Machine Learning Project
Project Proposal

Joëlle Bosman (s3794717)

Larisa Bulbaai (s3651258)

Wessel Poelman (s2976129)

Machine Learning project
Information Science
March 18, 2021

INTRODUCTION

In recent years, the problem of "sentiment classification" has been increasing attention (Kalaivani and Shunmuganathan, 2013). According to Kalaivani and Shunmuganathan (2013) large amounts of reviews and ratings can give important information for companies to market their product. The paper states that sentiment analysis helps monitor the public's mood about a particular product, service or object. This would prove to be helpful in assessing the success of a newly launched product, how a new version of a product is received by the customer and how the likes and dislikes are limited to a particular area.

For our research we will use an open dataset of Disneyland reviews[1]. Disneyland is a popular chain of theme parks that attracts a lot of visitors (Smith, 2020). Some of these visitors write a review on Trip Advisor in which they give their visit a rating and write a short story about their experiences. Other sources that do not yet have a rating could also be useful for analyzing how customers feel about the parks. When the reviews with ratings are used to train a model, reviews without ratings can be classified so a rating can be attached to them. This results in being able to combine reviews from multiple sources into the same rating system, which makes further analysis easier. The following research question will be used in the project: "How can textual reviews be categorized into the labels *negative*, *indifferent* or *positive*?"

BACKGROUND

The paper from Basari et al. (2013) uses support-vector machine (SVM) to classify opinions. According to this paper the SVM is a classifier that makes the greatest accuracy outcomes in text classification issues. The paper discusses SVMs using N-grams and feature weighting (TF and TF-IDF). Additionally, Se et al. (2016) states that opinions categorized into positive and negative are helpful in many fields. In this paper support-vector machines have been successfully used in language processing. Moreover, Kalaivani and Shunmuganathan (2013) studied online movie reviews using the following sentiment analyzing approaches: Naive Bayes, SVM and KNN. For our own research on classification we will be using a SVM and K-Nearest Neighbor (KNN).

The KNN algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve classification problems (Harrison, 2018). Baid et al. (2017) uses this method among other methods to identify the polarity of tweets. According to this paper KNN is the simplest of all machine learning algorithms. Because of this we will try out this method for our own research to discover if it gives a high accuracy score for our data.

MATERIAL

The dataset contains a little more than 42.000 reviews of three Disneyland parks, namely Disneyland Paris, Disneyland California, and Disneyland Hong Kong. The data in this collection has been taken from Trip Advisor.

For each review, the following information is available: the identification number of the review, when this review was posted on Trip Advisor, what the nationality of the visitor was, which of the three parks was visited, and a rating for the park. These ratings range from 1 to 5. If a visitor gives a 1, he or she is very dissatisfied about the visit; when a 5 is given, the visit was magnificent. Additionally, there is a

[1] https://www.kaggle.com/arushchillar/disneyland-reviews

review text in which the visitor writes a short story about his or her experiences of the visit to the park.

For our research, we only use the ratings and the review texts. Based on the text, we want to predict the rating the writer would give to the visit. Of course, these ratings are subjective and therefore, the difference between the various scores might not be very clear. Therefore, we split the data into the following three categories: negative (1, 2), indifferent (3) or positive (4, 5).

METHODOLOGY

The predicting will be done using the text of the review. There are multiple options to choose from when 'converting' text to machine learning features. The main option we want to try is creating TF-IDF vectors as input for the SVM classifier. Another option we want try is using an existing word-embedding model to convert the text to word-embedding vectors and using those as input. The main consideration for the embedding approach is the language of the reviews since an existing model for a certain language needs to be used. Luckily, all reviews in this dataset are in English, so this is not a problem.

The TF-IDF approach is more straightforward and most likely 'good enough', while the embedding approach is a bit more complicated, but could capture a lot more semantic information. This last point might have a noticeable effect on the performance of the system, this is something to experiment with. The reviews are not very long, which could be a detrimental for the embedding approach. The whole process will follow this rough outline: (1) Clean and normalize the text (lower, remove punctuation, tokenization, lemmatization, optionally remove stopwords), (2) Convert tokens to vectors (either TF-IDF or word-embedding vectors), (3) Train model (SVM or KNN), (4) Test and validate model.

These steps are subject to change and in the final project other steps could be taken to improve the accuracy of the model. This list is a starting point, other and more specific optimization techniques, (e.g. feature selection or parameter optimization) will probably also be used, but these are hard to define at this moment. These are quite dependent on the dataset and will probably become clear while working on the listed steps.

For every review in the dataset, the rating is known, this means that we have to create our own test set. This will be done by randomly selecting a portion of the dataset and removing the rating. This portion will either be 10% or 20%, this is something we will experiment with.

EVALUATION

For the evaluation the performance of our two models we will use a similar approach as Kalaivani and Shunmuganathan (2013). We will use the test set to evaluate how well our models predict the ratings. The evaluation of each model will be based on the accuracy, precision and recall. The accuracy is the overall accuracy of the model. Recall (positive) and Precision (positive) are the ratio and precision ratio for true positive reviews. Recall (negative) and Precision (negative) are the ratio and precision ratio for true negative reviews.

The model with the best score will be our final model for the classification of the movie reviews. We strive to achieve an accuracy of over 60%, which we believe must be possible for a SVM model since this has been achieved by Kalaivani and Shunmuganathan (2013); Basari et al. (2013); Se et al. (2016).

BIBLIOGRAPHY

Palak Baid, Apoorva Gupta, and Neelam Chaplot. 2017. Sentiment analysis of movie reviews using machine learning techniques. *International Journal of Computer Applications*, 179(7):45–49.

Abd Samad Hasan Basari, Burairah Hussin, I Gede Pramudya Ananta, and Junta Zeniarja. 2013. Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. *Procedia Engineering*, 53:453–462.

Onel Harrison. 2018. Machine learning basics with the k-nearest neighbors algorithm.

P Kalaivani and KL Shunmuganathan. 2013. Sentiment classification of movie reviews by supervised machine learning approaches. *Indian Journal of Computer Science and Engineering*, 4(4):285–292.

Shriya Se, R Vinayakumar, M Anand Kumar, and KP Soman. 2016. Predicting the sentimental reviews in tamil movie using machine learning algorithms. *Indian Journal of Science and Technology*, 9(45):1–5.

Craig Smith. 2020. How many people visit disney parks each year?