PROJECT 4

STEAM GAME REVIEWS – NATURAL LANGUAGE PROCESSING

ERIC HANSEN

FLATIRON DS; SELF PACED

4/3/2022

BACKGROUND - CONTEXT

- Classification of text reviews is a useful tool for many different fields and is an ideal application of Natural Language Processing (NLP) techniques within Machine Learning/Deep Learning.
- We investigate reviews of video games on the Steam platform (a digital distribution service and storefront developed by Valve). Launched in 2003, users can view ratings and reviews, and purchase and download games.
- In 2017, Steam earned $4.3 billion in sales.
- As of March 29, 2022, Steam reports over 27 million concurrent users at peak logged in to their service, with over 120 million total users.

BACKGROUND - STAKEHOLDERS

- Stakeholders in this process include (but aren't limited to)
  - Valve (or other digital distributors, e.g. Epic Games, Xbox Games Store, Apple App Store, Google Play),
  - video game developers/publishers,
  - video game consumers (including reviewers, critics, players, streamers, etc.),
  - computer hardware manufacturers (e.g., insight about game popularity could drive video card architecture)
  - cultural linguists (due to the nuanced and specialized vernacular exhibited by video gaming communities).

BACKGROUND – DATA

- Kaggle – Steam Game Review by Luiz Martins; sourced from Steam Digital Distribution
- Link to data

MODELS

- 32 NLP models, including on both CountVectorized and TF-IDF data, with/without Lemmatization: logistic regression, naïve Bayes, support vector machines, singular value decomposition, XGBoost, and multiple versions each of deep learning models LSTM and GRU (14 different DL architectures)

GOALS

- Applications
  - Validate new ratings (i.e. checking consistency between user text and user binary rating).
  - Prompt users to verify their intended binary rating if their text review seems inconsistent.
  - Notify distributor of potential abuse/manipulation of the rating system
  - Apply a rating to unlabeled reviews, acquired from off-platform sources
  - Generate continuous (non-binary) rating number for each review

METHODOLOGY DETAILS

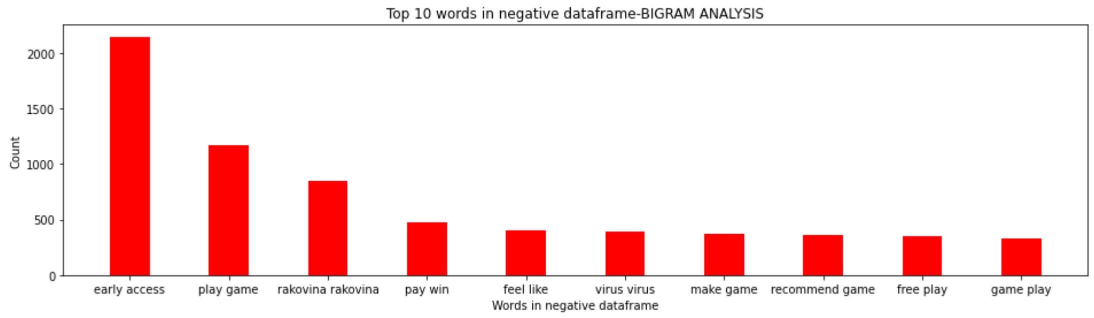
- 1) I performed Exploratory Data Analysis to identify characteristics of the data
- 2) I created numerous models, optimized them, and compared their performance on several metrics (f1, log-loss, AUC, accuracy)
- 3) I identified one of the high performing, interpretable models for word features of high importance
- 4) I identified the models with overall best performance on accuracy
- 5) I created a pickled version of the best model for portability and ease of future use
- 6) The conclusions: Logistic Regression on Lemmatized data categorized reviews best, with accuracy of 87%

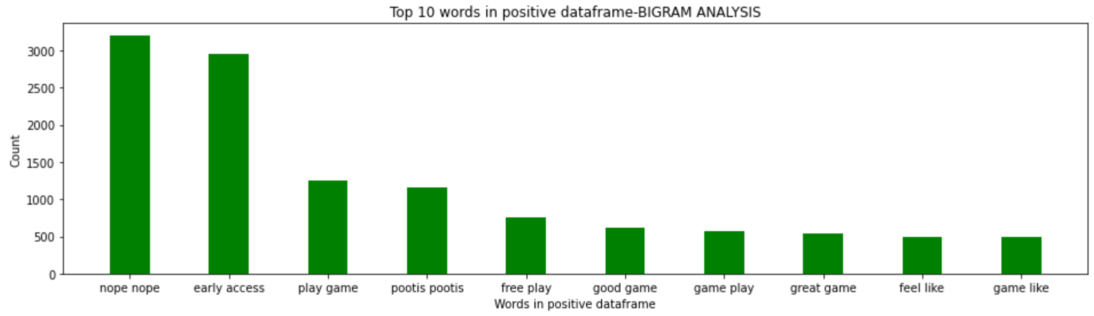
VISUALIZATIONS 1 – WORD CLOUD OF FREQUENTLY USED WORDS

- Highly used words for both positive and negative reviews are shown
- Commonly used words to describe games appear in both positive and negative reviews, but when there is difference, this can help develop insight

VISUALIZATIONS 2 – BIGRAM HISTOGRAM

MOST FREQUENT BI-GRAMS SHOWN IN HISTOGRAM FORM



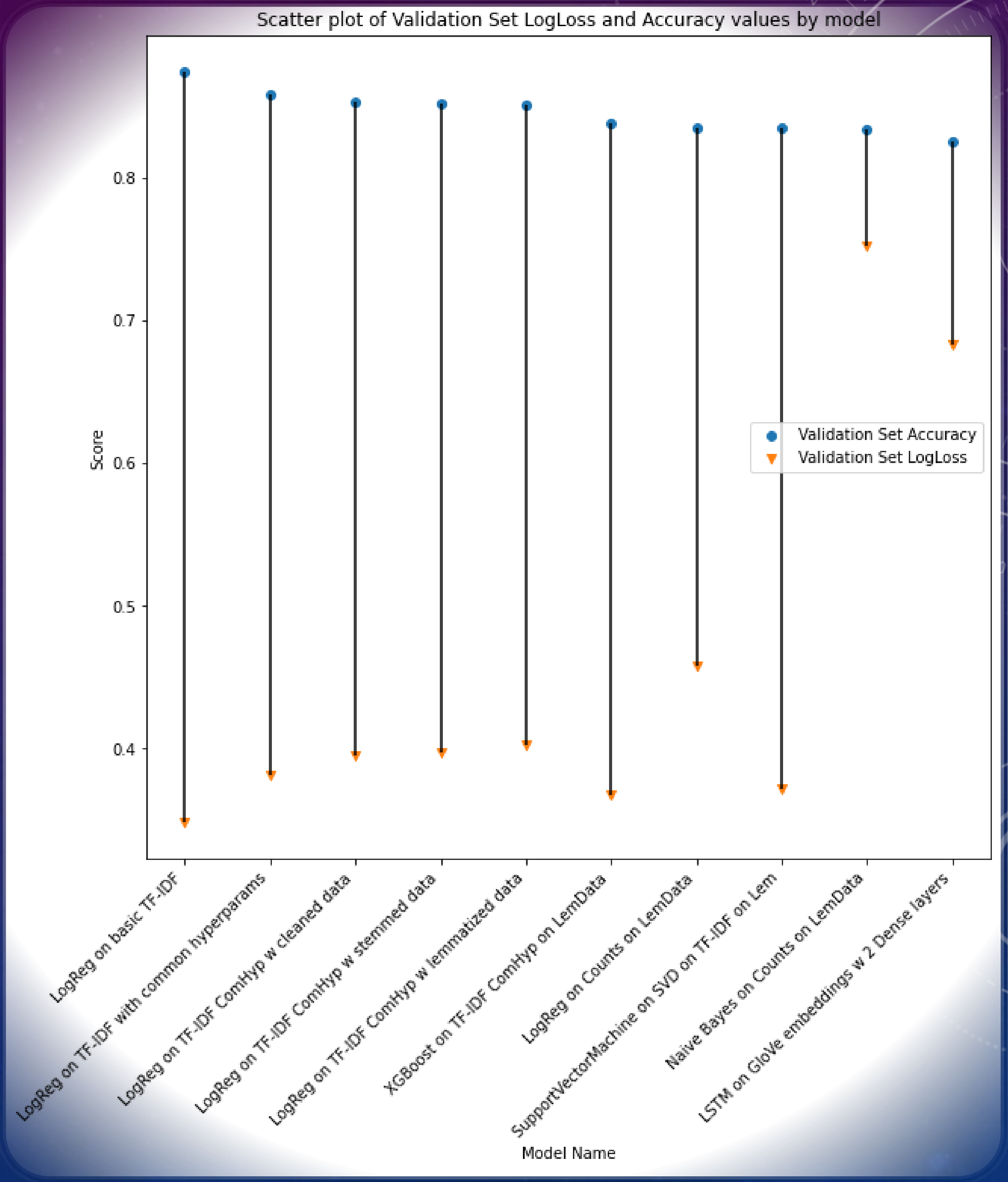


VISUALIZATIONS 3 – BIGRAM BUBBLEPLOT

- Positive count on x-axis
- Negative count on y-axis
- Overall frequency is size
- Color shows positivity
- Interactive version available

MODEL RESULTS

- 32 Models were produced
- Of these, the top 10 highest-accuracy (on validation set) models are presented
- Highest Accuracy Model is Logistic Regression on TF-IDF, with score of 87%
- (Industry best is mid-90%, but that doesn't factor in so much slang/vernacular)
- Naïve Bayes is also investigated further for its interpretability



NAÏVE BAYES: WORD SALIENCE

- The model that performed Naïve Bayes prediction on count vectorized data had respectable accuracy, but also is notable because the word weights are accessible and interpretable
- Upon word salience investigation, relative word importance sheds some illumination
- “Pootis” has a high correlation with positive reviews. It is a phrase from a game called Team Fortress; this is a phrase very localized to video gaming.
- “Rakovina” (a Slavic word for cancer) has a strong correlation with negative review. However, as above, this word is unlikely to appear in “usual” or pretrained models
- It is possible that the high presence and influence of these unusual words caused the models that were based on pre-trained vectors (e.g. GloVe) to underperform
- It is also possible that the unusual grammatical structures also contributed to the deep learning models’ lower performance.

CONCLUSIONS/QUESTIONS

- Steam reviews can be classified positive/negative with ~87% accuracy.
- NLP analysis of video game reviews is nuanced due to specific vernacular to that community
- Model could be extended to produce continuous "positivity" sentiment score, rather than just binary

CONCLUSIONS/QUESTIONS

- Actionable items using this model:
  - Validate user-generated scores
  - Bolster fraud/abuse checking
  - Incorporate generated scores into game ratings
    - Include off-platform reviews for more comprehensive score

CONCLUSIONS/QUESTIONS

- Potential further steps:
  - acquire more labeled data to further train models
  - use extracted observations about word frequencies over time to investigate emergent language behavior

CONCLUSIONS/QUESTIONS

- Questions?

CONCLUSIONS/QUESTIONS

- Thank you!