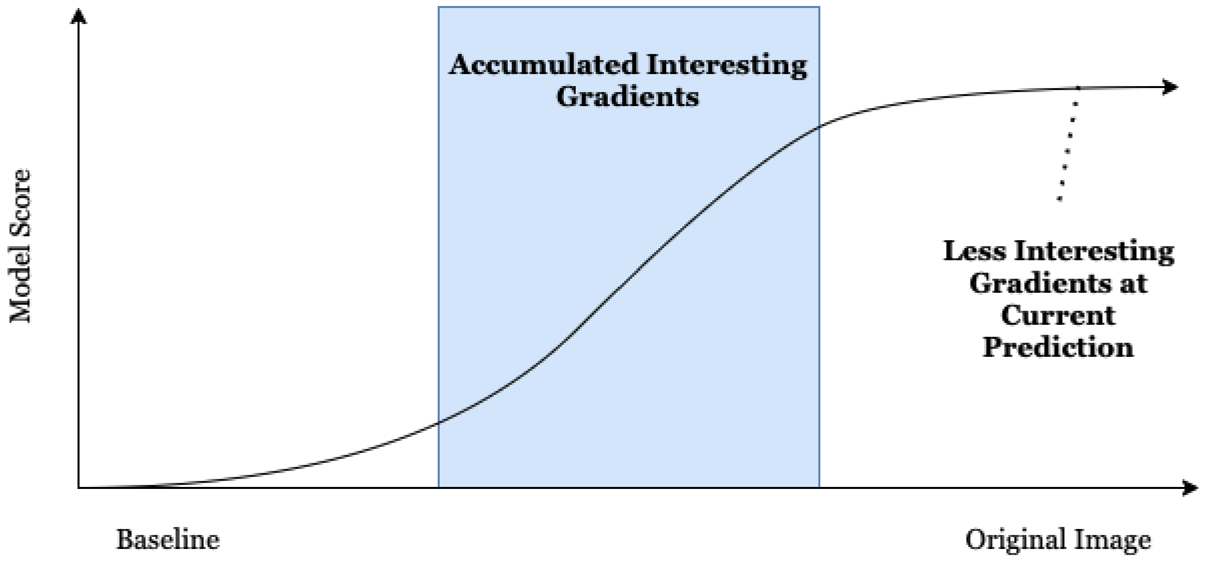
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Integrated Gradient

- **Baseline is an informationless input for the model**
  - *A person wants to sleep because he is sleepy, then he does not want to sleep when he is not sleepy.*
  - E.g., Black image for image models
  - E.g., Empty text or zero embedding vector for text model

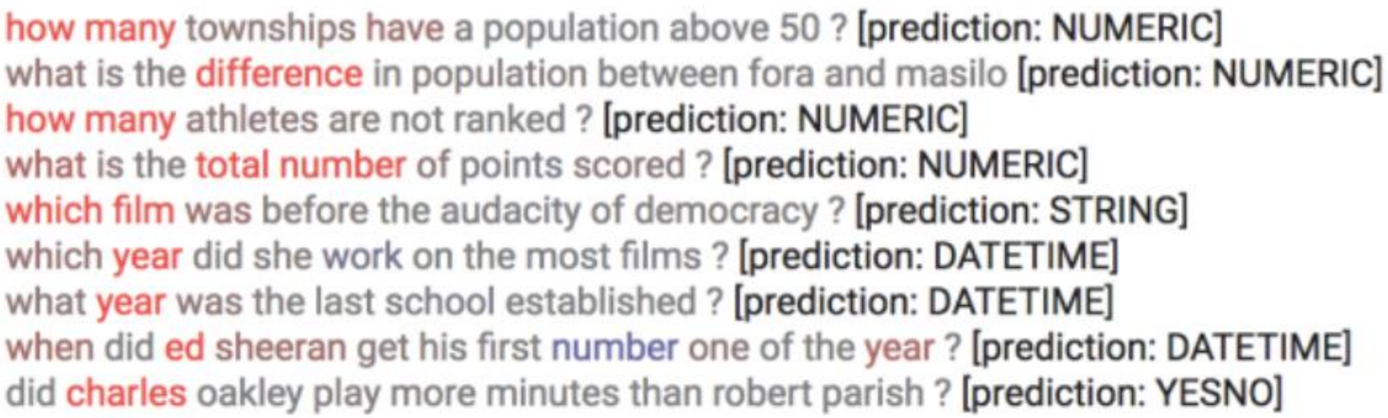
- **Integrated Gradients explain F(input) - F(Baseline) in terms of input features**

$$\begin{aligned}\phi_i^{IG}(f, \mathbf{x}, \mathbf{x}') &= \int_0^1 \frac{\delta f(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))}{\delta x_i} d\alpha (x_i - x_i') \\ &= (x_i - x_i') \int_0^1 \frac{\delta f(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))}{\delta x_i} d\alpha\end{aligned}$$

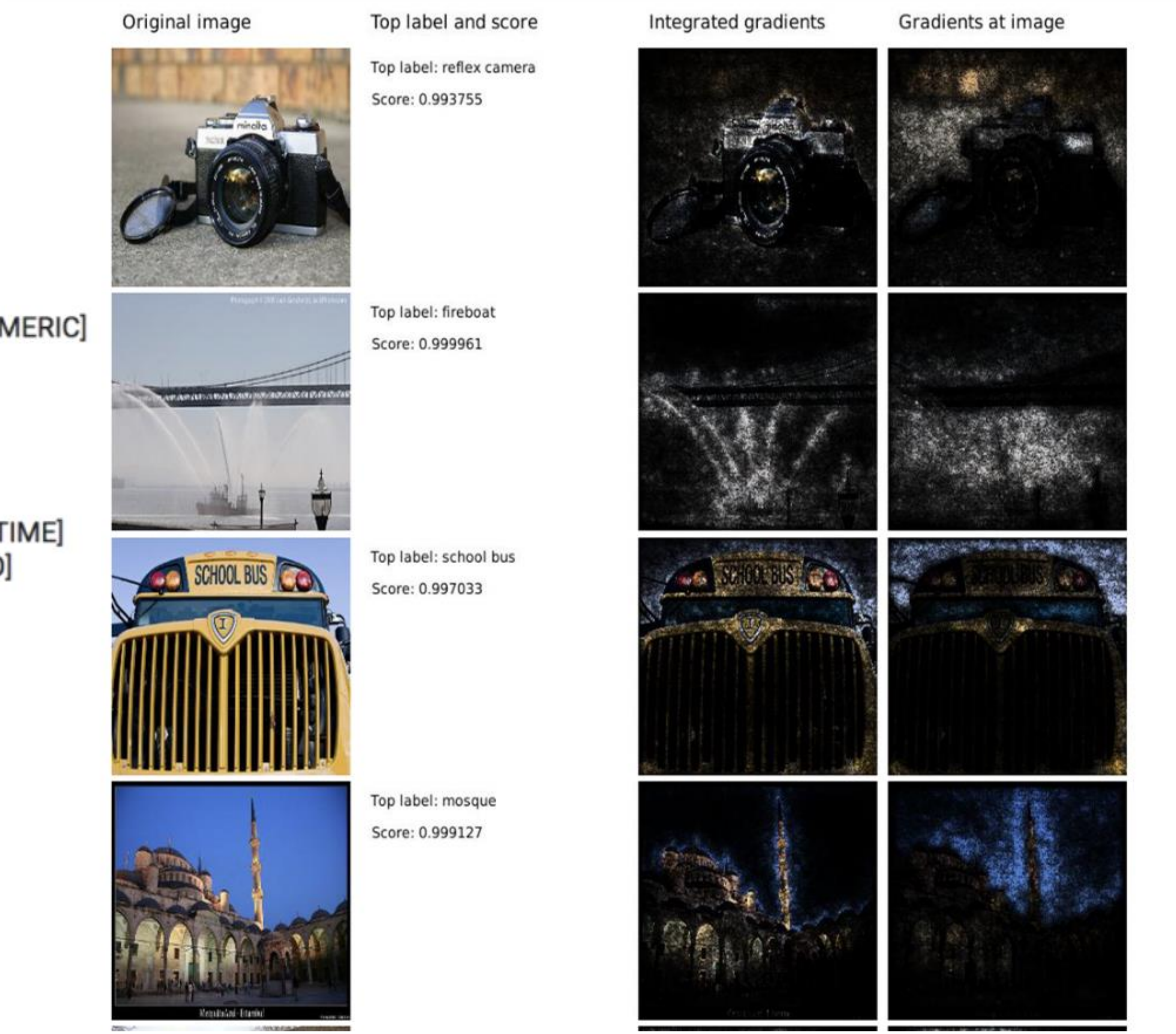

Accumulated Interesting Gradients
Less Interesting Gradients at Current Prediction
Model Score
Baseline
Original Image

Interpretability or Explainability

Results



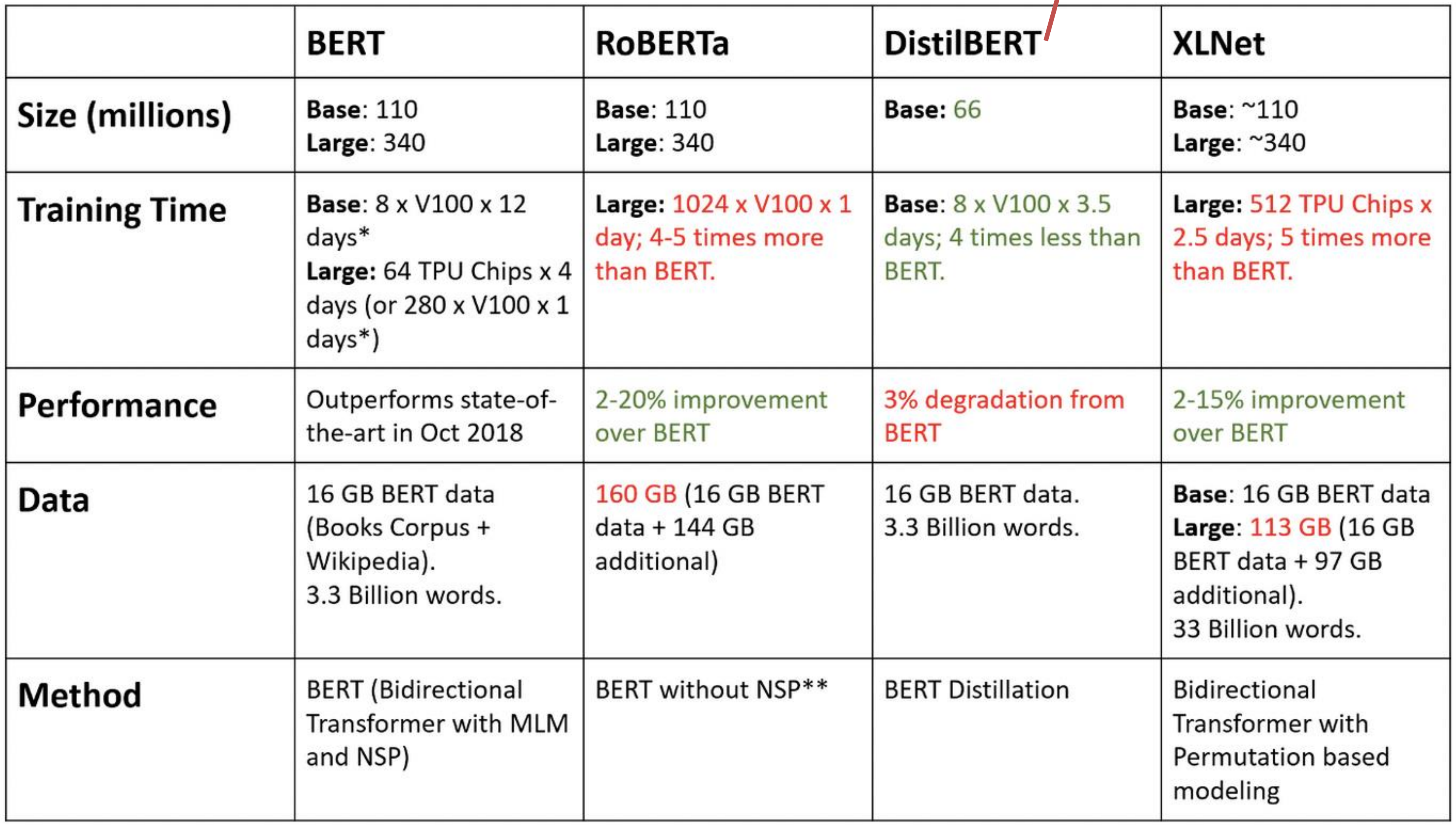
how many townships have a population above 50 ? [prediction: NUMERIC]
what is the difference in population between fora and masilo [prediction: NUMERIC]
how many athletes are not ranked ? [prediction: NUMERIC]
what is the total number of points scored ? [prediction: NUMERIC]
which film was before the audacity of democracy ? [prediction: STRING]
which year did she work on the most films ? [prediction: DATETIME]
what year was the last school established ? [prediction: DATETIME]
when did ed sheeran get his first number one of the year ? [prediction: DATETIME]
did charles oakley play more minutes than robert parish ? [prediction: YESNO]



Sentiment Analysis

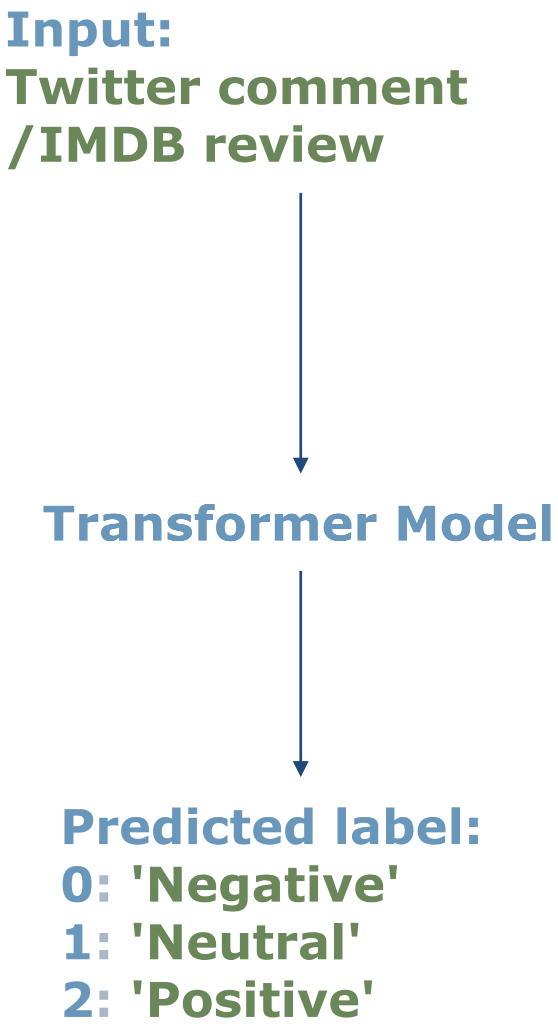
Base model Comparison

89% accuracy
IMDB test



	BERT	RoBERTa	DistilBERT	XLNet
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340	**Base**: 66	**Base**: ~110 **Large**: ~340
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.	**Base**: 8 x V100 x 3.5 days; 4 times less than BERT.	**Large:** 512 TPU Chips x 2.5 days; 5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	3% degradation from BERT	2-15% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.	**Base**: 16 GB BERT data **Large**: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation	Bidirectional Transformer with Permutation based modeling

Comparison of BERT and recent improvements over it



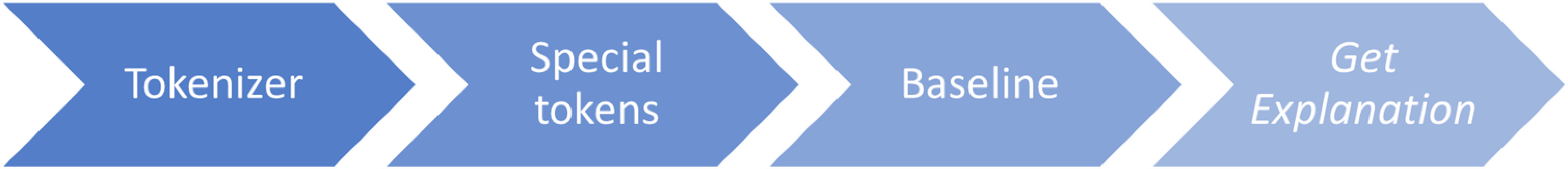
Input:
Twitter comment /IMDB review

Transformer Model

Predicted label:
0: 'Negative'
1: 'Neutral'
2: 'Positive'

Interpretability or Explainability

Process



Tokenizer

- distilBERT fine tuning
- RoBERTa Pretrain

Special tokens

- replacing the regular tokens with the [PAD]
- [CLS], [SEP], [UNK], [PAD]

Baseline

- embedding layer - attribution for those tokens will be 0

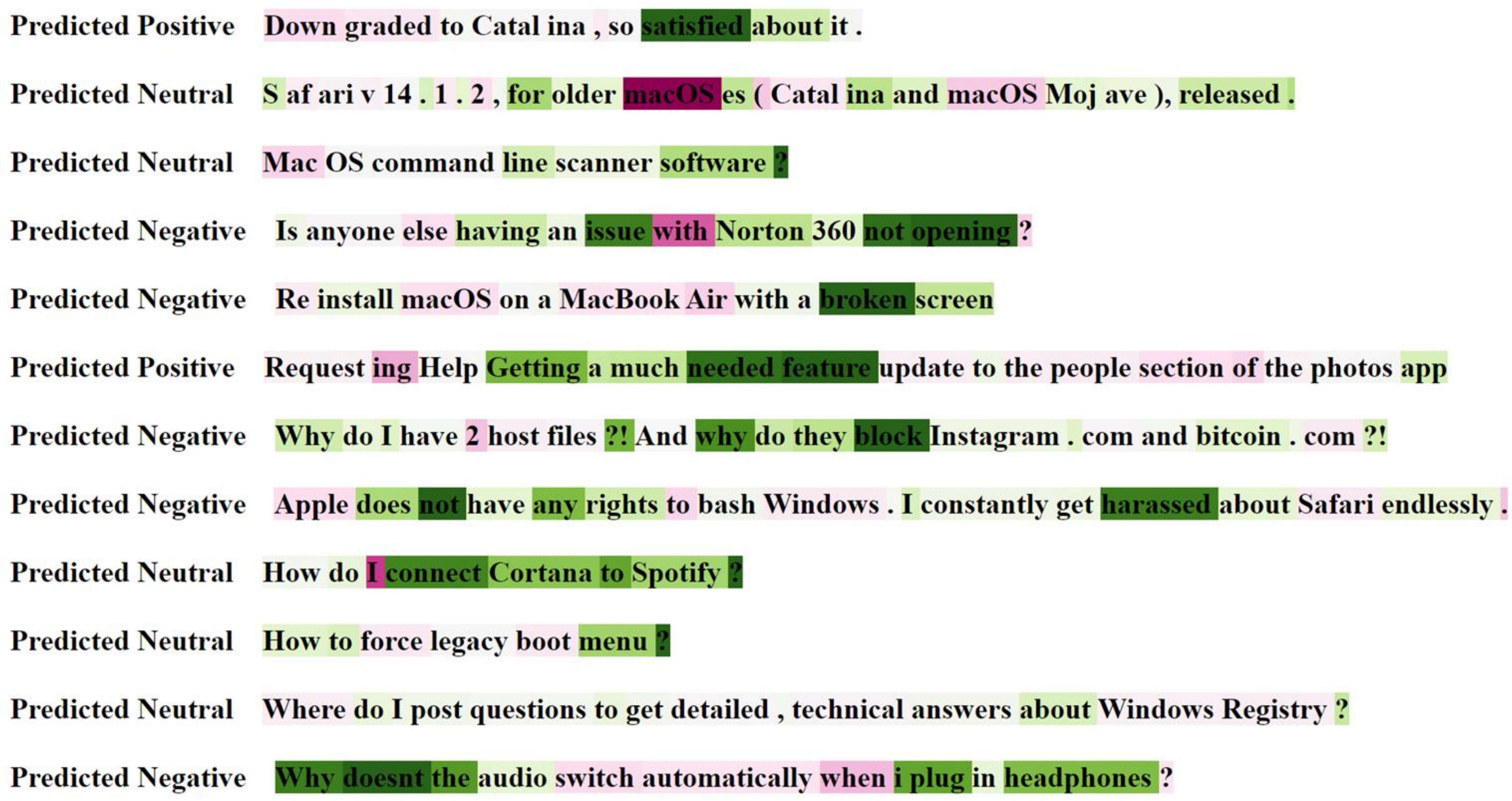
Get Explanation

- Output html with color

Interpretability or Explainability

Results

- Twitter-RoBERTa-base for Sentiment Analysis.



Predicted Positive	Down graded to Catal ina , so satisfied about it .
Predicted Neutral	S af ari v 14 . 1 . 2 , for older macOS es (Catal ina and macOS Moj ave), released .
Predicted Neutral	Mac OS command line scanner software ?
Predicted Negative	Is anyone else having an issue with Norton 360 not opening ?
Predicted Negative	Re install macOS on a MacBook Air with a broken screen
Predicted Positive	Request ing Help Getting a much needed feature update to the people section of the photos app
Predicted Negative	Why do I have 2 host files ?! And why do they block Instagram . com and bitcoin . com ?!
Predicted Negative	Apple does not have any rights to bash Windows . I constantly get harassed about Safari endlessly .
Predicted Neutral	How do I connect Cortana to Spotify ?
Predicted Neutral	How to force legacy boot menu ?
Predicted Neutral	Where do I post questions to get detailed , technical answers about Windows Registry ?
Predicted Negative	Why doesnt the audio switch automatically when i plug in headphones ?

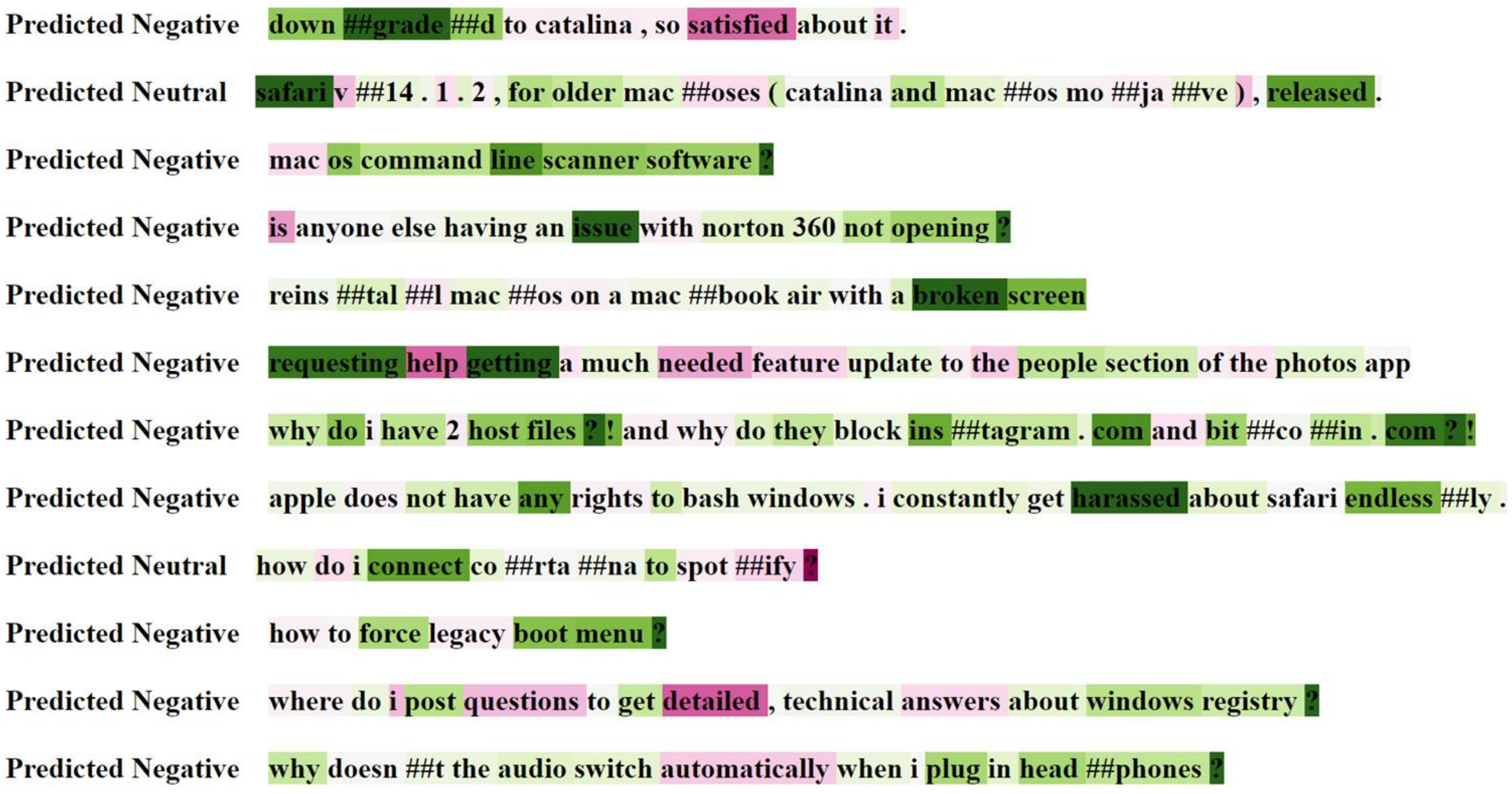
Interpretability or Explainability

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Results

- Based model: distilbert-base-uncased. Fine tuning on IMDB dataset



Predicted Negative down ##grade ##d to catalina , so satisfied about it .

Predicted Neutral safari v ##14 . 1 . 2 , for older mac ##oses (catalina and mac ##os mo ##ja ##ve) , released .

Predicted Negative mac os command line scanner software ?

Predicted Negative is anyone else having an issue with norton 360 not opening ?

Predicted Negative reins ##tal ##l mac ##os on a mac ##book air with a broken screen

Predicted Negative requesting help getting a much needed feature update to the people section of the photos app

Predicted Negative why do i have 2 host files ? ! and why do they block ins ##tagram . com and bit ##co ##in . com ? !

Predicted Negative apple does not have any rights to bash windows . i constantly get harassed about safari endless ##ly .

Predicted Neutral how do i connect co ##rta ##na to spot ##ify ?

Predicted Negative how to force legacy boot menu ?

Predicted Negative where do i post questions to get detailed , technical answers about windows registry ?

Predicted Negative why doesn ##t the audio switch automatically when i plug in head ##phones ?

Limitation and Future work

- Tokenization method did not match
- Training for DistilBert issue:only 1 epoch
- Training set different
- Methods for the integral approximation: riemann_left, riemann_right, riemann_middle, riemann_trapezoid, gausslegendre.
- A deeper study of the IG results.

THE GEORGE
WASHINGTON
UNIVERSITY

WASHINGTON, DC

Q&A?