Proposal: Sentiment Analysis on Steam Reviews using Hidden Markov Models

Ninja, Ton

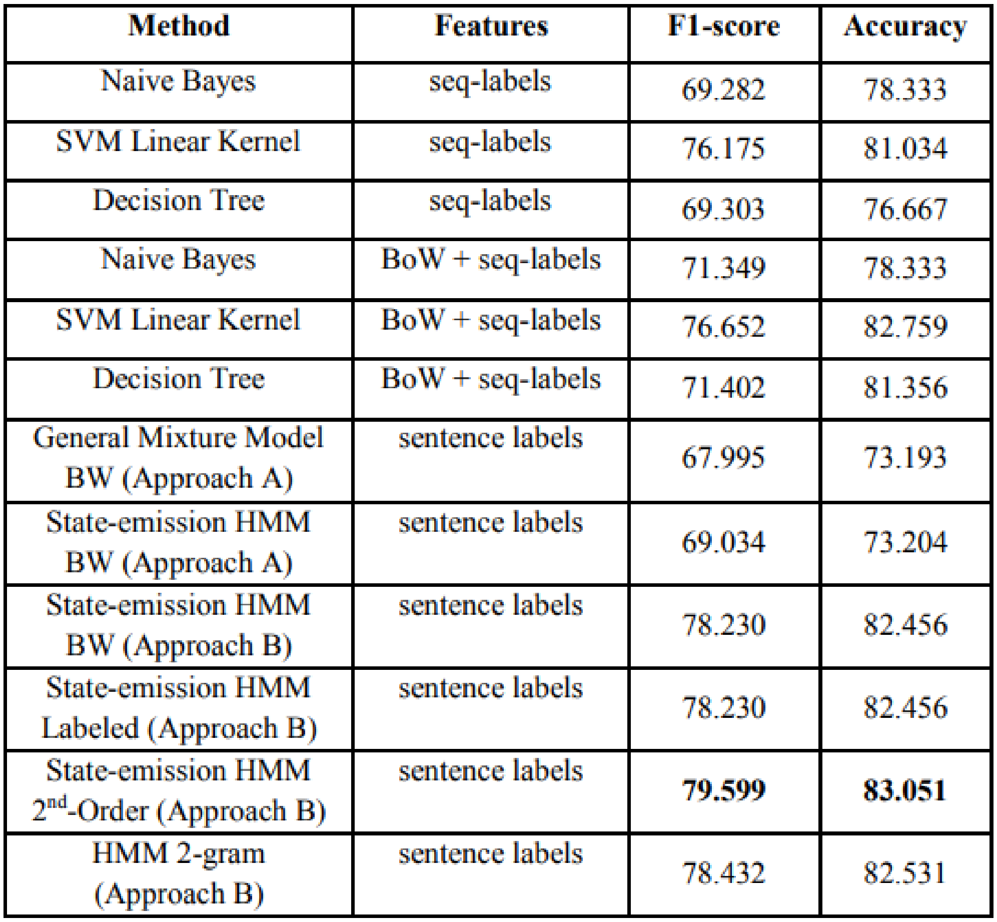
March 9, 2023

1 Introduction

The objective of this project is to perform sentiment analysis on Steam reviews using hidden Markov models (HMMs). Steam is a popular online gaming platform, and analyzing user reviews can provide insights into user preferences, satisfaction levels, and areas for improvement. The primary stakeholders for this project are game developers, who can use the analysis to improve their games, and Steam users, who can benefit from improved games.

The problem of sentiment analysis involves categorizing text into positive, negative, or neutral sentiment. This problem is ill-defined as it is challenging to determine the sentiment of text accurately. We plan to use HMMs to model the problem as it has shown some potential alongside the traditional method such as SVM, Naive Bayes, and others (e.g., TABLE IV in [1]).

TABLE IV. PERFROMANCE ON THE SPOT DATASET

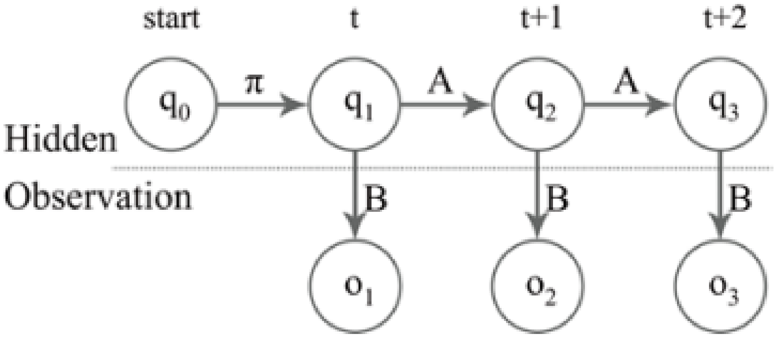


Method	Features	F1-score	Accuracy
Naive Bayes	seq-labels	69.282	78.333
SVM Linear Kernel	seq-labels	76.175	81.034
Decision Tree	seq-labels	69.303	76.667
Naive Bayes	BoW + seq-labels	71.349	78.333
SVM Linear Kernel	BoW + seq-labels	76.652	82.759
Decision Tree	BoW + seq-labels	71.402	81.356
General Mixture Model BW (Approach A)	sentence labels	67.995	73.193
State-emission HMM BW (Approach A)	sentence labels	69.034	73.204
State-emission HMM BW (Approach B)	sentence labels	78.230	82.456
State-emission HMM Labeled (Approach B)	sentence labels	78.230	82.456
State-emission HMM 2^{nd}-Order (Approach B)	sentence labels	**79.599**	**83.051**
HMM 2-gram (Approach B)	sentence labels	78.432	82.531

2 Model and Plan

Our preliminary plan is to adapt existing HMM-based sentiment analysis methods for movie and customer reviews to the Steam review dataset. We will begin with a small amount of training data and gradually increase it as we progress. We will use the NLTK library to perform preprocessing and feature extraction, including removing stop words, stemming, and tokenization.

First, we will assume that the previous word will have some degree of influences on the next word in order to satisfies the Markov properties. Next, we will then design the HMM with hidden states that represent sentiment and observation states that represent the words in the reviews as shown.



, where q_i is the hidden state denotes positive or negative sentiment. And o_i be the corresponding words.
Next we will randomly create many sequences of these chain per review and use the Viterbi algorithm to decode the most likely sequence of hidden states for each review. Then, the overall sentiment of each review, will be the greater proportion of the sentiment according to the hidden states in the most likely chain.

We anticipate using Python as the programming language. The computational considerations we expect to encounter include memory constraints and computational time depends on the number of chains per review.

We have collected a dataset of Steam reviews, via Steam review in kaggle.com by Larxel, that contain both text and corresponding sentiment labels, and we plan to use this dataset for training and evaluation.

Finally, we will evaluate the performance of our model using metrics such as accuracy, precision, recall, and F1 score. We will also vary the number of chains per review and observe how the performance of the models. Furthermore, as time permits, we will change the hidden states and study the suitability of each of the models on Steam review.

3 Conclusion

In this project, we aim to develop an HMM-based sentiment analysis model for Steam reviews. Our approach will involve adapting existing methods for movie and customer reviews and gradually increasing the amount of training data. We will use Python for development and evaluate our model using metrics such as accuracy, precision, recall, and F1 score. We anticipate encountering computational considerations such as memory constraints and computational time. By the end of this project, we hope to provide insights into user preferences and satisfaction levels on Steam and help game developers improve their games.

References

[1] Isidoros Perikos & et al.(2019) *The Hidden Markov Models for Sentiment Analysis in Social Media* , Honolulu, Hawaii.