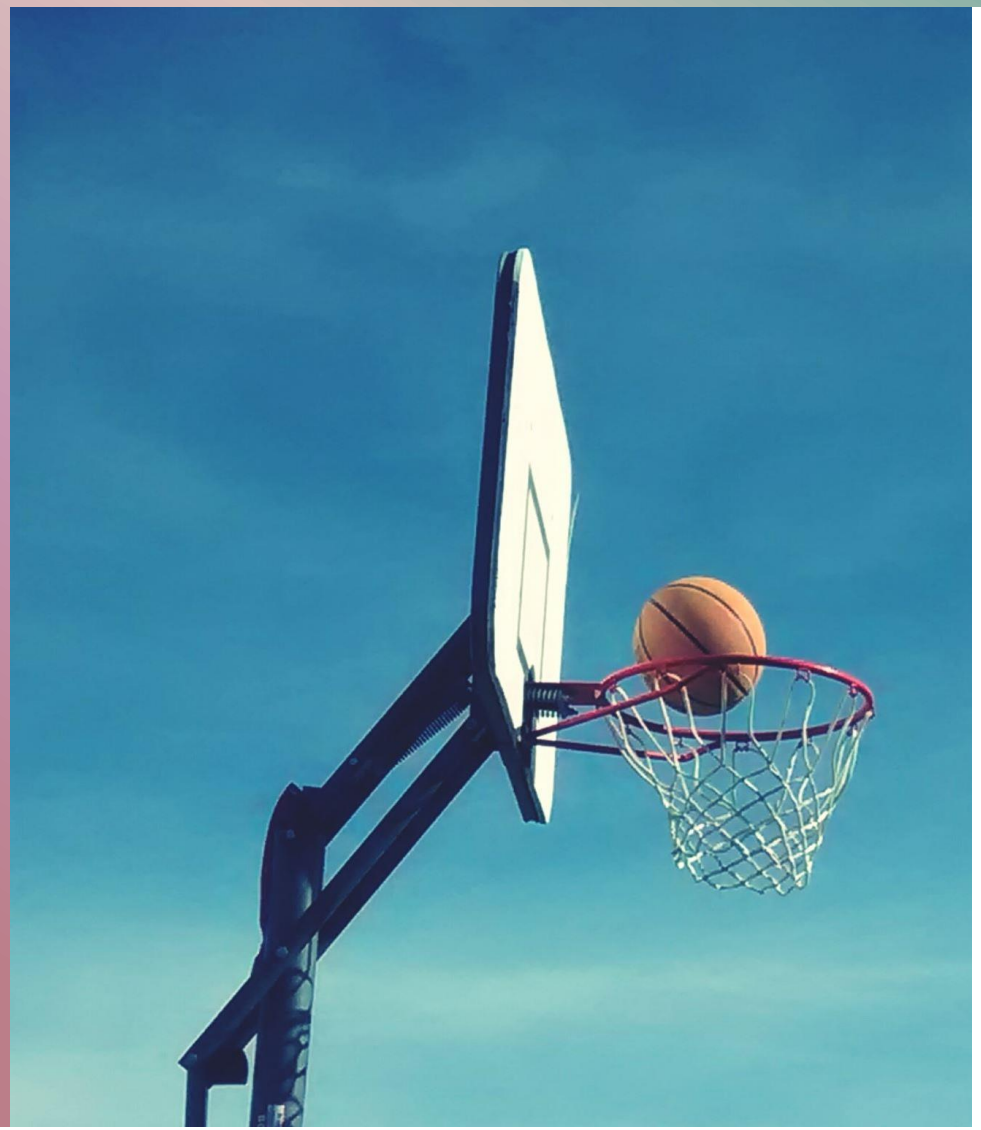
Basketball Post Sentiment Analysis

By Daniel Cohen, Kayvan Khoobehi, Gabriel Kusiatin, Abe Zaidman and Jack Whalen

Agenda

- Project Objective
- Data Collection and Processing
- Model Construction
- Google Cloud Incorporation
- Streamlit Cloud Application



Objective

- Our objective is to build a Streamlit app that predicts whether a r/NBA user post is positive or negative.
- NBA officiating, player actions and corporate leadership have been under large levels of scrutiny in recent months. We want to keep track overall of public opinion on these matters as the NBA playoffs continue.

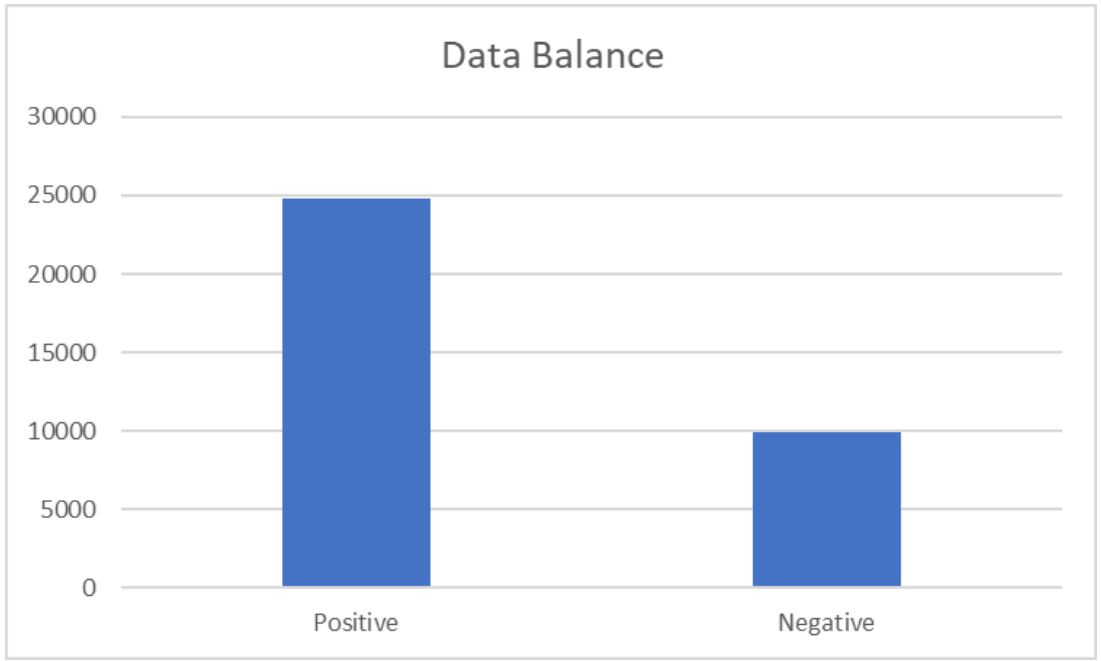
Collecting Training Data

- We used a database from Kaggle that contained over 30,000 Reddit posts about basketball.
- We removed all columns besides Title and Body, which contained the text data.
- We filled selftext NA values with blank spaces.

Preprocessing the data

- We tokenized the text using NLTK's word_tokenize() function.
- We removed URLs using regular expressions.
- We removed stop words using NLTK's set of stop words.
- We lemmatized tokens using NLTK's WordNetLemmatizer().
- We removed non-alphabetic characters and words less than 3 characters.

Data Balance



Sentiment Analysis

+ We used NLTK's SentimentIntensityAnalyzer() to conduct sentiment analysis on the training data.
+ We assigned a positive or negative label to each post based on the processed title and selftext of the post.

Building the Model

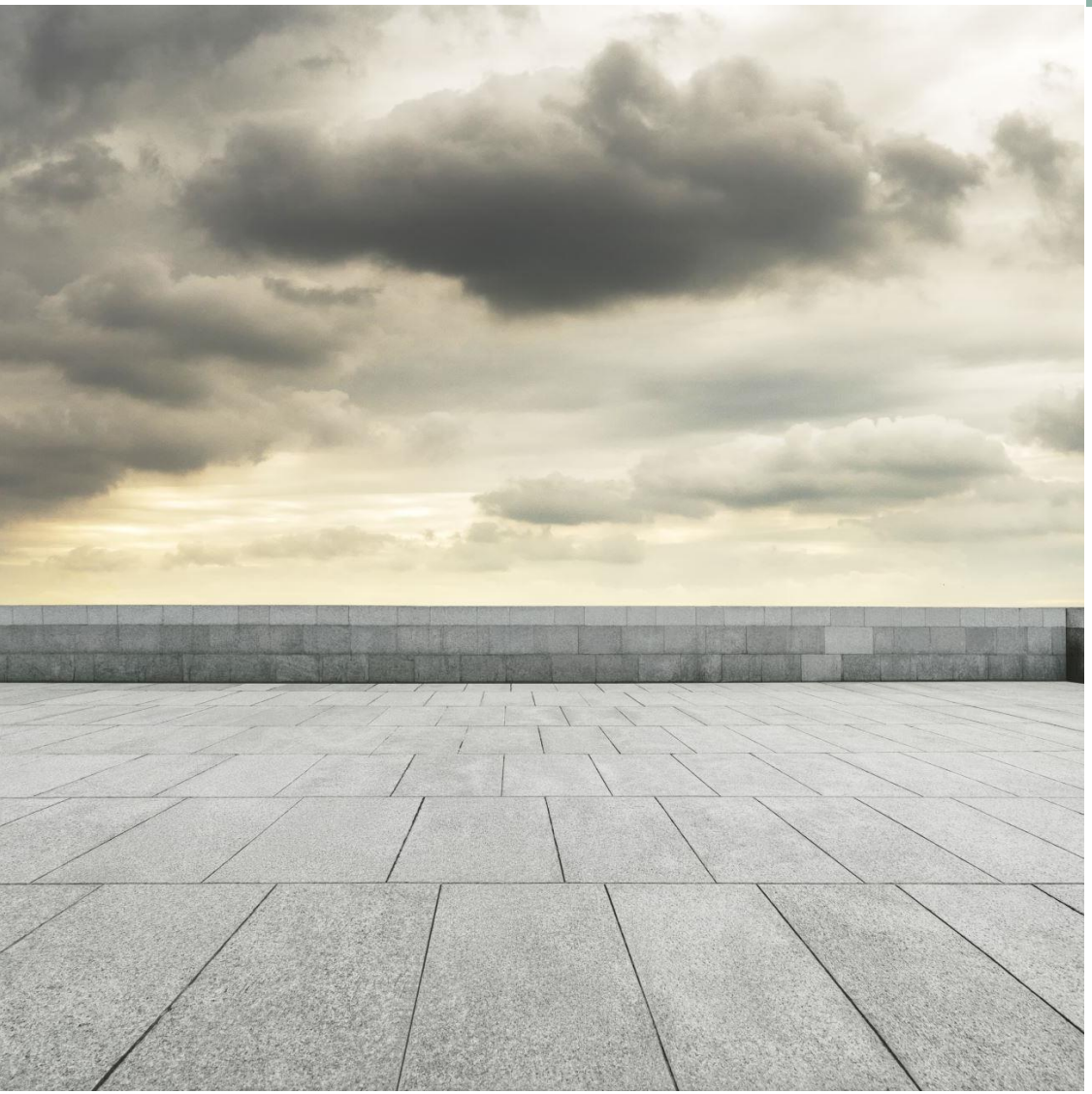
- We split the data into training and test sets (80% train, 20% test).
- We built a model pipeline using CountVectorizer to convert the text into numerical features and Logistic Regression to make predictions.
- We trained our model using a dataset of random Reddit text data

Model Performance on Testing Data

- Test Dataset Accuracy = .9067
- Test F1 Score = .9358
- .22 False Negative Rate
- .034 False Positive Rate

Future Model Improvement

- The data was slightly imbalanced, with around 25,000 positive posts and 10,000 negative posts.
- Further data pre-processing steps could be taken to increase model accuracy.
- Using a different Vectorizer, such as TF-IDF, may yield better results.
- Incorporating more features, such as information on comments, upvotes, and subreddits, may improve the model's performance.



Google Cloud Storage

- We decided to pull posts from the NBA reddit and upload them to the Google Cloud Firestore
- Google Cloud Storage allows us to update our Streamlit app in real time
- This means that as new posts are added to the NBA subreddit, they are automatically updated in our database and reflected in our app

Streamlit App

- Our app allows users to see samples of training data, model performance scores, and test the model themselves with both manually inputted text and copied headlines/articles to analyze

Dataset Sample and Model Performance

Sample Data from `combined.csv`

	Unnamed: 0	ID	is_Original	Flair	num_comments	Title
0	0	fcpbui	False	None	13	Can't have sex the s
1	1	fbtk1w	False	None	3	How to break up wit
2	2	1681jg	False	None	4	[MODPOST] [META]
3	3	fcmxds	False	None	0	The most important
4	4	f9gz1a	False	None	15	My dad found my pc

Model Performance Metrics

- Accuracy Score: 0.9067
- F1 Score: 0.9358
- False Negative Rate: 0.22
- False Positive Rate: 0.034

Thank You!
Any Questions?