Annotation Project 4
INFO259
Group 1
Jingshu Rui, Yang Yang, Tianhao Wu
April 21, 2022

Part A: Build Predictive Models

Baseline Results of binary classification models

Model	Test accuracy for best dev model	95% CIs
MajorityClass	0.77	/
Logistic Regression	0.765	[0.706 0.824]
Logistic Regression(improved)	0.790	[0.734 0.846]
BERT (Results differ each time)	0.800	[0.745 0.855]

Based on the notebooks provided by the teaching staff, we tried to improve the baseline model's test accuracy by incorporating knowledge from our guidelines to empower features in the logistic regression model and using more advanced architecture like the Bert model.

Many new features proved to be useless or even do harm to the test accuracy, like counting the number of occurrences of capital letters in a title, the title including 'not clickbait' is a clickbait, change bow from occurrence to word count. However, some superior knowledge contributes to constructing effective features. For example, the feature that superlative adjectives and adverbs probably lead to a clickbait helps to increase accuracy from 0.765 to 0.79. Because our data is to some extent imbalanced(the majority class takes up 77%), we also tried to overcome this by adjusting the hyperparameter class weight of logistic regression to 'balanced' to penalize more for predicting the majority class. But the effect was not good as test accuracy decreased.

BERT model didn't improve performance a lot. Actually only a little above the results we got from the improved version of logistic regression + bag of words. It's likely that the upper limit of this task is caused by some inconsistency in human labeling. We tried to avoid it as much as possible, but sometimes subjectivity still exists. For instance, titles

starting with ‘most viewed XXX’ can be classified as clickbait or non-clickbait depending on the content or the annotator’s interests. It is also a situation that we can add to our guidelines.

All the revises of the models can be found in the Jupyter Notebooks and PDFs we hand in.

Part B: Analysis

Following are the feature weights results generated from logistic regression using a binary bag of words as features:

1	2.183	most
1	2.157	world
1	1.932	best
1	1.641	debunk
1	1.538	10
1	1.310	top
1	1.175	$
1	1.142	this
1	1.117	moments
1	1.061	universe
0	-1.306	from
0	-1.048	|
0	-0.914	building
0	-0.832	bbc
0	-0.829	comedy
0	-0.777	review
0	-0.711	vs
0	-0.706	probably
0	-0.701	with
0	-0.696	my

Impactful positive weighted words such as *most, world, best, top, universe, 10, $,* are expected according to our guideline. Such words would be flagged as highly suspicious of click-baiting according to the following rules listed: superlative adjectives and adverbs, extreme numbers such as money.
Word *“debunk”* does not fall under any rules from our guidelines, but we are considering adding it as a new rule —- mysteries / revealing mysteries.

For negative weighted words, some stop words and symbols such as *from,* |*, with, my* should be removed from features since they should not indicate whether the text is click-baiting or not. Words such as *review*, *probably, and bbc* are leaning towards objective statements in nature, and thus not click-baiting.

Since YouTube titles are relatively short text documents, 1000 annotations might be deficient in features to train and justify the model. To improve the modeling, we need more annotated data. Increasing data reduces overfitting toward some features, and might increase the accuracy of the model.

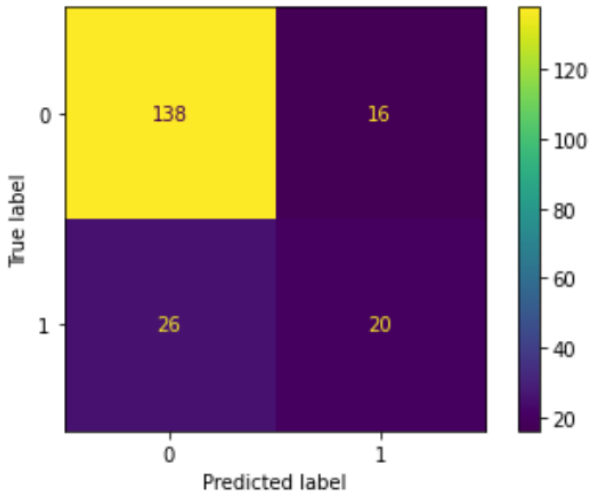
The top influential features obtained from the improved version of logreg model are listed below. The results are basically the same except the ‘est_count’ feature we construct. The positive related features make sense while the negative related ones don’t reveal any useful patterns. The word ‘debunk’ may be a strong evidence of the title to be a clickbait, which helps us to rethink the category boundaries.(660 features in total)

1	2.283	most
1	1.727	world
1	1.673	debunk
1	1.648	10
1	1.450	est_count
1	1.343	top
1	1.243	best
1	1.198	$
1	1.161	this
1	1.071	moments
0	-1.268	from
0	-0.917	|
0	-0.889	comedy
0	-0.859	building
0	-0.819	bbc
0	-0.818	review
0	-0.727	with
0	-0.701	my
0	-0.696	honest
0	-0.696	trailers

Our dataset has a moderate level of imbalance with 77% of the majority class. More appropriate scenario for using an oversampling or downsampling method should be a more extreme imbalance like 1:10. So we adjusted the class weight hyperparameter to achieve the same effect. The test accuracy decreased after applying this method.

Measure	Value	Derivations
Sensitivity	0.4348	TPR = TP / (TP + FN)
Specificity	0.8961	SPC = TN / (FP + TN)
Precision	0.5556	PPV = TP / (TP + FP)
Negative Predictive Value	0.8415	NPV = TN / (TN + FN)
False Positive Rate	0.1039	FPR = FP / (FP + TN)
False Discovery Rate	0.4444	FDR = FP / (FP + TP)
False Negative Rate	0.5652	FNR = FN / (FN + TP)
Accuracy	0.7900	ACC = (TP + TN) / (P + N)
F1 Score	0.4878	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	0.3624	TP*TN - FP*FN / sqrt((TP+FP)* (TP+FN)*(TN+FP)*(TN+FN))

Printing out the confusion matrix, we can see that the data is pretty unbalanced, even if the accuracy of the prediction seems to be decent, the precision and recall are not very satisfying. This is mainly because the data are unbalanced, so there is considerable accuracy in just guessing the category of maximum probability.



Furthermore, we printed out the misclassified data as below. Examining the categories of those errors: out of the 16 false-positive error instances (type one error), there are 5 “blog”s, 3 “science”s, 3 “tech”s, 2 “food”s, 2 “videogames”s, 1 “informative”; and out of the 26 false-negative errors, there are 5 “food”s, 4 “videogames”s, 3 “automobile”s, 3 “blog”s, 3 “entertainment”s, 3 “science”s, 3 “tech”s, 2 “comedy”s, 2 “news”s (some video is multi-labeled, i.e. belong to more than one categories). BLOG is the category that has the most false positives and FOOD is the category that has the most false negatives. This is likely due to the different

distribution of click baits between categories in our dataset. Food and entertainment categories' clickbait rates are high because they're likely to be emotionally inflammatory or suspicious of the use of extreme rhetoric titles. While categories like blog, science and tech are inherently serious, knowledge-dense, and factual, thus less likely to be clickbait. In addition, by investigating the content we found patterns for both errors. False-negative errors are made mainly because the model failed to learn the overgeneralized and over affirmative criteria, with "every", and "real" being the indicator. Also, 10 out of 26 types one error cases contain numbers, and the judgment of whether the number is an exaggeration is based on common sense and background knowledge which the model doesn't have. Almost all false positives, on the contrary, have question marks or exclamation marks, which might trick the model to think that it's a clickbait.

prediction: 0 truth: 1 (type II error: false-negative conclusion)		**prediction: 1 truth: 0 (type I error: false-positive conclusion)**	
Automobile	2021 Ford F-150 POWERBOOST Review - INCREDIBLE!	**Blog**	Jake Paul SWATTTED! & Dropped by Disney! #DramaAlert Jake Paul DOXXED Post Malone! H3H3!
Automobile	Living With A Renault Twizy: What It's REALLY Like	**Blog**	Japan's Biggest Gaming Obsession Explained | Pachinko
Automobile	How To Be A BMW Driver	**Blog**	Jake Paul says N-Word! #DramaAlert Logan Paul SUED! & Roasted by Maze Runner!
Blog	26 Traits Japanese Girls Want in a Guy	**Blog**	I Joined Team 10 So You Don't Have To
Blog,Comedy	The Real Mighty Thirsty	**Blog**	Logan Paul ROASTED by Dolan Twins! #DramaAlert RiceGum , PewDiePie , Jacksepticeye & Much More!
Blog,Entertainment	Going Through The Same Drive Thru 1,000 Times	**Food**	Chicken Soup Impresses Gordon! | Hell's Kitchen
Blog,Science	The Value of F*** YOU money | Joe Rogan and Lex Fridman	**Food,Entertainment**	Gabriel Iglesias Does Wrestling Trivia While Eating Spicy Wings | Hot Ones
Comedy,Entertainment	Kim & Kanye's Unborn Baby Makes A Run For It | CONAN on TBS	**Informative**	Floating City DEBUNK
Entertainment	$27 Cake Vs. $1,120 Cake	**Science**	The Edge of an Infinite Universe
Food	Penis Pesto Pizza Taste Test | FOOD FEARS	**Science**	If the Universe is expanding, where is the centre?
Food	How To Fillet Every Fish | Method Mastery | Epicurious	**Science**	Is TON-618 the Largest Black Hole in the Universe? [OOTW]
Food	Kids Try 100 Years of Brown Bag Lunches from 1900 to 2000	**Tech**	Can You Actually Game in 8K? (RTX 3090 Gameplay!)
Food	Every Way to Cook an Egg (59 Methods) | Bon Appétit	**Tech**	Is a $100 Game Console Worth It?
Food,Entertainment	Ken Jeong Performs a Physical While Eating Spicy Wings | Hot Ones	**Tech**	Cool Keyboards You May Have Never Heard Of!
News	Live PD: Top 6 Worst Liars | A&E	**VideoGames**	Try Not To Laugh Challenge #16
News	We Decoded The Nuclear Weapons At North Korea's Military Parades | Decoded	**VideoGames**	Try Not To Laugh Challenge #5
Science	5 REAL Possibilities for Interstellar Travel		
Science	Why Snatch Blocks are AWESOME (How Pulleys Work) - Smarter Every Day 228		
Science	Making 500,000 VOLT ARC with Marx Generator		
Tech	Double or Triple Your Internet Speed - This Method Actually Works!		

Tech,Comedy	Don't Be A Programmer		
Tech,News	HA! No One's Buying iPhones!		
VideoGames	I Built an Airport of Suffering Where Nobody Is Safe - SimAirport		
VideoGames	MASSIVE APEX LEGENDS PACKS OPENING (200+ PACKS)		
VideoGames	PUBG IN VR!!!		
VideoGames	WOOF		

Bert model's performance is not significantly improved compared with Logistic Regression, probably because there's very limited data and the 12 layer attention model is too complicated to train. Out of the 200 test cases, there're 24 true-positives, 131 true-negatives, 23 false-positive, and 22 false-negative predictions.

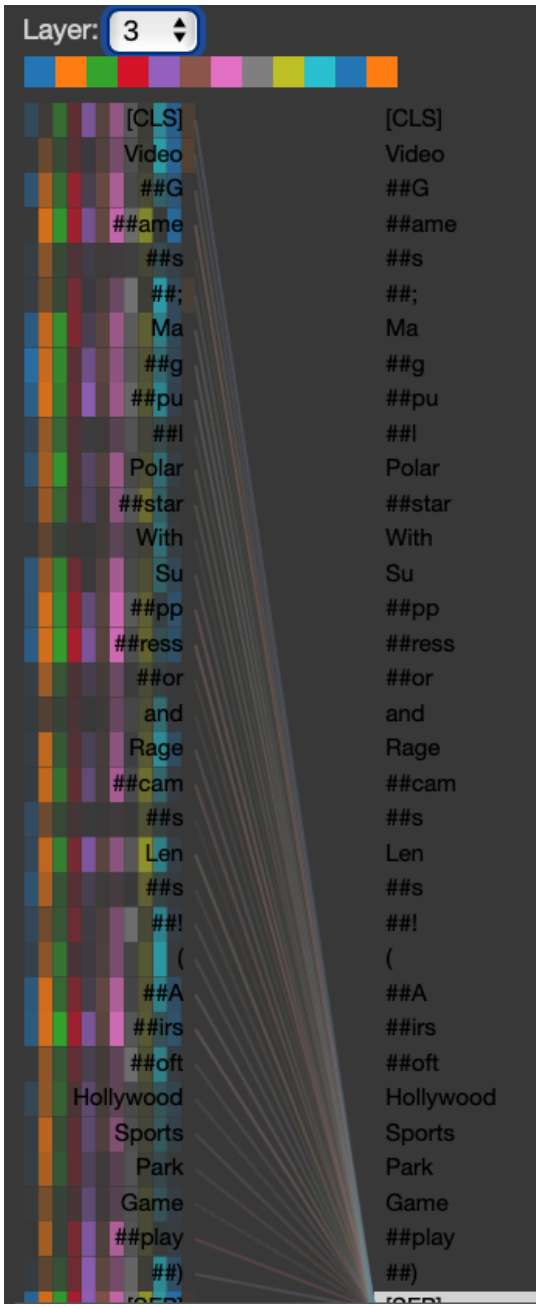
Measure	Value	Derivations
Sensitivity	0.6304	TPR = TP / (TP + FN)
Specificity	0.8377	SPC = TN / (FP + TN)
Precision	0.5370	PPV = TP / (TP + FP)
Negative Predictive Value	0.8836	NPV = TN / (TN + FN)
False Positive Rate	0.1623	FPR = FP / (FP + TN)
False Discovery Rate	0.4630	FDR = FP / (FP + TP)
False Negative Rate	0.3696	FNR = FN / (FN + TP)
Accuracy	0.7900	ACC = (TP + TN) / (P + N)
F1 Score	0.5800	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	0.4437	TP*TN - FP*FN / sqrt((TP+FP)* (TP+FN)*(TN+FP)*(TN+FN))

Analyzing the component of errors, BERT makes similar errors as the logistic regression model. This also justifies our previous conjecture. The data that caused the two models to make errors overlapped in a significant portion, indicating that these titles themselves were the edge cases or were difficult for the machine to identify.

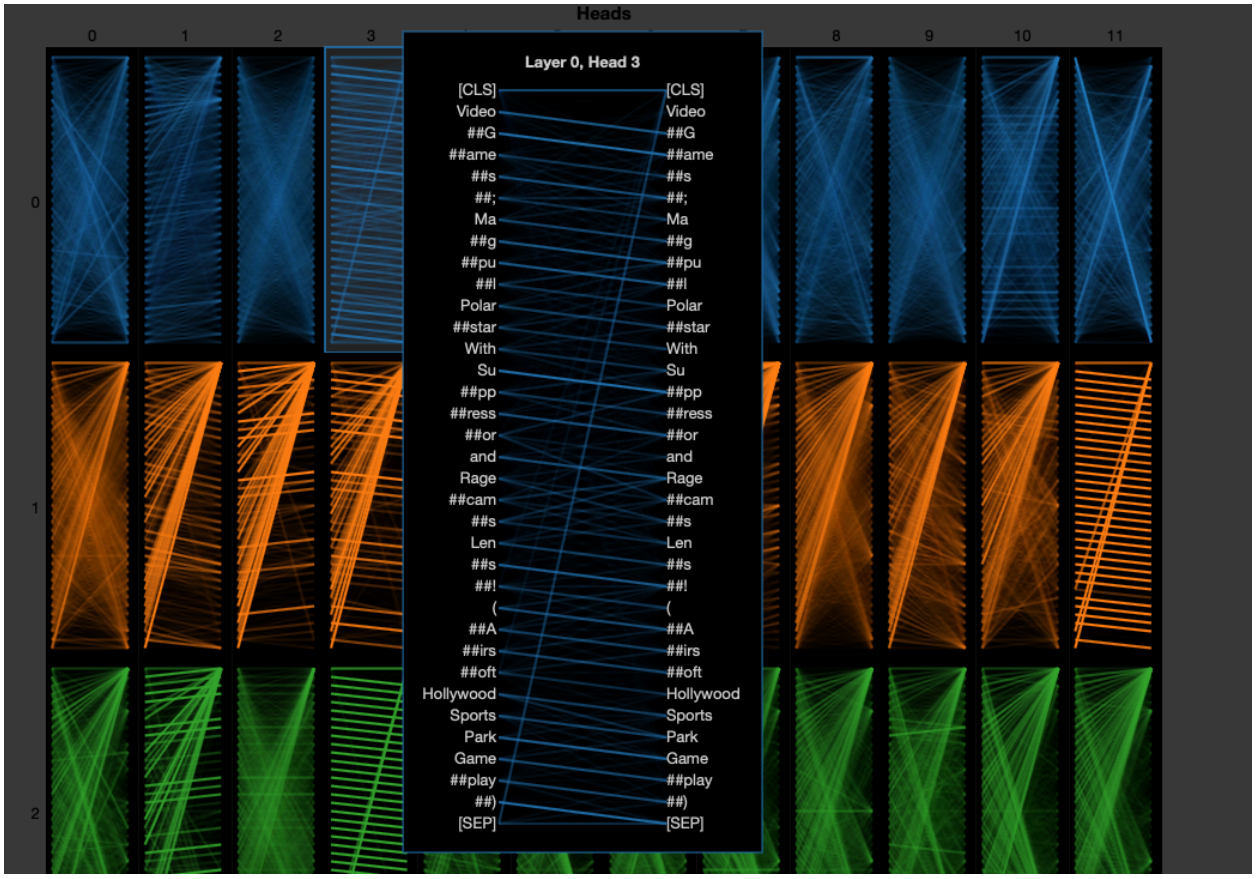
prediction: 0 truth: 1 (type II error: false-negative conclusion)		**prediction: 1 truth: 0 (type I error: false-positive conclusion)**	
Automobile	2021 Ford F-150 POWERBOOST Review - INCREDIBLE!	**Blog**	I Joined Team 10 So You Don't Have To
Automobile	Living With A Renault Twizy: What It's REALLY Like	**Blog**	Logan Paul ROASTED by Dolan Twins! #DramaAlert RiceGum , PewDiePie , Jacksepticeye & Much More!

Automobile	How To Be A BMW Driver	**Blog**	Japan's Biggest Gaming Obsession Explained | Pachinko
Blog	I Try EVERY Japanese Ramen	**Blog**	Jake Paul SWATTTED! & Dropped by Disney! #DramaAlert Jake Paul DOXXED Post Malone! H3H3!
Blog,Science	The Value of F*** YOU money | Joe Rogan and Lex Fridman	**Food**	The Diet of a Champion Female Bodybuilder
Comedy,Entertainment	Kim & Kanye's Unborn Baby Makes A Run For It | CONAN on TBS	**Food**	I Made This 96-Hour Ox Tongue Stew From Worth It
Food	Penis Pesto Pizza Taste Test | FOOD FEARS	**Food**	How a Master Chef Runs the Only Las Vegas Restaurant Awarded 3 Michelin Stars — Mise En Place
Food	Every Way to Cook an Egg (59 Methods) | Bon Appétit	**Food**	$379 McDonald's Filet-O-Fish Taste Test | FANCY FAST FOOD
Food	How To Fillet Every Fish | Method Mastery | Epicurious	**Informative**	Guilty until proven innocent.
Food	Kids Try 100 Years of Brown Bag Lunches from 1900 to 2000	**Informative**	LG OLED TV rolls up like a piece of paper
Food,Entertainment	Ken Jeong Performs a Physical While Eating Spicy Wings | Hot Ones	**Informative**	Floating City DEBUNK
Informative	The Brazen Bull (Worst Punishment in the History of Mankind)	**Informative**	Captain D's Definitive Guide to TRICK SHOTS
News	We Decoded The Nuclear Weapons At North Korea's Military Parades | Decoded	**Science**	Trying to Catch a 1,000 MPH Baseball - Smarter Every Day 247
Science	Why Snatch Blocks are AWESOME (How Pulleys Work) - Smarter Every Day 228	**Science**	How Safe Is the SHOWER HEAD OF DOOM?!
Science	5 REAL Possibilities for Interstellar Travel	**Science**	Why Alien Life Would be our Doom - The Great Filter
Science	Making 500,000 VOLT ARC with Marx Generator	**Science**	Building a Marsbase is a Horrible Idea: Let's do it!
Tech	Double or Triple Your Internet Speed - This Method Actually Works!	**Science**	The Edge of an Infinite Universe
Tech	I Bought EVERY Console at GameStop...	**Science**	Is TON-618 the Largest Black Hole in the Universe? [OOTW]
Tech,Comedy	Don't Be A Programmer	**Science**	Boarding a US NAVY NUCLEAR SUBMARINE in the Arctic - Smarter Every Day 240
VideoGames	WOOF	**Tech**	Is a $100 Game Console Worth It?
VideoGames	PUBG IN VR!!!	**Tech**	It's All Gone
VideoGames	I Built an Airport of Suffering Where Nobody Is Safe - SimAirport	**Tech,News**	I'm ALMOST sad for Apple...
		VideoGames	USING AN EYE TRACKER IN PUBG

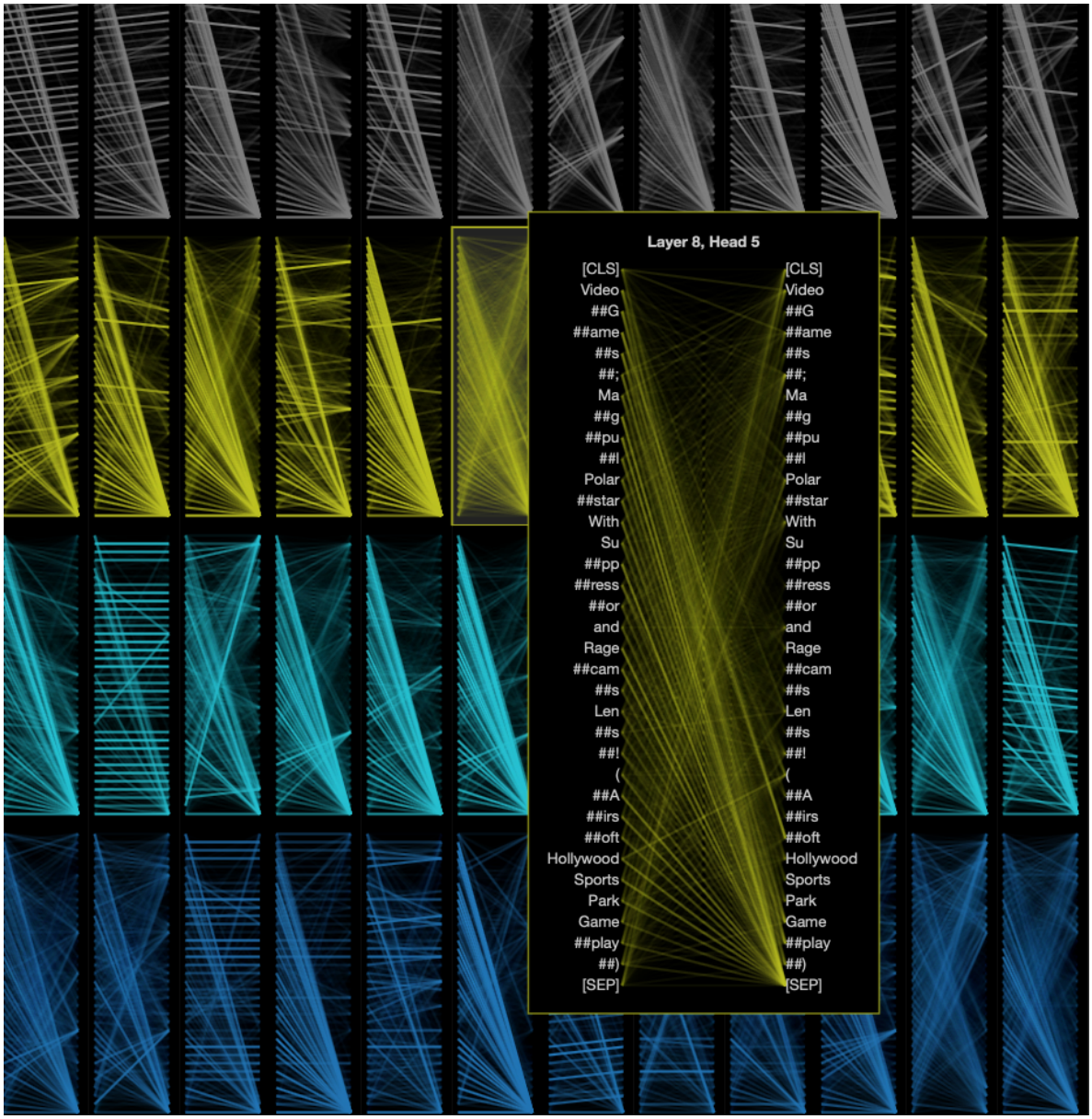
We also visualized the attention for the BERT model, see below for a screenshot for the "head_view" feature from the "bertviz" package. This view visualizes attention as lines connecting the word being updated on the left-hand side with the word being attended to on the right. The color intensity reflects the attention weight, dark lines suggest heavier weights and faint lines correspond to weights close to zero. Here, we highlight the [SEP] symbol, a special separator token that indicates a sentence boundary, to see the attention from it only.



The BERT model we use here has 12 layers and 12 heads, multiplying to a total of 144 distinct attention mechanisms. The “model_view” function in the “bertviz” package allows for us to visualize attention in all heads at once. Take the below image as an example, “Layer 0, Head 3” sub-cell shows the attention pattern for head 3 in layer 0. We can see that some specific word pairs have higher attention weights than others, especially the weight of attention tends to be significant between two consecutive tokens in order, which makes sense because adjacent words are often the most relevant for understanding a word’s meaning in context. This cell is a classic example of next-word attention patterns, almost all the attention is focused on the next word in the input sequence, except at the [SEP] and [CLS] tokens. But there are some word pairs that also have a somewhat significant reverse weight as well.



Another common pattern is the bag-of-words attention pattern, where attention is fairly evenly across all words. This pattern shows because BERT calculates a bag-of-words embedding by taking an almost unweighted average of the word embeddings.


Layer 8, Head 5
[CLS]
Video
##G
##ame
##s
##;
Ma
##g
##pu
##l
Polar
##star
With
Su
##pp
##ress
##or
and
Rage
##cam
##s
Len
##s
##!
(
##A
##irs
##oft
Hollywood
Sports
Park
Game
##play
##)
[SEP]
[CLS]
Video
##G
##ame
##s
##;
Ma
##g
##pu
##l
Polar
##star
With
Su
##pp
##ress
##or
and
Rage
##cam
##s
Len
##s
##!
(
##A
##irs
##oft
Hollywood
Sports
Park
Game
##play
##)
[SEP]