Computer Science in Collective Intelligence 3

Sentiment Analysis for Product Review: A Case of Video Games Startup

Instructor: Dr. Manal EL AKROUCHI

Group members:

Ghizlane Goubraim

Kawtar Oukhouya

Chyamae Bennouri

University Mohammed VI Polytechnic - 2023/2024

Table of content

Introduction

Natural Language Processing (NLP) is a highly engaging field with visible applications in various aspects of our daily lives. NLP projects have already permeated our surroundings, contributing to the advancement of technology. Notable examples are conversational agents (Amazon Alexa), sentiment analysis (Hubspot's customer feedback analysis feature), language recognition and translation (Google Translate), spelling correction (Grammarly), and much more.

In this project, we will focus on sentiment analysis, which is the task of determining the emotional value of a given expression in natural language. Sentiment analysis has applications in a wide variety of domains, including analyzing user reviews, tweet sentiment, etc.

These are some examples:

- **Movie reviews:** analyzing online movie reviews to get insights from the audience about the movie.
- **News sentiment analysis:** analyzing news sentiments for a particular organization to get insights.
- **Social media sentiment analysis:** analyze the sentiments of Facebook posts, Twitter tweets, etc.
- **Online food reviews:** analyzing sentiments of food reviews from user feedback.

Literature review

(Fang & Zhan, 2015); In this paper, they tackled the problem of sentiment polarity categorization, which is one of the fundamental problems of sentiment analysis. A general process for sentiment polarity categorization is proposed with detailed process descriptions. Data used in this study are online product reviews collected from Amazon.com.

Nowadays, people are relying on online products, and the importance of reviews is increasing. When selecting a product, a customer needs to go through thousands of reviews to understand the product. But in this prosperous day of machine learning, going through thousands of reviews would be much easier if a model were used to polarize those reviews and learn from them. **(Johar & Mubeen, 2020)** used a supervised learning method on a large-scale Amazon dataset to polarize it and get satisfactory accuracy.

Understanding the emotions and experiences of video game players is complex. As a first step, players can share their experiences with video games by writing reviews. By investigating these reviews, the emotions, experiences, concerns, and opinions of players can be understood. The goal of **(Guzsvinecz & Szűcs, 2023)** is to provide a deeper insight for video game developers to understand the players' emotional reactions regarding top-level genres. Besides, as subgoals, it can be observed whether a connection exists between word numbers and review type (positive/negative); whether there is a difference between the types of emotions in the case of different genres; or whether the emotional valence changes during writing a review. In this paper, overall 35, 983, 481 reviews of 11 top-level video game genres are studied using natural language processing methods. The results show that people write negative reviews earlier than positive ones, and no correlation exists between the time of review and their word number. A connection can also be observed between word numbers and whether a review is positive or negative: the median review length is 40 words in the case of negative reviews, while it is 19 words in the case of positive ones.

Motivation

According to the latest data, there are approximately 3.32 billion active video gamers worldwide(Appendix 1). Yet, SuperScale's "Good Games Don't Die" report reveals that a significant 83% of launched mobile games face failure within three years, while 43% are halted during the development phase. Most games fail solely because they haven't done pre-production properly.

Analyzing sentiments in user reviews of similar video games can provide valuable insights into enhancing products, addressing concerns, and predicting success. Understanding the features that contribute positively or negatively to player satisfaction is a key factor in improving the accuracy of new game development and increasing the likelihood of success.

Idea/Project description

Business case

As part of the Growth Hacking Team of a freshly launched startup that is introducing a new video game to the market, one of the key targets of a growth hacking team is to enhance the massive growth of early startups in a short time. To do so, the team introduces strategies to acquire as many customers as possible at the lowest possible cost. The goal of the team is to map the field of the video game market and find out how customers evaluate competitors' products, namely what they like and dislike in a video game. Knowing what makes a video game attractive to a gamer helps the startup improve its products, and the marketing team can articulate the product's message more effectively.

To achieve this task, we will employ NLP methods to get a deeper understanding of customer feedback and opinions.

Data Acquisition

For this project, we will Stream the video games dataset, which can be found on this website (Appendix 2).
The dataset contains the review text with user recommendations and other information related to each game for 64 game titles.

About Data Source: Steam Platform

- train.csv

review_id --> Unique ID for each review

title --> Title of the game

year --> Year in which the review was posted

user_review --> Full Text of the review posted by a user

user_suggestion --> (Target) Game marked Recommended(1) and Not Recommended(0) by the user

- game_overview.csv

title --> Title of the game

developer --> Name of the developer of the game

publisher --> Name of the publisher of the game

tags --> Popular user-defined tags for the game

overview --> Overview of the game provided by the publisher.

Data Cleaning

The first step is to clean the dataset in order to prepare it for our machine-learning model. We are going first to convert our textual data from its raw form into a numerical representation so that our model can make sense of it. However, before doing this, we need to remove all the noise from the data.

Spelling corrections: We are dealing with used reviews on a platform. Spelling mistakes are very common in that case. To overcome this problem we are considering creating a dictionary of the most common mistype words and replacing these common mistakes with the correct word to make sure this will not bias our analysis in the coming steps.
Special characters: These special characters add no value to text understanding and induce noise into algorithms. We are going to use regular expressions to get rid of these characters.

Data Preprocessing

Tokenization: Tokenization is the process of segmenting the text into a list of tokens; for our case, we will tokenize each word of user review. This process involves breaking down the review text into individual words.

Lemmatization: To reduce words to their base form while keeping the meaning, which can help reduce the vocabulary size and simplify the text, we will use lemmatization.

Normalization: We are going to convert everything to lowercase.

Stop words: Stop words are commonly occurring words in comments such as “the”, “and”, “a”, “an”, etc. We are considering removing them because they do not carry much meaning and can cause noise in the data.

Feature Engineering

The representation of features in a way that makes the computer understand text has helped NLP grow. However, Various techniques exist for feature representation, each offering distinct advantages. In our specific context, we have opted for the word embedding approach, a method that holds several key benefits. Word embeddings involve representing words as numeric vectors, allowing words with similar meanings to share common representations. This not only facilitates a more nuanced understanding of semantic relationships but also condenses the representation of words into a lower-dimensional space, aiding computational efficiency.

One noteworthy advantage of word embeddings is their ability to approximate meaning effectively, contributing to enhanced generalization and performance, particularly when confronted with limited training data. This becomes especially pertinent in real-world applications where data availability might be constrained. Additionally, word embeddings align seamlessly with the training of deep learning models, such as Long Short-Term Memory Networks (LSTMs), which have demonstrated significant success in NLP tasks like sentiment classification.

Model Building

Our objective is not only to classify the user reviews to distinguish the bad from the good games but to extract the most relevant information and give recommendations to the Growth Hacking Team. To achieve this goal we will use Long short-term memory (LSTM).

Long Short-Term Memory is an improved version of the recurrent neural network designed by Hochreiter & Schmidhuber. A traditional RNN has a single hidden state that is passed through time, which can make it difficult for the network to learn long-term dependencies. LSTMs address this problem by introducing a memory cell, which is a container that can hold information for an extended period. LSTM networks are capable of learning long-term dependencies in sequential data, which makes them well-suited for tasks such as language translation, speech recognition, and time series forecasting. LSTMs can also be used in combination with other neural network architectures, such as Convolutional Neural Networks (CNNs) for image and video analysis.

The memory cell is controlled by three gates: the input gate, the forget gate, and the output gate. These gates decide what information to add to, remove from, and output from the memory cell. The input gate controls what information is added to the memory cell. The forget gate controls what information is removed from the memory cell. And the output gate controls what information is output from the memory cell. This allows LSTM networks to selectively retain or discard information as it flows through the network, which allows them to learn long-term dependencies.

The advantages of LSTM (Long-Short Term Memory) are as follows:

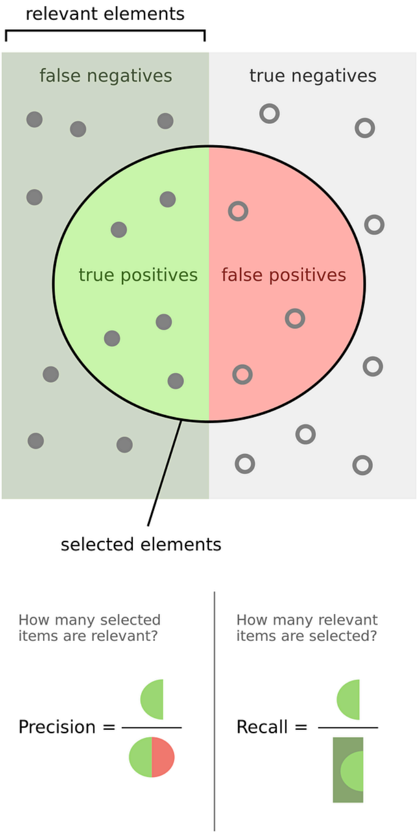
- Long-term dependencies can be captured by LSTM networks. They have a memory cell that is capable of long-term information storage.
- In traditional RNNs, there is a problem of vanishing and exploding gradients when models are trained over long sequences. By using a gating mechanism that selectively recalls or forgets information, LSTM networks deal with this problem.

- LSTM enables the model to capture and remember the important context, even when there is a significant time gap between relevant events in the sequence. So where understanding context is important, LSTMS are used. eg. machine translation.

Evaluation

we will evaluate the model by using different metrics so that we can look at the three main performance metrics:

- **Accuracy**: Refers to the percentage of the total predictions our model makes that are completely correct.
- **Precision**: Describes the ratio of true positives to true positives plus false positives in our predictions.
- **Recall**: Describes the ratio of true positives to true positives plus false negatives in our predictions.



Deployment

This is the last step for finalizing our project. We are considering deploying our model on an Open-Source ML Platform, leveraging TensorFlow for its flexibility and control while ensuring compliance. while also serving as an Ideal platform for projects requiring customization, control, and low costs, and providing access to the latest innovations, and community support.

Results/Discussion

The model for understanding the emotion of the video game review from the text of the reviewer is built using LSTM which classifies the reviews very well. This strategy will be useful to improve the video game by getting to know which content has a good reception by the audience. Our results will show that the Long Short Term Memory Networks algorithmic standard beats others as far as exactness.

We are proposing a sentiment classification model that helps in classifying video game reviews based on the emotion of the sequence text data. This system is different and efficient from others as it catches up in classifying long sequence data with the help of LSTM which makes use of long-term memory and thus can handle the long-term dependencies very effectively. By using larger datasets we can still improve the performance of the model. In the future examination, we will in general complete classifying video games not just reviews by using further analysis of the internal results of each review corresponding to a game movie. We tend to additionally expect to improve the execution of both video game classification techniques and video game review classification models. By integrating these two models into one we can create a model with which we can classify a collection of video games by classifying the reviews and each video game and further analyzing the results we can classify video games. Thus by passing a set of video games, we can classify them all at a time.

limitations

The incorporation of dictionaries, which encompass emojis, abbreviations, and slang, introduces a significant challenge in comprehending and extracting meaningful insights from user comments. The informal and dynamic nature of language employed by users creates a noisy dataset, further complicated by the prevalence of abbreviations, emojis, and typographical errors. These linguistic nuances pose a formidable obstacle in preprocessing the data for machine learning models. Deciphering the intended meaning behind slang and understanding the context of emojis becomes a complex task. Furthermore, the abundance of abbreviations, often specific to online discourse, necessitates careful handling to avoid misinterpretation or loss of critical information. The need for a nuanced and adaptive approach in processing such diverse linguistic elements is paramount to ensuring the accurate extraction of insights from user-generated content.

Another notable limitation lies in the heterogeneous use of multiple languages within a single user comment. Users frequently intermingle languages, creating a mixed-language text that adds a layer of complexity to language processing tasks. This multilingual aspect can confound traditional language models, as they may struggle to seamlessly switch between languages and accurately interpret the intended meaning. Managing such mixed-language texts requires specialized techniques for language identification, translation, and context preservation. Failure to address these challenges may result in misinterpretations, leading to inaccurate analyses and insights. Consequently, the incorporation of robust mechanisms to handle mixed-language content is essential for enhancing the efficacy of natural language processing approaches in the context of user comments and online discourse.

Foresight

To foresee the future of the video game startup - and considering the dynamic nature of the gaming industry and the broader business environment - we will use scenario planning that is ideal in the gaming startup context due to the industry's rapid changes. When looking at how fast-evolving the gaming industry is, 5 years seems to be a reasonable time frame for our foresight.

Scenario 1: Gaming can become more popular, which can help the startup's gaming products get famous because people will love them as a result of the startup's learning from the product review analysis.

➔ The startup will need to make more products quickly to keep up with all the new customers. While always keeping in mind that it is not only about quantity but also quality and making sure that customers have a good experience.

Scenario 2: While the video gaming industry keeps growing, tons of companies will start to make and sell video games, which can make it a bit difficult for the startup to stand out among all the competition.

➔ The startup will need to find a way to make its products very special. They will need to think about what can make them unique and different while also using creative ways of marketing.

Scenario 3: Considering the development of AI, virtual reality, and holograms... A big change in gaming technology might happen. (e.g.: everyone playing games in virtual reality). By then, people will start liking products that use new technology.

➔ The startup should keep an eye on new tech trends and be ready to change its products based on what's new and exciting. The startup can develop a research committee that will be responsible for updating the team with the new technologies.

References

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal Of Big Data*, *2*(1). https://doi.org/10.1186/s40537-015-0015-2

Guzsvinecz, T., & Szűcs, J. (2023). Length and sentiment analysis of reviews about top-level video game genres on the steam platform. *Computers In Human Behavior*, *149*, 107955. https://doi.org/10.1016/j.chb.2023.107955

Johar, S., & Mubeen, S. (2020). Sentiment analysis on large scale Amazon product reviews. *IJSRCSE*, *8*(1), 7-15.

Appendix

https://explodingtopics.com/blog/number-of-gamers

https://www.kaggle.com/datasets/piyushagni5/sentiment-analysis-for-steam-reviews