Data Mining and Machine Learning Project

PODCAST REVIEWS ANALYSIS

Ludovica Cocchella

OBJECTIVES

Objective 1: Sentiment Analysis

The main goal of this project is to conduct Sentiment Analysis on podcast reviews, employing classification techniques to categorize the reviews into positive, neutral, or negative sentiments. The project aims to systematically evaluate and compare different classifiers to determine the most effective one for the analysis

Objective 2: Streaming Analysis

The second obejctive is to implement Streaming Analysis which is dynamic approach that tracks the evolution of sentiments toward podcasts over time. The main aim is to continuously assess the efficacy of our classifier in adapting to changing trends and the evolving emotions expressed by listeners.

DATASET

Source:

- https://www.kaggle.com/datasets/thoughtvector/podcastreviews

Size:

- 2 million reviews for 100,000 podcasts.

Content:

- User reviews on podcasts.

Attributes:

- Podcast ID
- Title
- Content
- Rating
- Author ID
- Created_at

SENTIMENT ANALYSYS

STEPS

1
Dataset
Preprocessing

2
Text
Preprocessing

3
Classifiers
Evaluation

4
Model
Selection

DATASET CLEANING

1: Handling Missing Values:

- Remove missing values

2: Handling Duplicates Reviews:

- Remove duplicates reviews

3: Text Preprocessing:

- **Emoticon Removal:** Eliminating elements like :), :(, etc.
- **URL Removal:** Removing strings containing URLs or web addresses.
- **HTML Tag Removal:** Ensuring a clean textual representation.
- **Unicode Normalization:** Removing accents or non-ASCII special characters.
- **Removal of Special Characters, Numbers, and Punctuation:** Retaining only alphabet letters and spaces.
- **Conversion to Lowercase:** Ensuring standardization.
- **Trimming Whitespaces:** Removing leading and trailing whitespaces.
- **Removal of Extra Spaces:** Reducing multiple whitespaces to a single space.

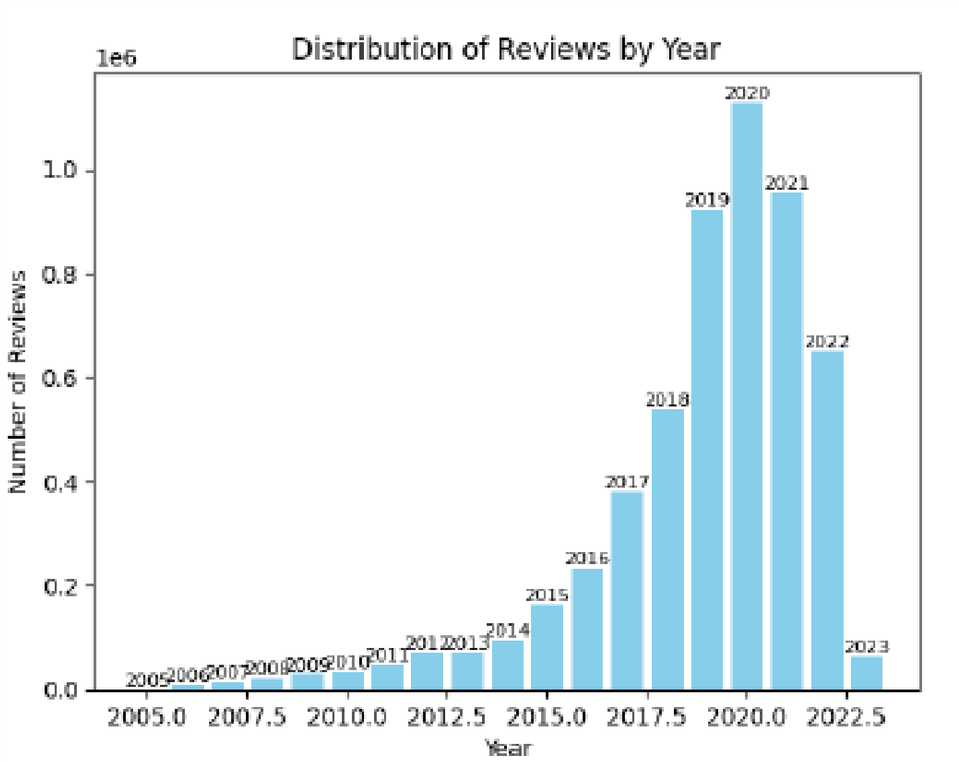
DATA REDUCTION

Dimensionality reduction:

- Attribute selected:
  - Content: Text of reviews used to build the classifier.
  - Rating: Used to establish a ground truth, dividing comments into three classes.

Numerosity reduction:

- Initial vs. Selected Reviews:
  - Initial Reviews: 5,430,620.
  - Year 2020 Reviews: 1,127,656.



TRAINING SET BUILDING

Exclusion of Ambiguous Ratings:

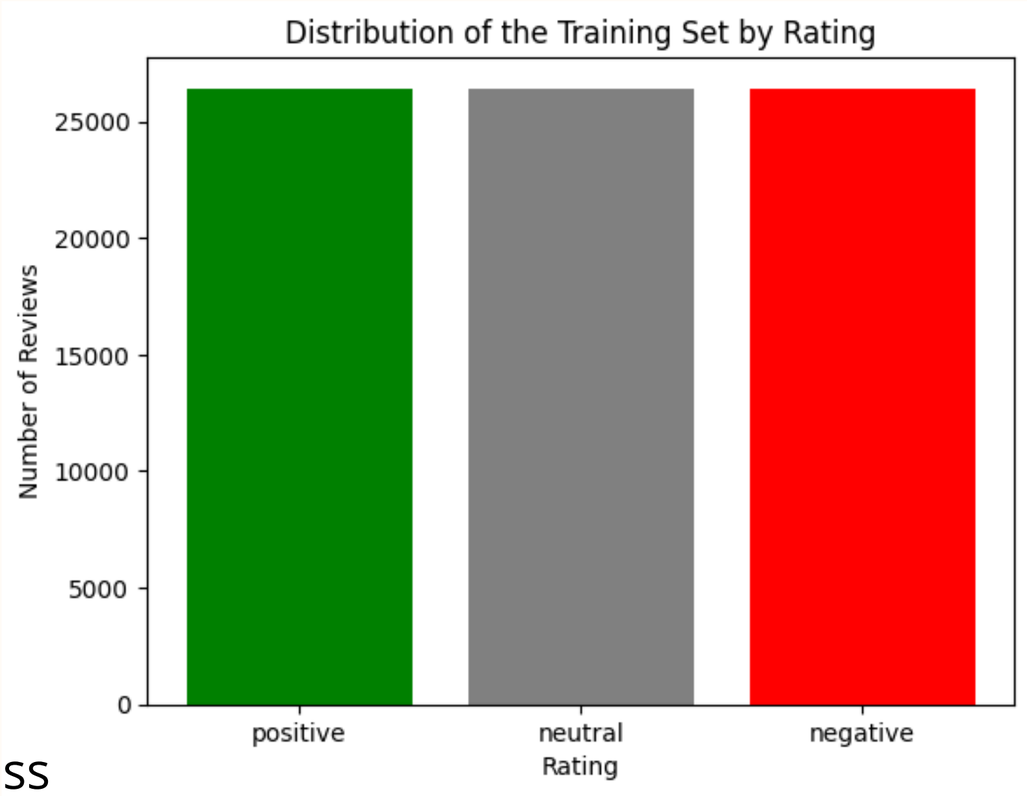
- Excluded comments with ratings 2 or 4 (they don’t have a well-defined class).

Mapping Ratings:

- Rating = 1 negative comment
- Rating = 3 neutral comment
- Rating = 5 positive comment

Balanced Training Set:

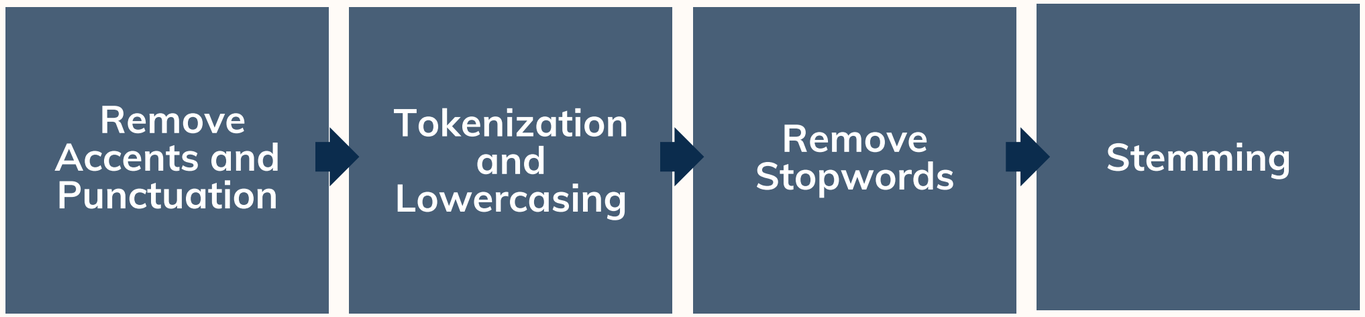
- Balanced set with 26,412 reviews for each class



TEXT PREPROCESSING

Custom Analyzer:

- implement the following step of text preprocessing



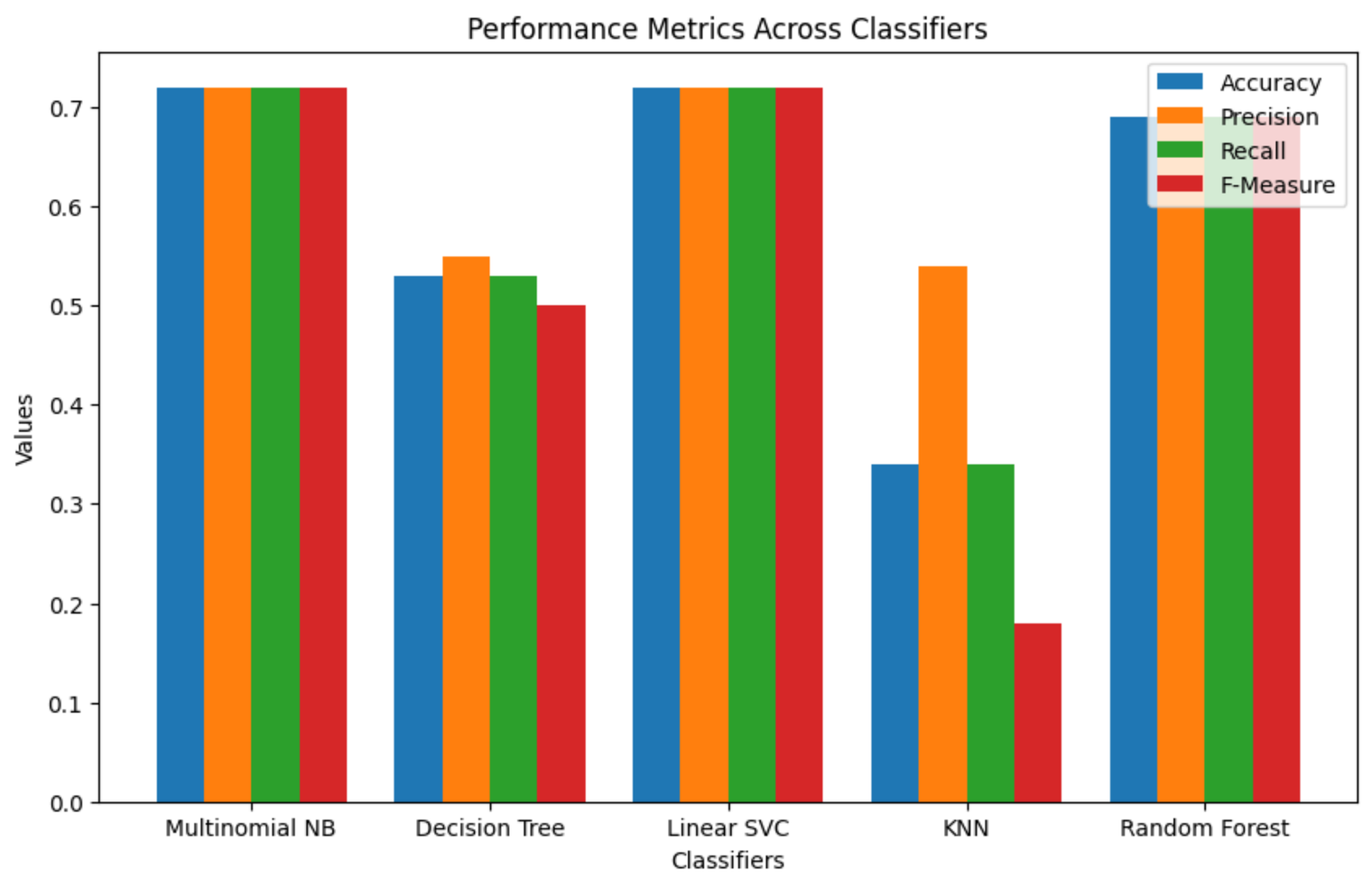
CountVectorizer:

- uses **custom analyzer** to process and tokenize the text before assigning a numerical value to each word based on its frequency of occurrence in the documents in order to transform the comments into a Bag of Words representation

TfidfTransformer:

- converts the count matrix generated by CountVectorizer into a TF-IDF representation. This step is useful for assigning greater importance to less common words and reducing the importance of common words.

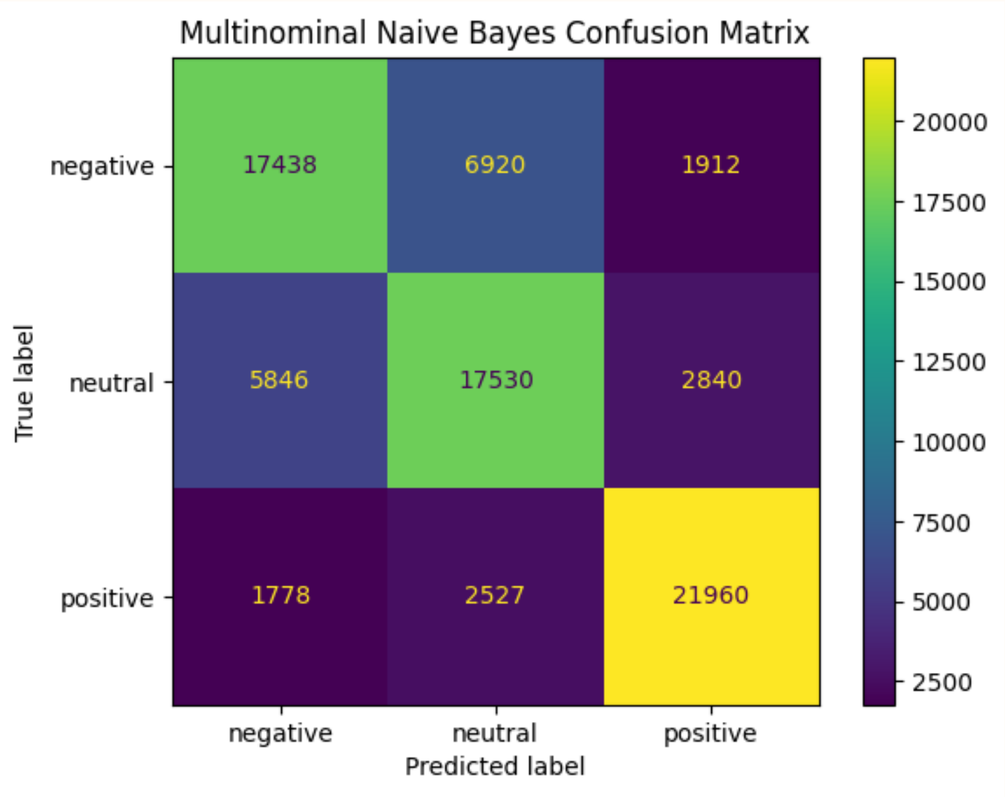
CLASSIFIERS



MODEL SELECTION 1/3

- **Multinomial Naive Bayes**

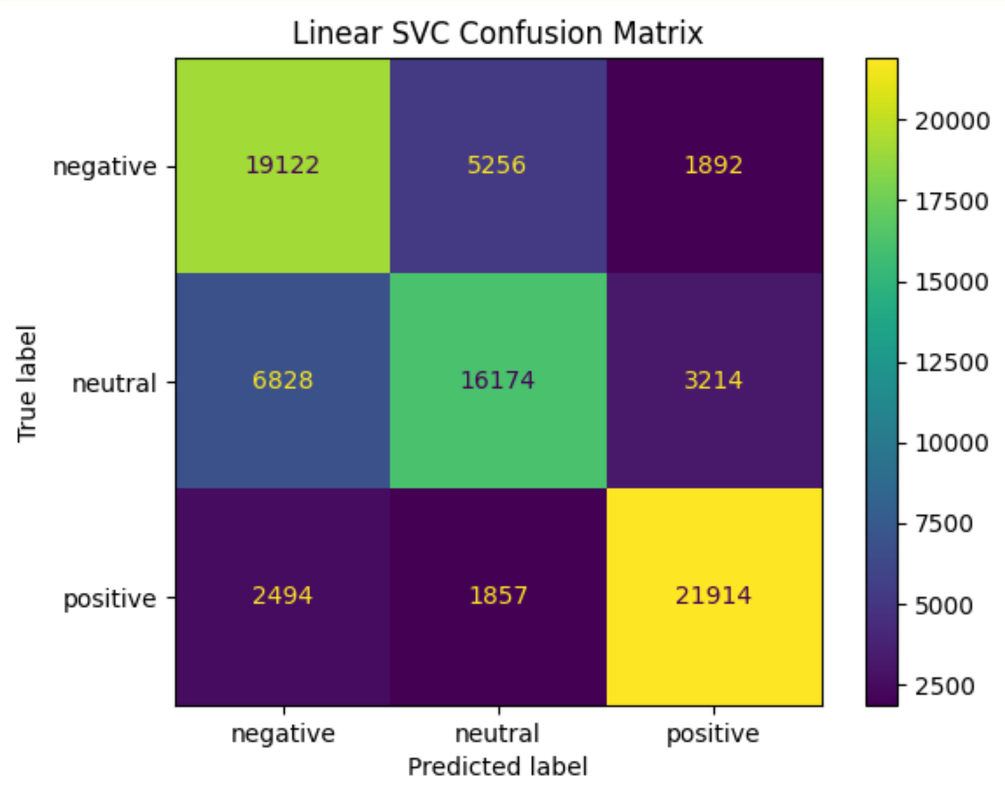
Class	Precision	Recall	F1-Score	Support
Negative	0.70	0.66	0.68	26270
Neutral	0.65	0.67	0.66	26216
Positive	0.82	0.84	0.83	26265
Accuracy			0.72	78751
Macro Avg	0.72	0.72	0.72	78751
Weighted Avg	0.72	0.72	0.72	78751



MODEL SELECTION 2/3

- **Linear Support Vector (SVC)**

Class	Precision	Recall	F1-Score	Support
Negative	0.70	0.66	0.68	26270
Neutral	0.65	0.67	0.66	26216
Positive	0.82	0.84	0.83	26265
Accuracy			0.72	78751
Macro Avg	0.72	0.72	0.72	78751
Weighted Avg	0.72	0.72	0.72	78751



MODEL SELECTION 3/3

- **Statistical Test:** Conducted a paired t-test on two leading classifiers: Multinomial Naive Bayes and Linear SVC
- **Test Outcome:** Obtained a p-value of 0.00116, signaling p-value < significance level ($\alpha = 0.05$)
- **Null Hypothesis:** Null hypothesis rejected, indicating significant performance differences between the models
- **Model Choice:** Chose the Linear SVC over Multinomial Naive Bayes based on accuracy assessment because Linear SVC exhibits a lower error rate, ensuring superior accuracy in sentiment classification

STREAMING ANALYSYS

STEPS

1
Extract new comments

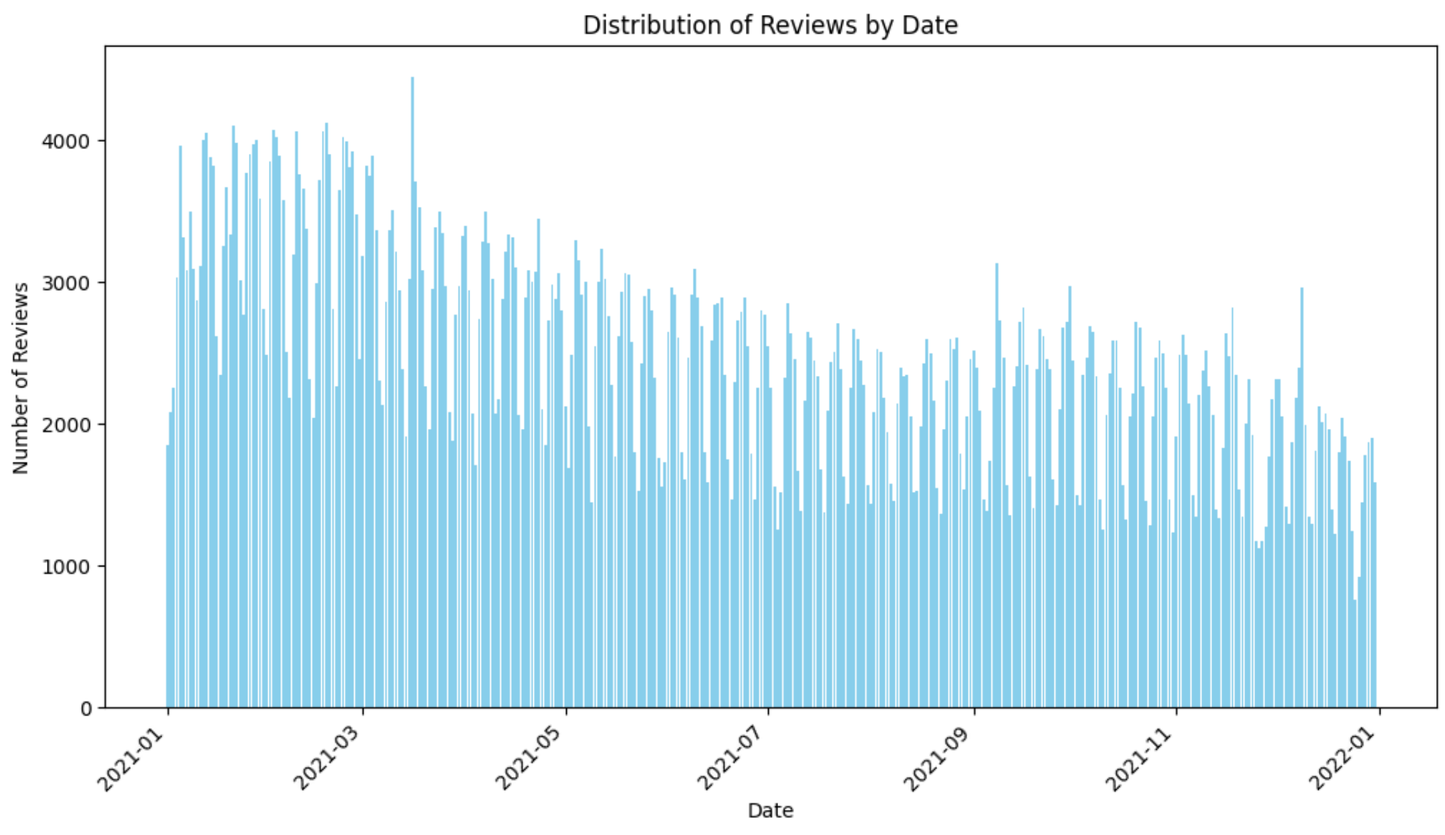
2
Identify key events

3
Build Models

4
Model Selection

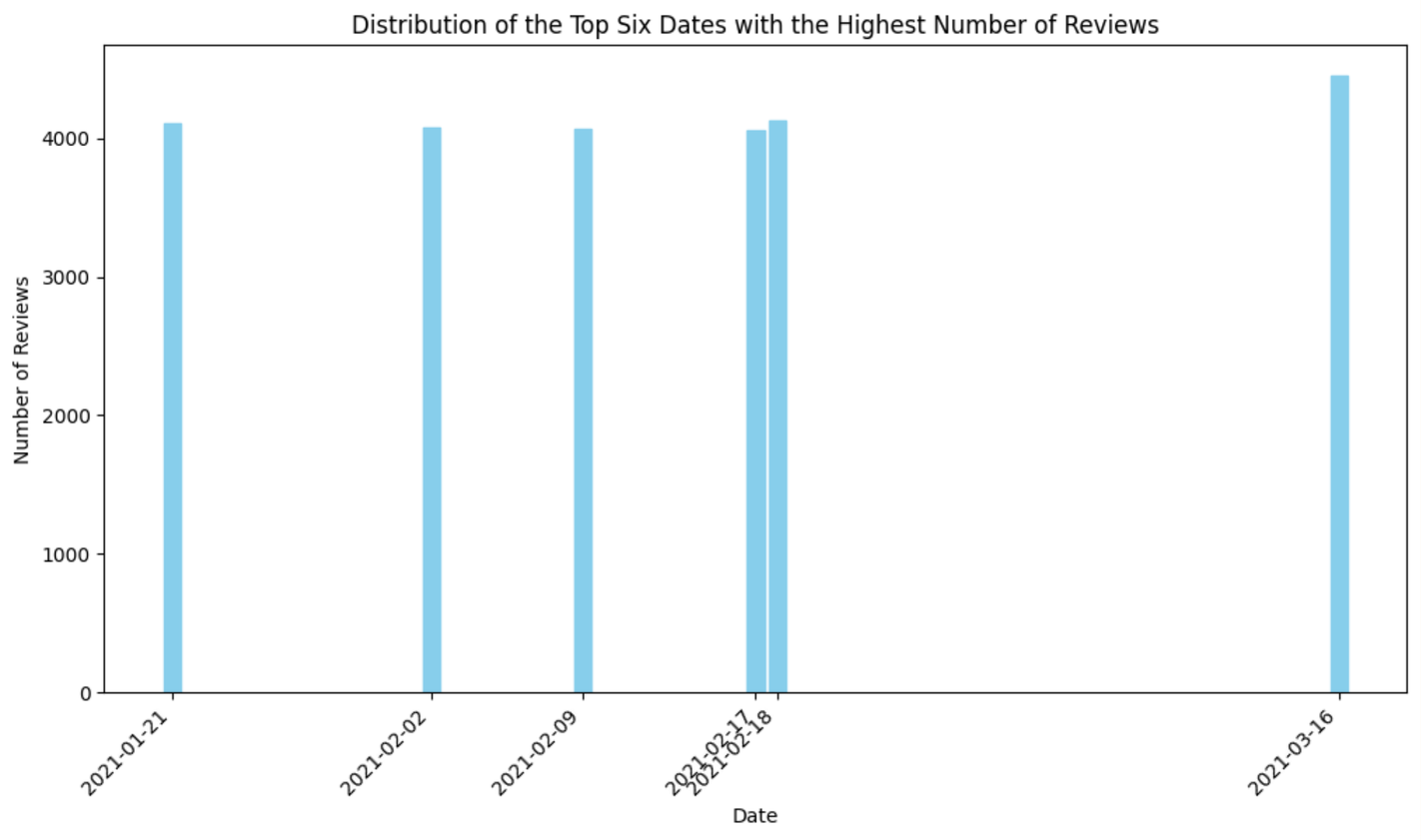
NEW COMMENTS

- Time Windows: 2021



EVENTS

- Events: Peak of comments and extract the 6 key events on the timeline

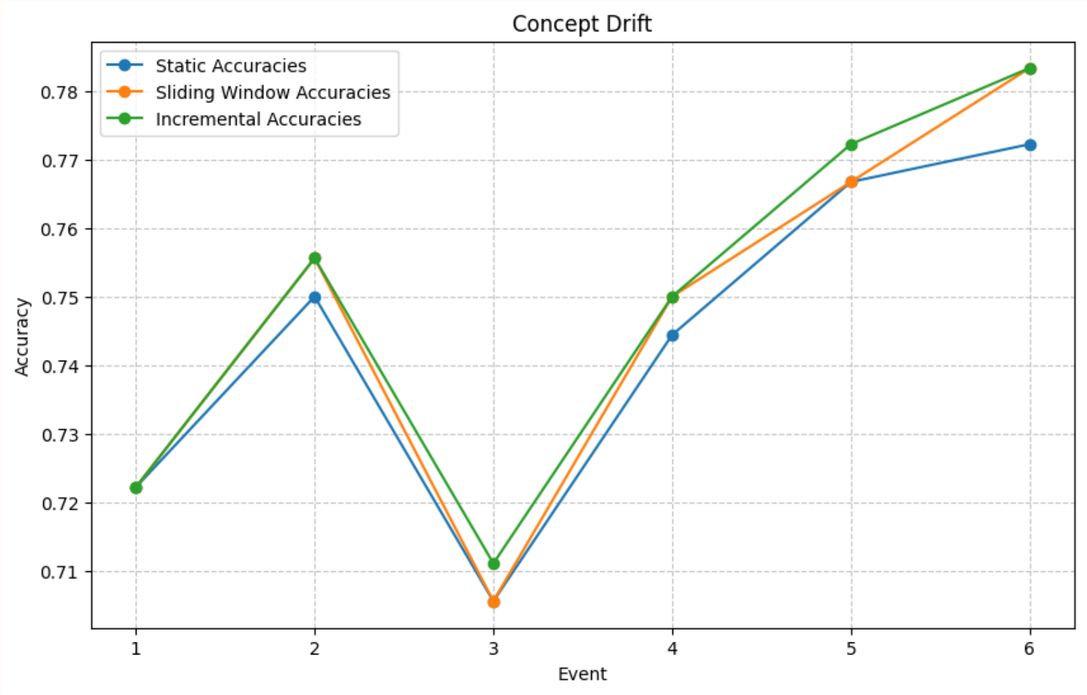


MODELS

For each event, 60 new comments were selected for each class and labeled. These comments were then used as a test set for three different learning settings:

- **Static Model:**
  - The initial model was constructed using the first training set.
- **Sliding Model:**
  - This model was retrained periodically with the most recent comments of the training set. In each iteration, the oldest 60 comments were removed, and the newest 60 comments were added.
- **Incremental Model:**
  - Trained with the initial training set adding the labelled data of all previous events.

RESULTS



The static model has a fixed dictionary, while the sliding and incremental models update theirs over time. However, neither the sliding nor incremental model shows an increase in vocabulary size.

The analysis reveals that the accuracies of the three models remain stable over time, with no apparent signs of deterioration.