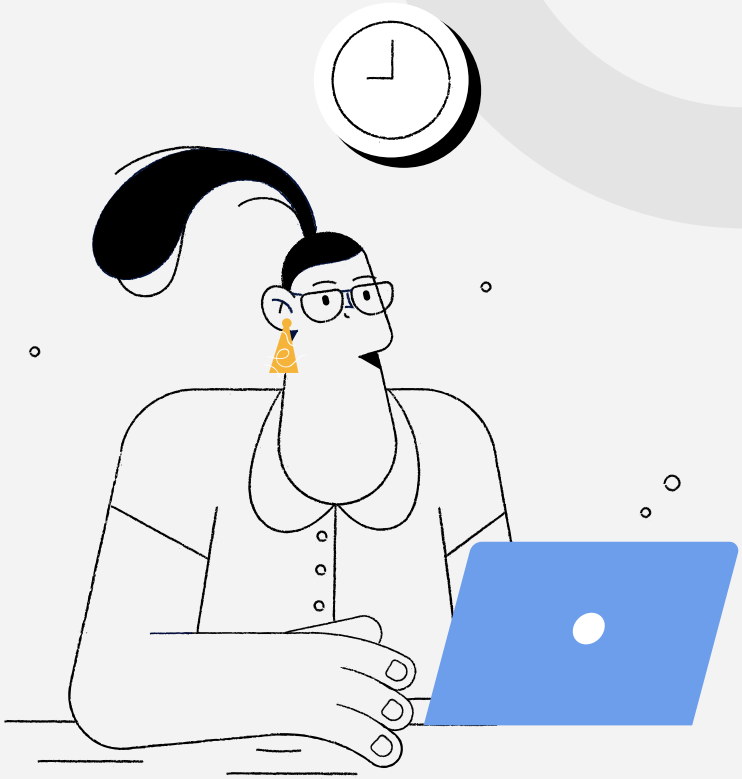
ROBOT WARS

Text Classification of Real
& Computer-Generated Amazon Reviews

Table of contents



01

Background

What’s the problem?

80%

Of U.S consumers indicate that they use online reviews before purchasing a product [1]

~50%

Of U.S consumers trust online reviews as much as recommendations from family & friends [2]

5-9%

Increase in revenue for an extra star on Yelp [3]

Sources

1. https://www.pewresearch.org/internet/2016/12/19/online-shopping-and-e-commerce/
2. https://www.brightlocal.com/research/local-consumer-review-survey/#in-summary
3. https://www.hbs.edu/faculty/Pages/item.aspx?num=41233

Paid Reviews

- Consumers being offered free products/monetary incentives to provide a favourable review

Computer Generation

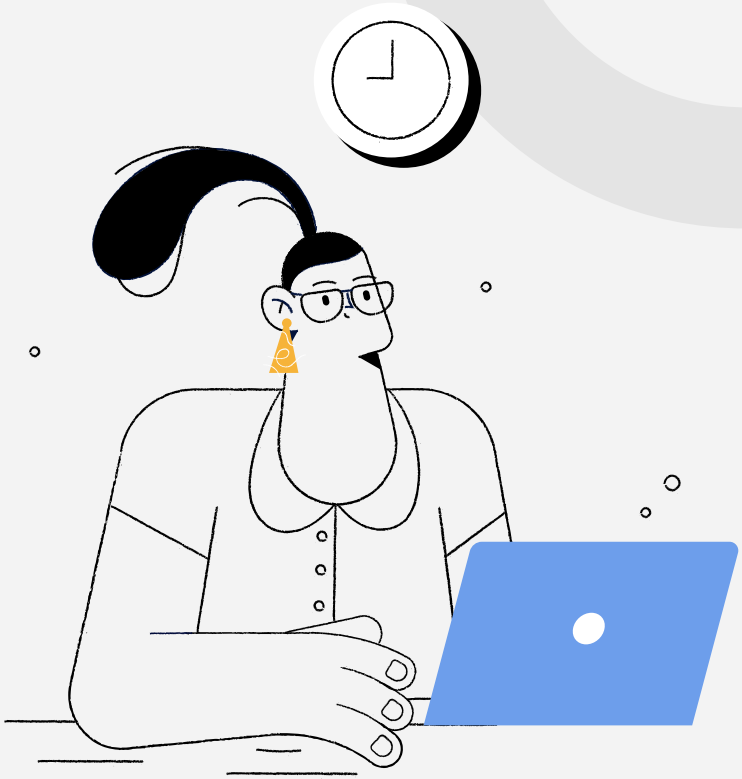
- Using text generation models to mimic human reviewers
- Can be generated at scale and at low cost
- Likely trajectory for fraudsters

52-65%

Efficacy of humans in detecting fake reviews [4]

Source: https://www.sciencedirect.com/science/article/pii/S0969698921003374?via%3Dihub

Can a machine beat a machine?



02

Methodology

How are we doing this?

Preprocessing

EDA

Modelling

40k Reviews
50% REAL
50% CG

Data Cleaning

- **Remove HTML tags**
- **Remove web addresses**
- **Remove symbols**
- **Remove numerical values**
- **Expanded contractions**

Preprocessing

EDA

Modelling

40k Reviews
50% REAL
50% CG

Data Cleaning

- Remove HTML tags
- Remove web addresses
- Remove symbols
- Remove numerical values
- Expanded contractions

Exploratory Data Analysis

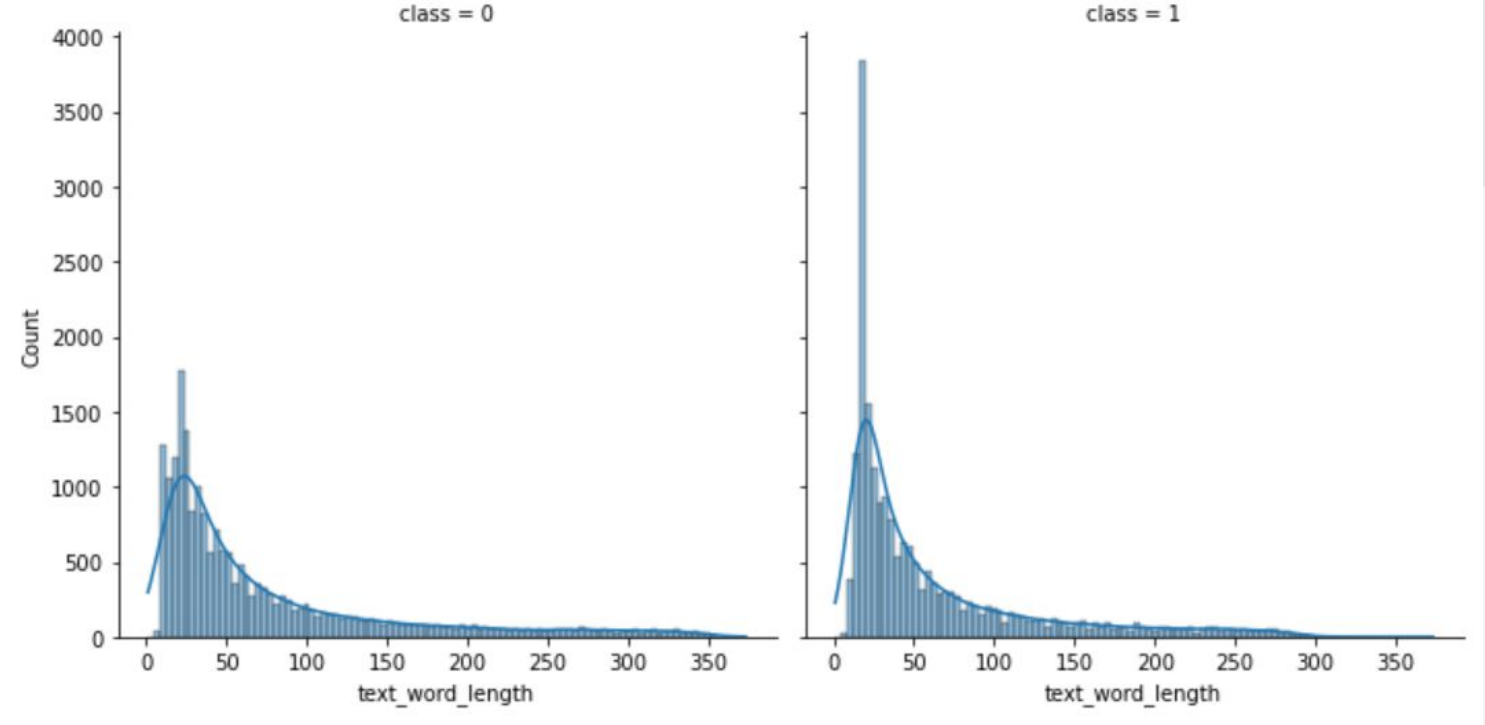
- Sentence Lengths
- Top occurring words
- Bigrams
- Class Distributions across columns

Preprocessing

EDA

Modelling

Text Word Length

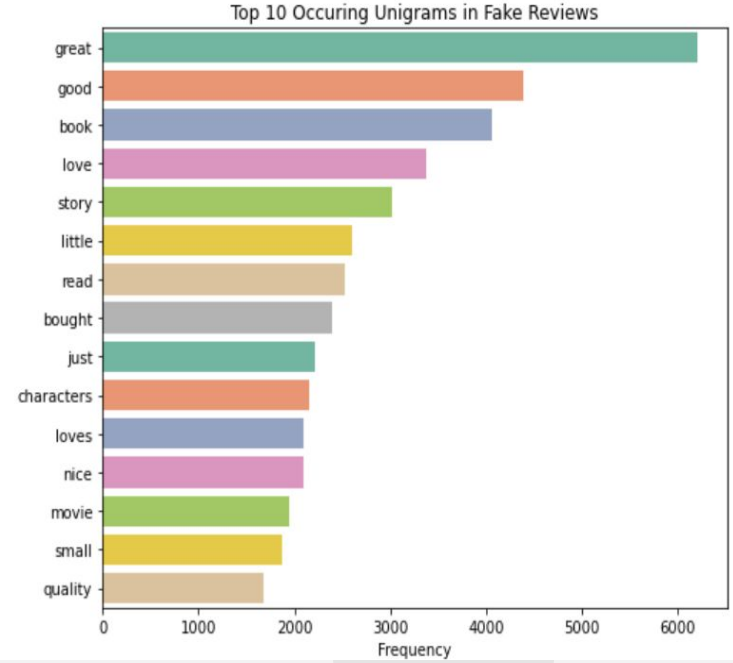


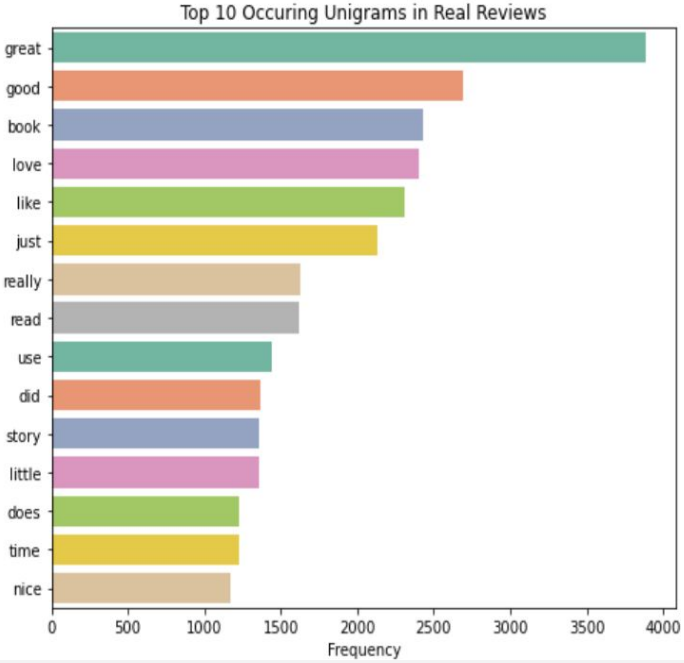
Preprocessing

EDA

Modelling

Unigrams Analysis



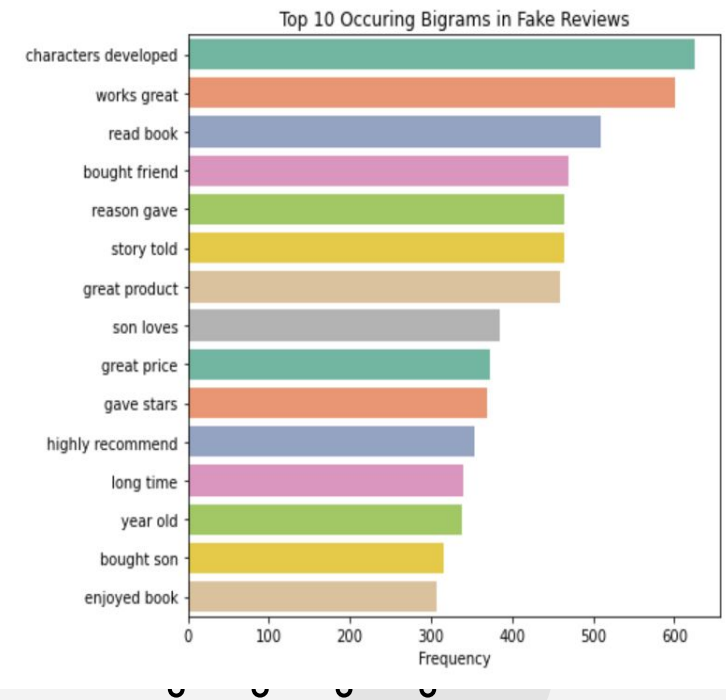


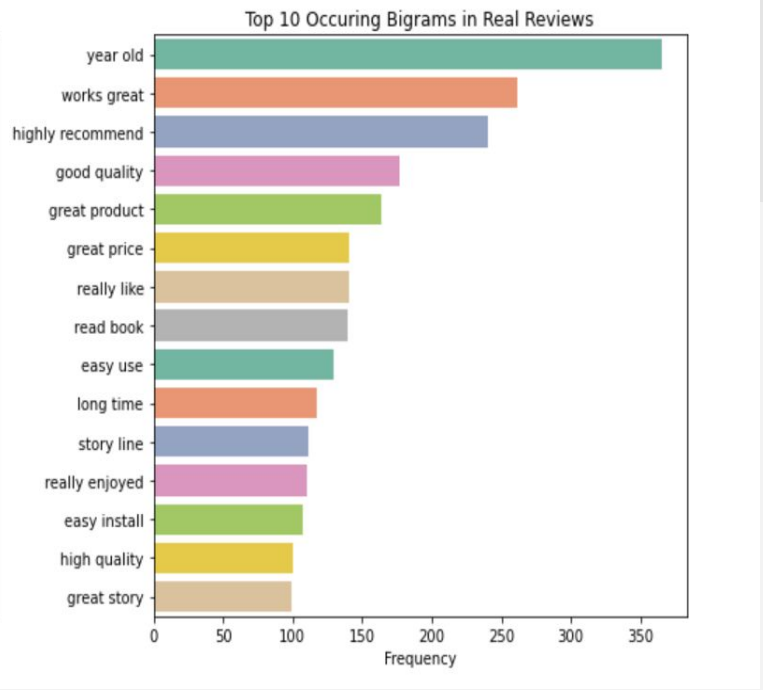
Preprocessing

EDA

Modelling

Bigrams Analysis



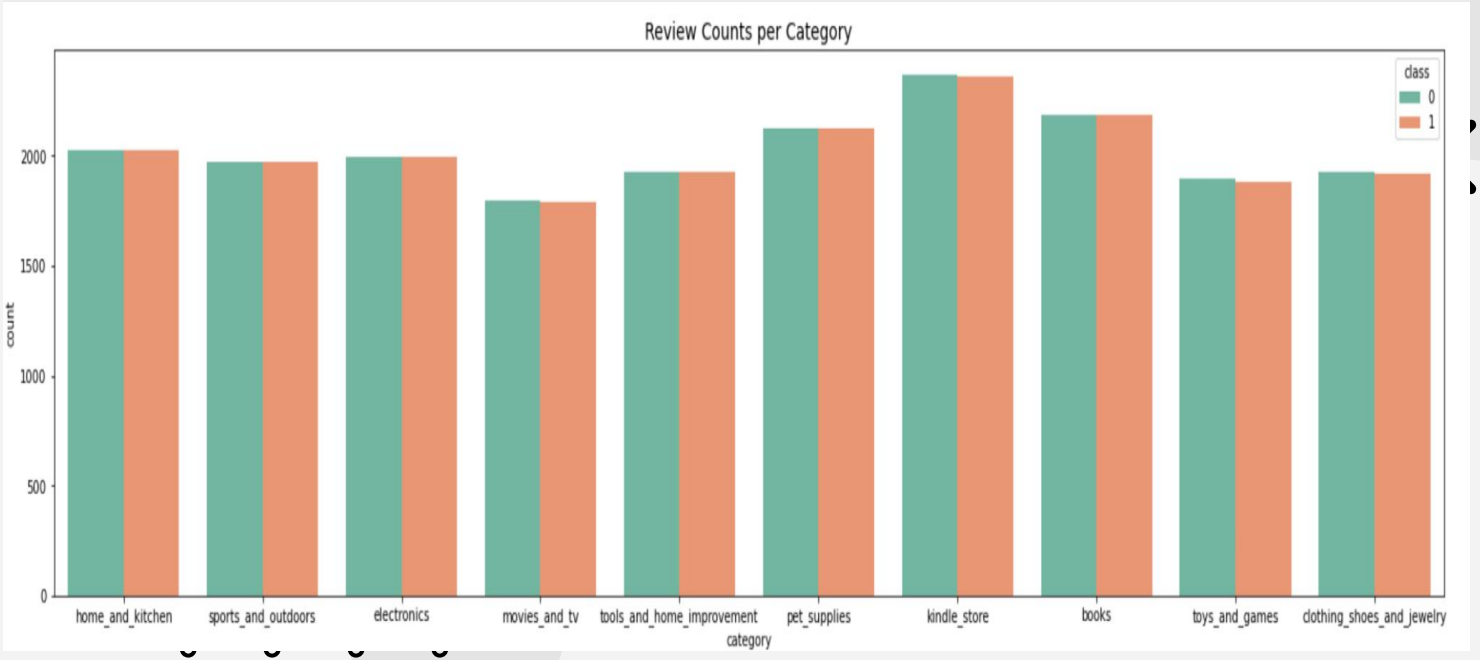


Preprocessing

EDA

Modelling

Class Distribution by Category



Preprocessing
EDA
Modelling

40k Reviews
50% REAL
50% CG

Data Cleaning

- Remove HTML tags
- Remove web addresses
- Remove symbols
- Remove numerical values
- Expanded contractions

Exploratory Data Analysis

- Sentence Lengths
- Top occurring words
- Bigrams
- Class Distributions across columns

Modelling

- Machine Learning Models
- Deep Learning Models

Preprocessing
EDA
Modelling

Modelling

- **Machine Learning Models**
- **Deep Learning Models**

Evaluation Metric:

F1 Score (>0.9): Weighted average of Precision and Recall

Machine Learning

BoW Approach

- CountVectorizer
- TF-IDF Vectorizer

Models

- Multinomial Naive Bayes
- Logistic Regression
- XGBoost
- Random Forest

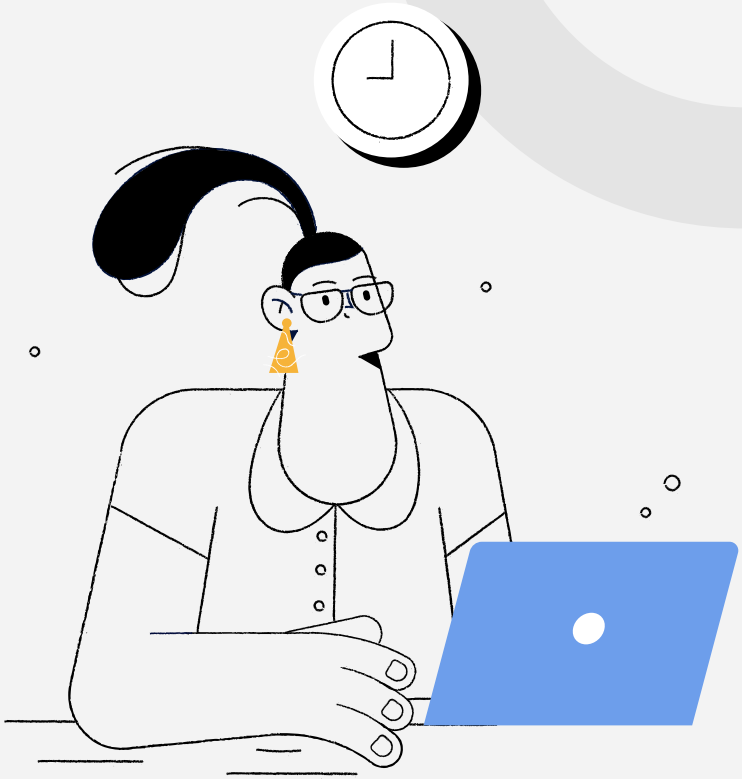
Deep Learning

Word Embeddings

- Model Embeddings
- Pre-Trained Embeddings

Models

- Bidirectional + GRU
- Bidirectional + LSTM
- DistilBERT Linear Classifier



03

NLP Primer

A quick revision of essential concepts

- BoW
- Word Embeddings
- DistilBERT

BoW: CountVectorizer

	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

Source: https://miro.medium.com/max/880/1*hLvya7MXjsSc3NS2SoLMEg.pngBag-of-Words

Bag-of-Words: TF-IDF Vectorizer

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

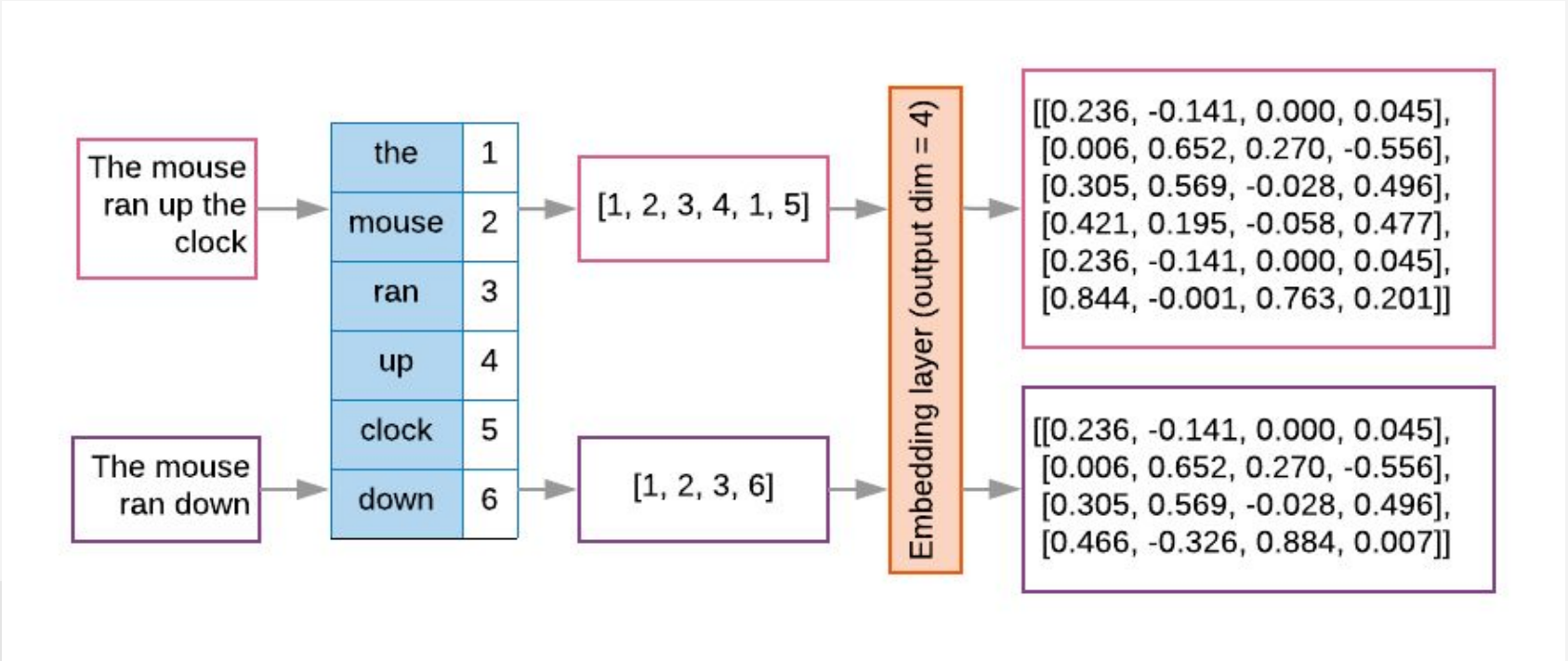
$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Source: https://miro.medium.com/max/1200/1*V9ac4hLVyms79jl65Ym_Bw.jpeg

Word Embeddings



Source: https://developers.google.com/machine-learning/guides/text-classification/images/EmbeddingLayer.png

Pre-Trained Word Embeddings

GloVe: Global Vectors for Vector Representation

- Trained on Wikipedia, 6B tokens, 400k vocab
- Dimension: 300

FastText: Global Vectors for Vector Representation

- Trained on Wikipedia, 16B tokens, 1M vocab

DistilBERT: Distilled Bidirectional Encoder Representation for Transformers (BERT)

- Trained on Wikipedia (2.5k million words) and Book Corpus (800 million words)
- Lightweight, 40% less parameters than BERT, runs 60% faster and retains 95% performance on benchmarks

Pre-Trained Word Embeddings

GloVe: Global Vectors for Vector Representation

- Trained on Wikipedia, 6B tokens, 400k vocab
- Dimension: 300

FastText: Global Vectors for Vector Representation

- Trained on Wikipedia, 16B tokens, 1M vocab

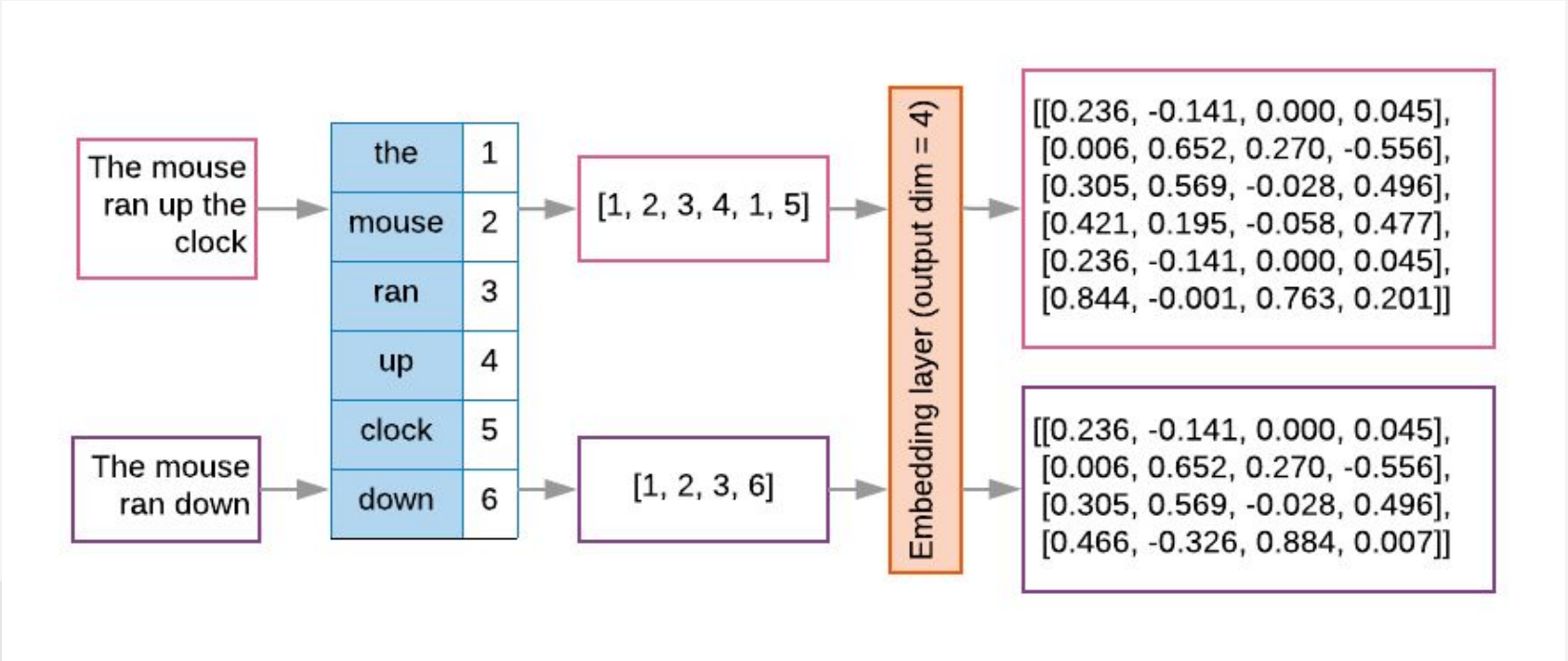
DistilBERT: Distilled Bidirectional Encoder Representation for Transformers (BERT)

- Trained on Wikipedia (2.5k million words) and Book Corpus (800 million words)
- Lightweight, 40% less parameters than BERT, runs 60% faster and retains 95% performance on benchmarks

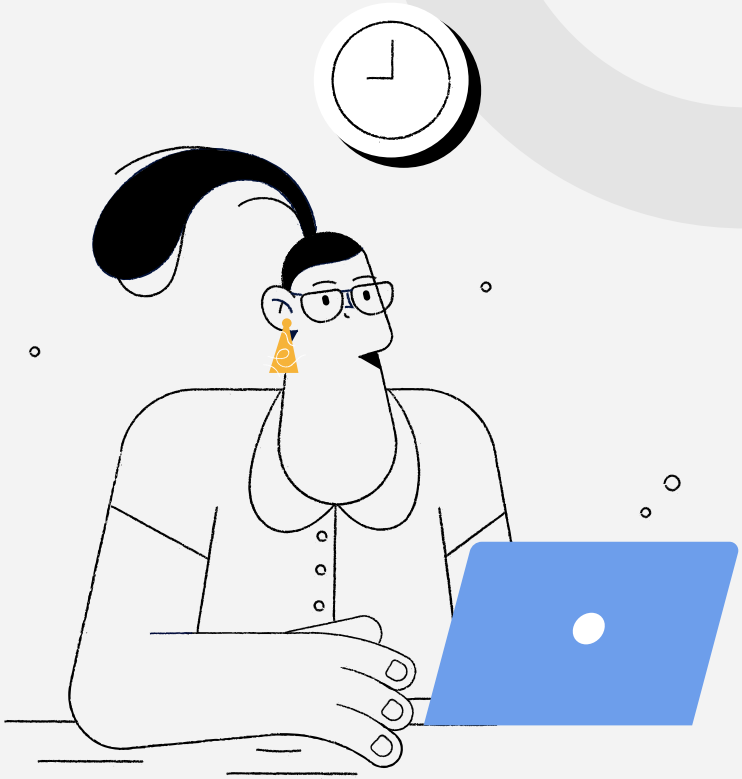
“ I robbed a bank”

“ I sat on the river bank”

Pre-Trained Word Embeddings



Source: https://developers.google.com/machine-learning/guides/text-classification/images/EmbeddingLayer.png



04

Results

What’s good?

Machine Learning Models (with Hyperparameter Tuning)

Model	Variants	Accuracy	F1 Score
Multinomial NB	CVEC	0.830	0.842
Logistic Regression	CVEC	0.889	0.890
	TVEC	0.902	0.891
XGBoost	CVEC	0.825	0.846
	TVEC	0.921	0.753
Random Forest	CVEC	0.869	0.805
	TVEC	0.832	0.820

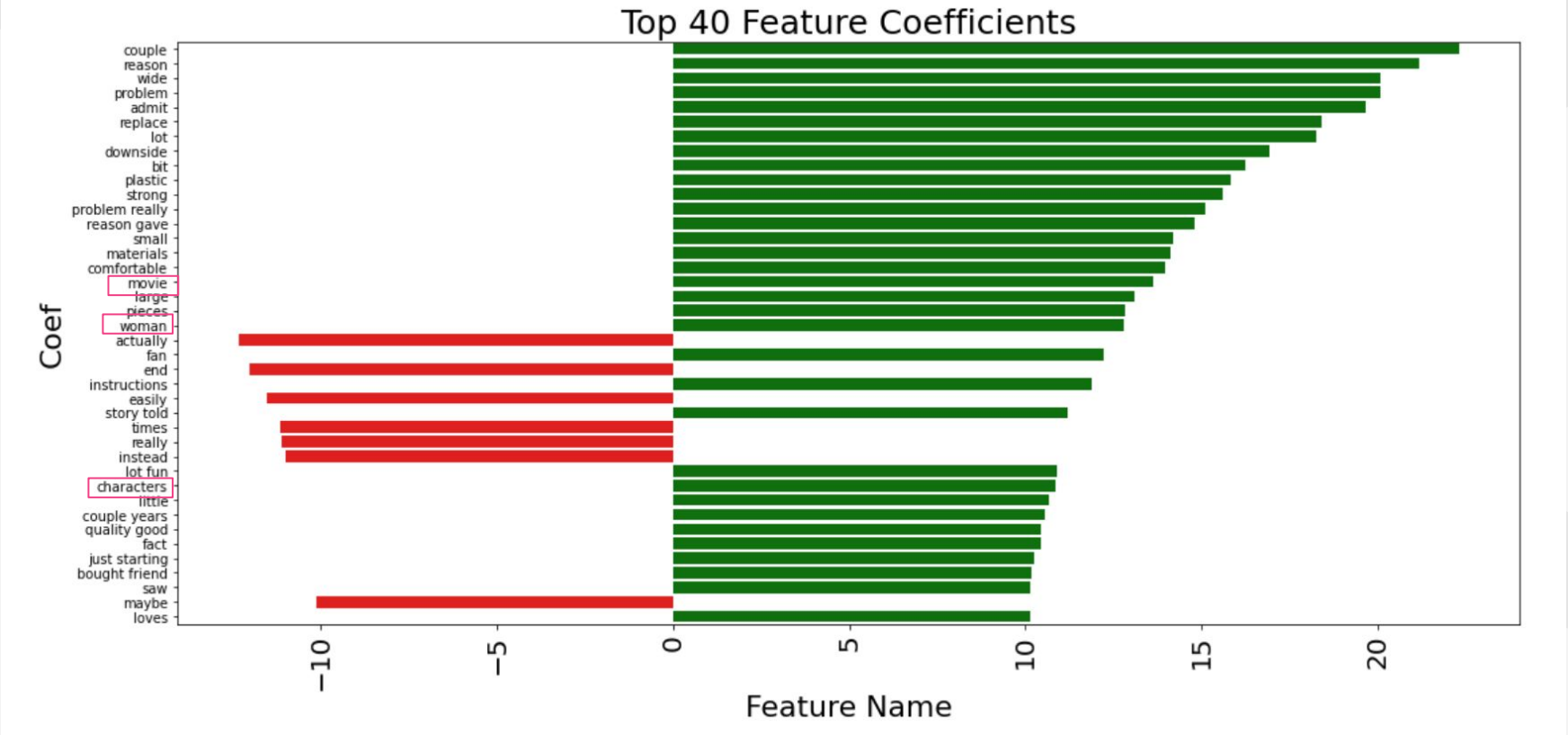
Machine Learning Models

Model	Variants	Accuracy	F1 Score
Multinomial NB	CVEC	0.830	0.842
GRU	Word Embedding	0.756	0.747
	GloVe Embedding	0.747	0.739
	FastText Embedding	0.745	0.736
Bidirectional GRU	Word Embedding	0.941	0.941
	GloVe Embedding	0.915	0.915
	FastText Embedding	0.926	0.926

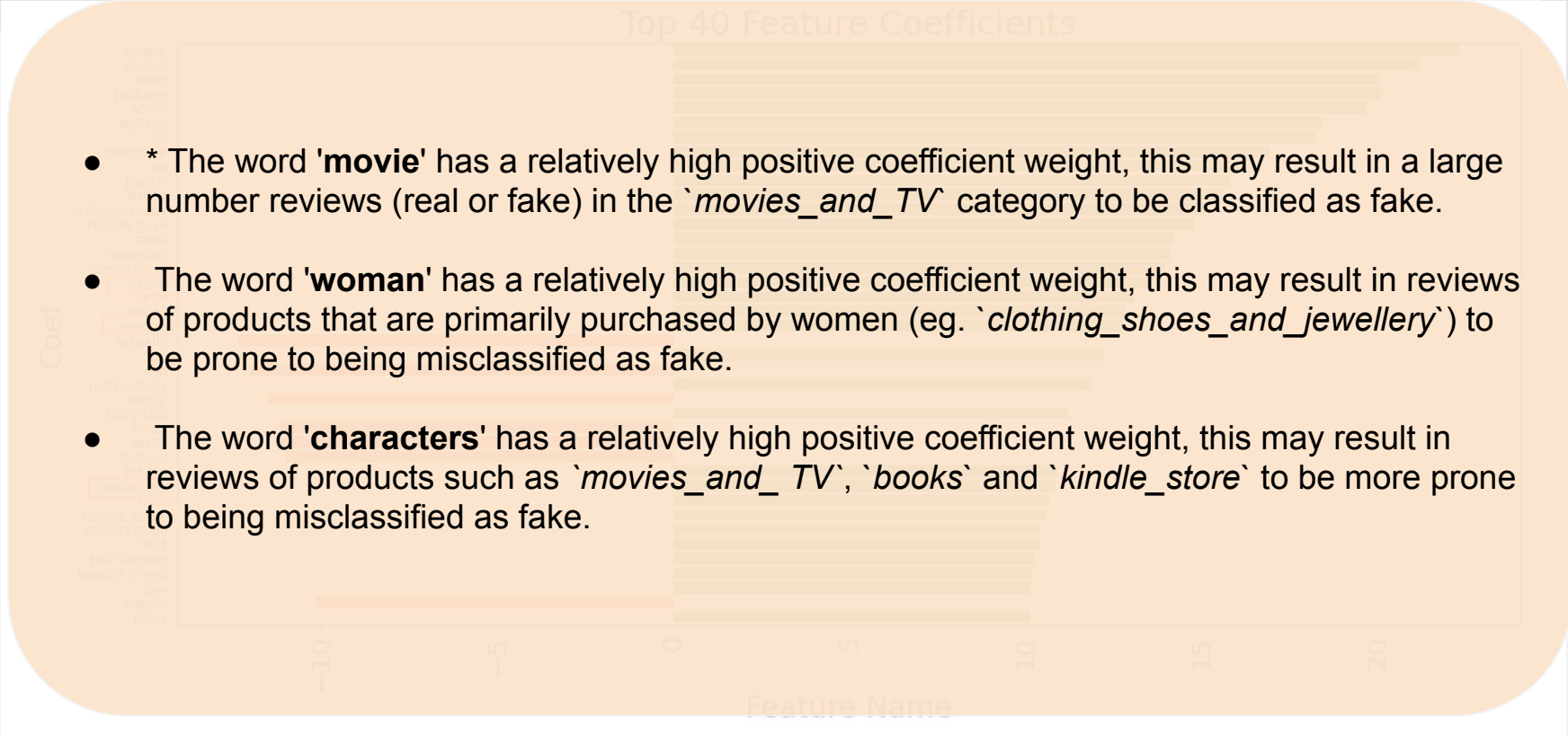
Machine Learning Models

Model	Variants	Accuracy	F1 Score
Multinomial NB	CVEC	0.830	0.842
LSTM	Word Embedding	0.758	0.750
	GloVe Embedding	0.750	0.740
	FastText Embedding	0.747	0.739
Bidirectional LSTM	Word Embedding	0.925	0.925
	GloVe Embedding	0.920	0.920
	FastText Embedding	0.923	0.923
DistilBERT Classifier	-	0.97	0.97

ML Observations:
Logistic Regression/ TF-IDF

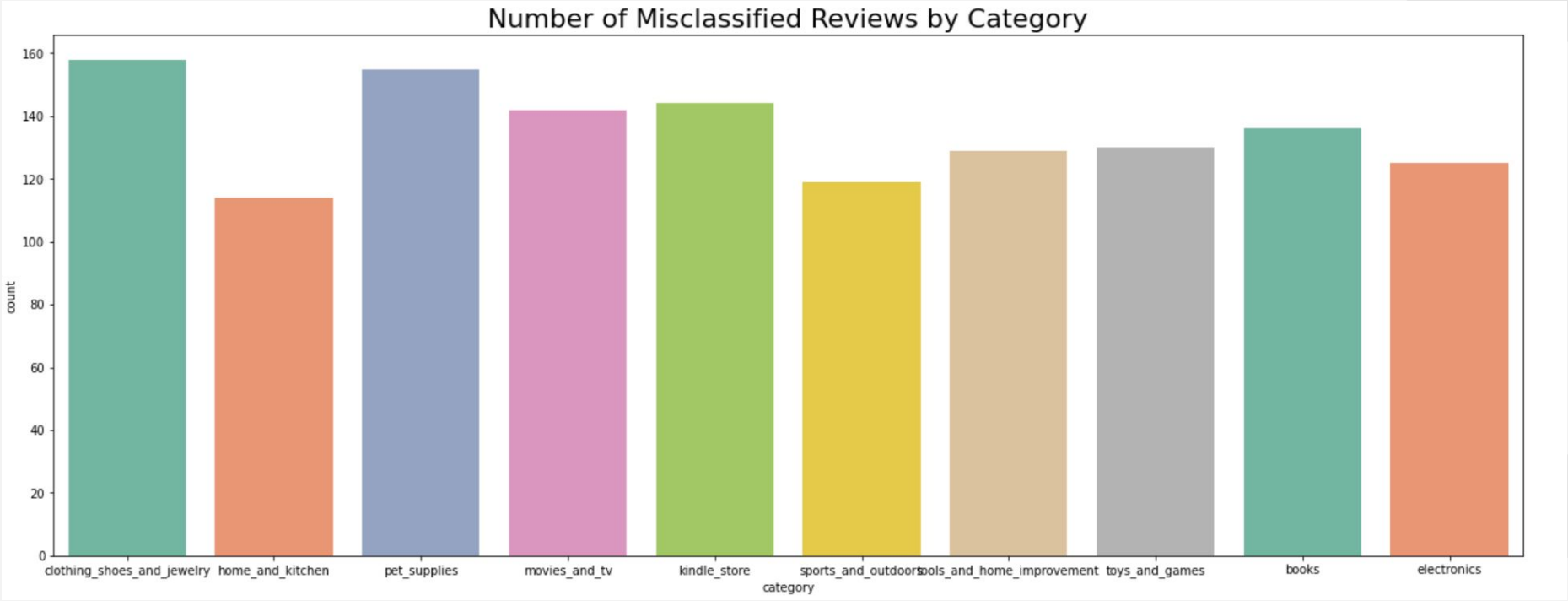


ML Observations:
Logistic Regression/ TF-IDF

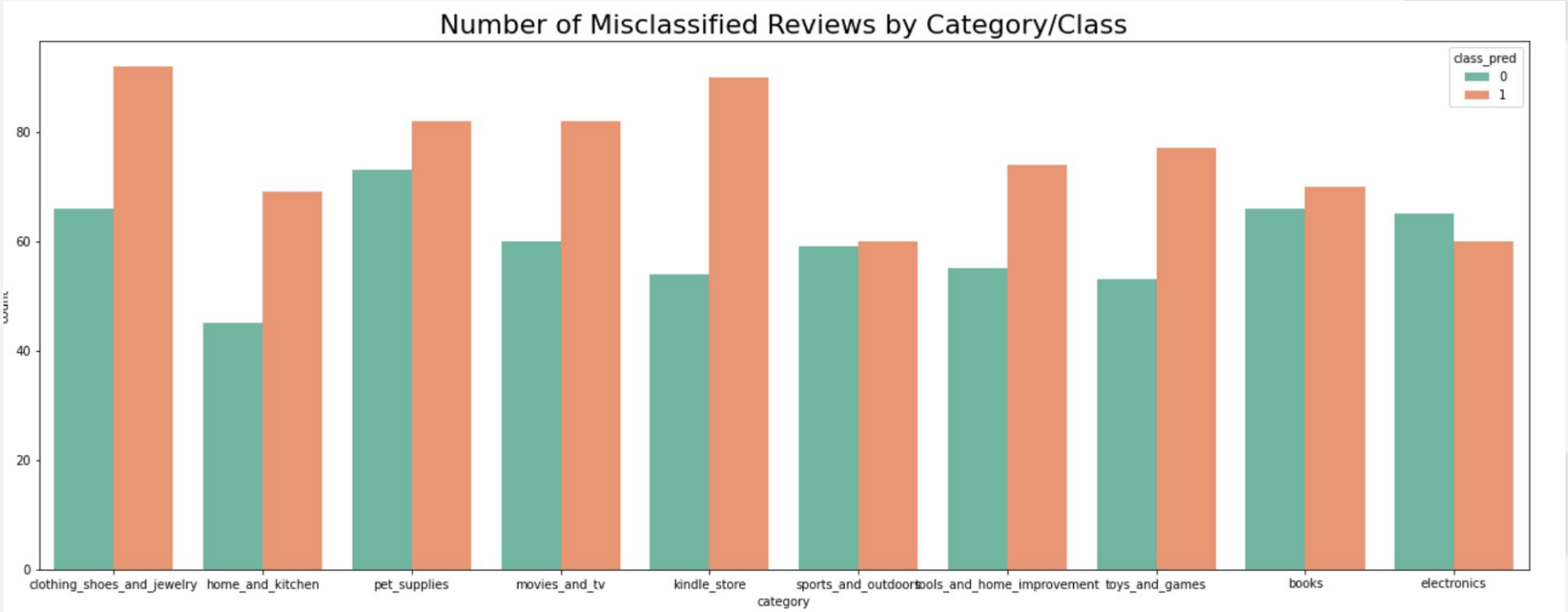


- * The word '**movie**' has a relatively high positive coefficient weight, this may result in a large number reviews (real or fake) in the `*movies_and_TV*` category to be classified as fake.
- The word '**woman**' has a relatively high positive coefficient weight, this may result in reviews of products that are primarily purchased by women (eg. `*clothing_shoes_and_jewellery*`) to be prone to being misclassified as fake.
- The word '**characters**' has a relatively high positive coefficient weight, this may result in reviews of products such as `*movies_and_ TV*`, `*books*` and `*kindle_store*` to be more prone to being misclassified as fake.

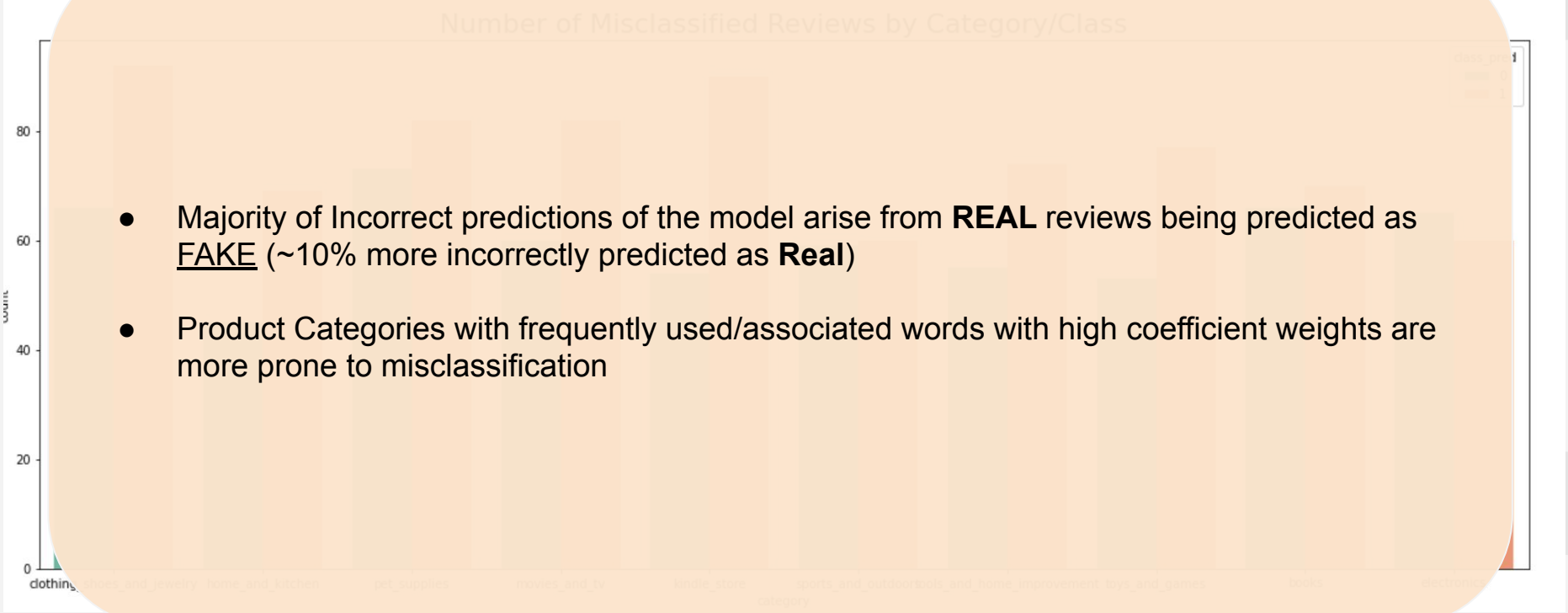
ML Observations:
Logistic Regression/ TF-IDF



ML Observations:
Logistic Regression/ TF-IDF



ML Observations:
Logistic Regression/ TF-IDF



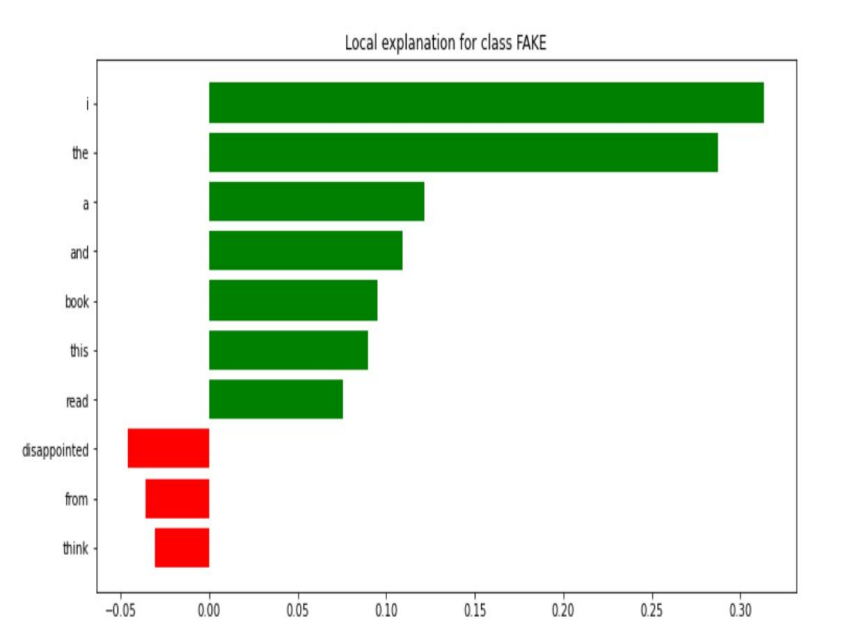
- Majority of Incorrect predictions of the model arise from **REAL** reviews being predicted as FAKE (~10% more incorrectly predicted as **Real**)
- Product Categories with frequently used/associated words with high coefficient weights are more prone to misclassification

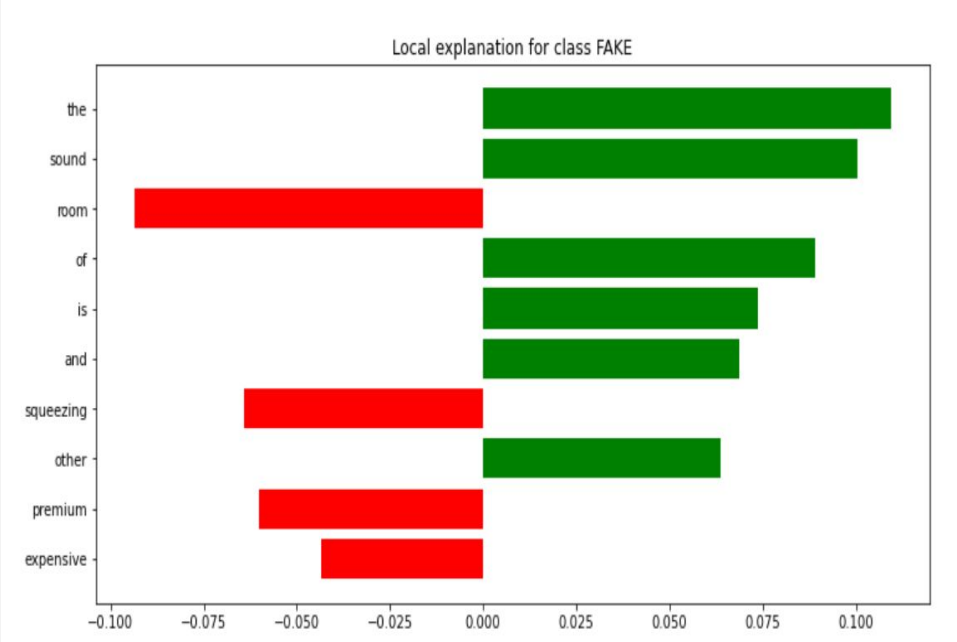
DL Observations:

- The Bidirectional GRU and LSTM have marked performance improvement over the one-directional versions (close to 20% improvement in both accuracy and F1 score). This means that having future information is important in the prediction of fake reviews.
- The performance of models using both GloVe and FastText achieved similar scores.
- As a whole the pre-trained embeddings do not perform better than the 'vanilla' word embeddings used in the model, a possible reason might be due to the number of OOV words which impeded performance.

Pre-Trained Embedding	Vocabulary Size	OOV Words
GloVe	400,000	471
FastText	999,994	884
DistilBERT	30,552	0

DL Observations:





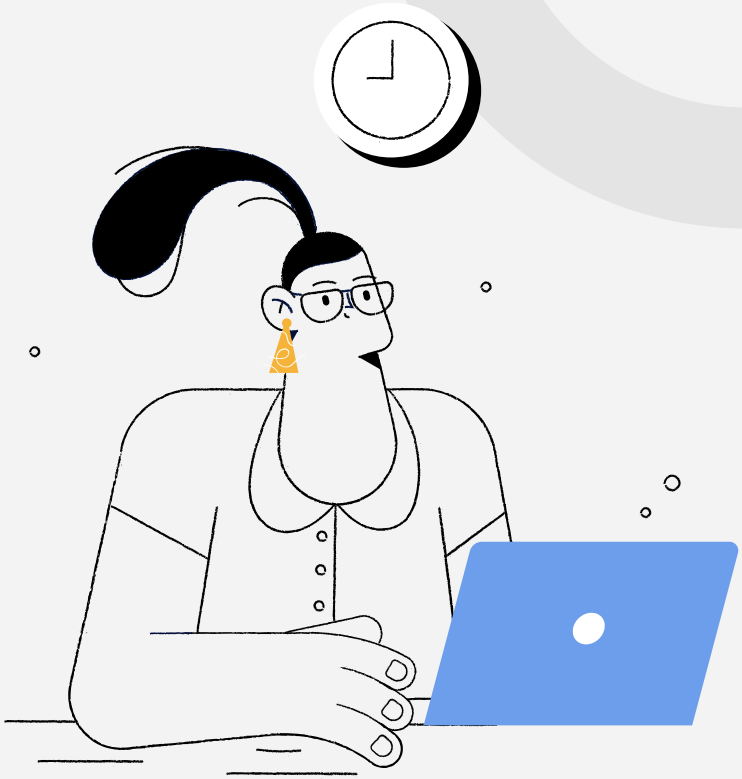
- Reviews with adjectives (eg. premium, expensive, surprised, disappointed) are likely to be real reviews.
- **'I'** has relatively higher positive weights. This means that fake reviews are likely to be written in the first-person.
- **'the'** is the most commonly used word in the english language, having it being a strong predictor a fake review means that there will be a larger number of reviews being incorrectly classified as fake, which is confirmed to be the case (refer to Confusion Matrix for the DistilBERT Classifier)

Recommendation

Model	Variants	Accuracy	F1 Score
Multinomial NB	CVEC	0.830	0.842
Logistic Regression/ TF-IDF Vectorizer	TVEC	0.902	0.891
DistilBERT Classifier	-	0.970	0.970

- If immediate deployment is required, recommend to adopt the DistilBERT model as it has the highest performance, model training time is also fast (~30 minutes for a mini batch of 2.8k batches for 4 epochs)

- If time allows, preferable to use Logistic Regression due to better interpretability, lower complexity and training time
(< 5 minutes), will likely meet F1 score requirements after further hyperparameter tuning



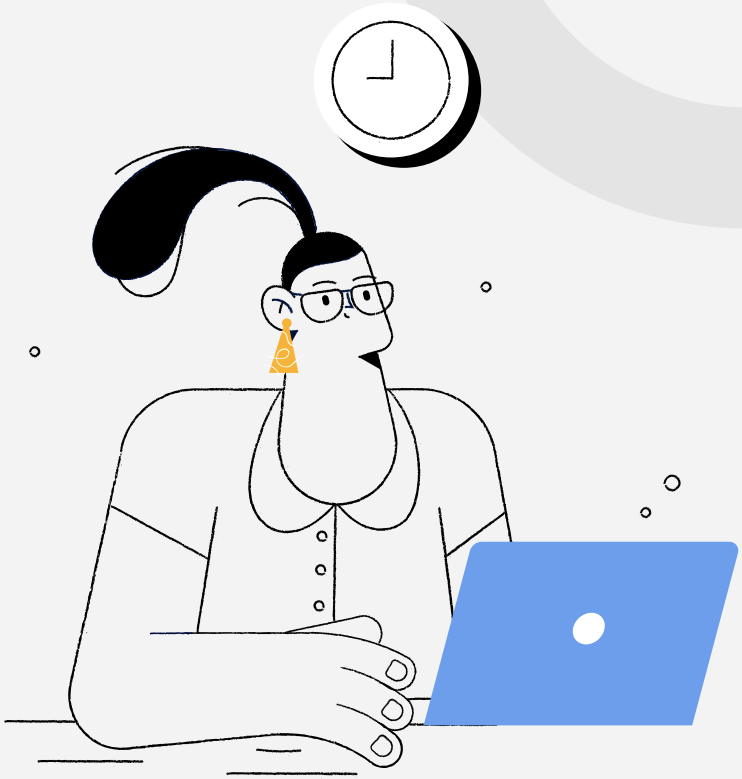
05

Limitations

What can be better?

Limitations

- For DistilBERT model, max length of a review is kept to 256 tokens due to local memory issues (max length for input to the DistilBERT model is 512 tokens), performance may increase if more tokens are added.
- Analysis using LIME can be more exhaustive. Constrained by time.
- Initialize OOV words in Pre-trained embeddings with random values instead of zero vectors, may yield better results.
- More comprehensive hyperparameter tuning of both Machine Learning and Deep Learning models may yield additional performance benefit. Tuning done in this project is only indicative of model performance.
- Additional types of ML and DL models may provide better result (K Neighbors, Gradient Boosting, RNN, Stacked LSTM, GRU), and will be attempted should timeline permit.



06

Challenges

Knocks along the way

Main Challenges

Reproducibility of DL Models

Although no impact to conclusions drawn, unable to obtain totally reproducible DL results in Pytorch

Learning a new DL Framework

Pytorch not taught in class, have to self-learn (get used to it!)

Hardware is key for DL

6GB GPU can only handle so much.

Can a machine beat a machine?

YES! (For now)

Doing pretty good for now, we are able to detect a large majority of computer generated text but it is an arms race as text generation gets more and more advanced (i.e. GPT-2 => GPT-3)

Now for some fun..

This pillow saved my back. I love the look and feel of this pillow.

Love this! Well made, sturdy, and very comfortable. I love it!Very pretty

Easy to review. The kid DDS love them! Thanks

Bought as a gift for a cousin. Very nice.

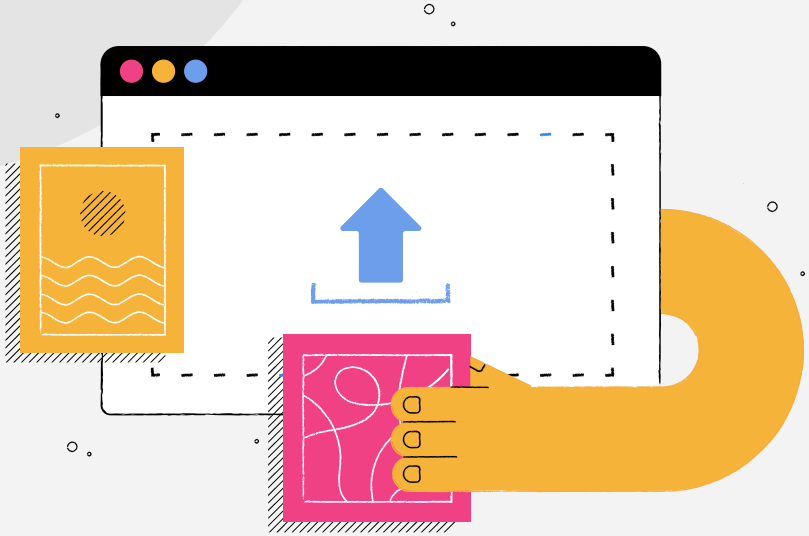
Now for some fun..

This pillow saved my back. I love the look and feel of this pillow.

Love this! Well made, sturdy, and very comfortable. I love it!Very pretty

Easy to review. The kid DDS love them! Thanks [REAL]

Bought as a gift for a cousin. Very nice. [REAL]



Thanks!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**