AMAZON REVIEW SENTIMENT ANALYSIS

Abeselom Fanta

April 20, 2022

OUTLINE

- BUSINESS AND DATA UNDERSTANDING

- MODELING

- RESULTS

- CONCLUSION

- NEXT STEPS

OVERVIEW

- Sentiment analysis is one the of the most important tasks to understand user satisfaction.

- Most websites that offer products and services have various means of keeping track of user satisfaction criteria, such as stars-based system.

- However, most users are disincentivized to provide accurate rating for the products or services they purchased. In addition, manually sorting through users' comments to determine if the comments left by users/clients is positive or negative takes a lot of work.

- Therefore, the problem necessitates an automated way to determine sentiment analysis of clients.

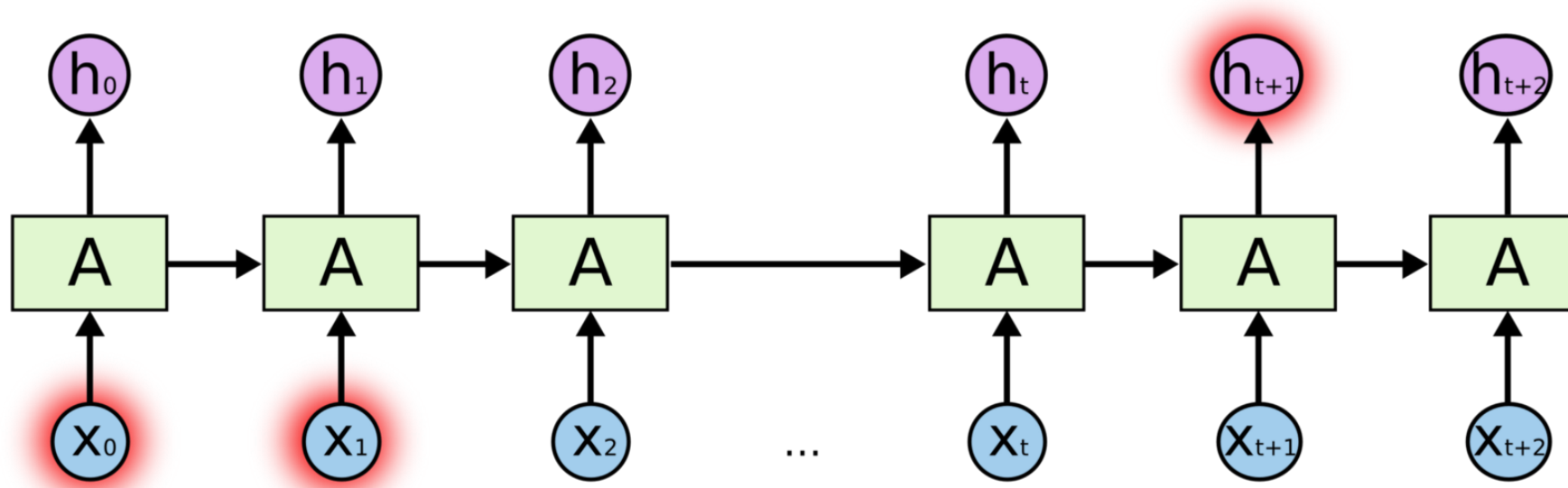# BUSINESS AND DATA UNDERSTANDING

- The data used in this project was obtained from Kaggle.

- The dataset contains four million comments (3.6 million training and 0.4 million test datasets).

- The files are presented in fastText format, which will be parsed to the required type of data for processing.

- Both the training and test datasets are labeled, which will help in quantify how the predictions measure with true labels.

MODELING

- Text data was cleaned
  - De-contracted
  - Spell checked
  - Email and URL removed
  - Lemmatized and tokenized
- Models used
  - Shallow ML (naïve Bayes, logistic regression and random forest)
  - Gradient Boost (XGBoost and CatBoost)
  - Deep Learning (LSTM, LSTM with GloVe and DistilBert tokenized CNN)

RESULTS

MODEL COMPARISON

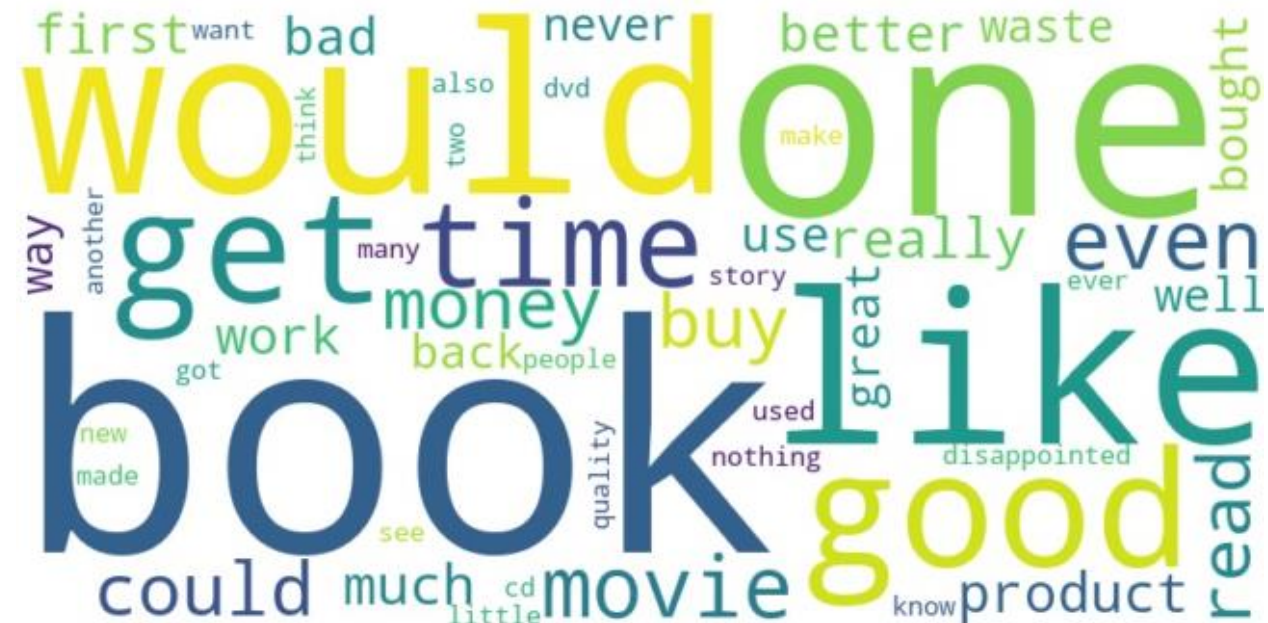LSTM model achieved 91.4% accuracy on a test dataset

MODEL PREDICTION

Model was able to identify key words for positive or negative sentiment classification

True Positives

True Negatives

WORD FREQUENCY

Predictions from the final model were used to display the word frequency. The 50 most frequent words in the true positive and true negative are shown as follows.

CONCLUSION

- We built a binary text classification model LSTM that is able to perform Amazon reviews sentiment analysis with 91.4% accuracy.

- Our model was also able to assign probabilistic ratio for each word used from a completely unseen review.

- Word Clouds were generated to show which words are the appear the most in both true positive and negative reviews.

- It was also observed that LSTM with GloVe and BERT tokenized models also performed strongly.

- More deep learning architectures could be considered to improve the accuracy of test data.

NEXT STEPS

- The next steps include considering more robust models for an improved performance.

- Implement transformer models for a significantly higher performance compared to the traditional LSTM model.

- The dataset used in this project have binary classes. It would be interesting to test the model performance on a star-based rating, instead of positive or negative reviews.

- Created a web deployable application for an interactive feedback.

THANK YOU

Abeselom Fanta

GitHub: @afanta-fi