Sentiment analysis based on Women E-commerce store reviews dataset

Project report

Prepared by: Claudia Słaboń

Analyzed problem

Sentiment analysis based on review and rating given by customer.

Introduction

As it is crucial for business to have happy customers companies need to monitor the quality of their product. Based on the historical reviews and rating given by client it is possible to train the model which will discover customer's sentiment based on review. If the business discover that its clients are unsatisfied management staff can investigate and take actions in order to improve the product and service.

Dataset description:

The dataset is a real commercial data taken from anonymized women's clothing e-commerce store. It contains 10 variables and around almost 23 500 rows.

List of available variables:

1. Clothing ID
2. Age: age of the customer
3. Title: title of the review
4. Review Text: text of the review
5. Rating: score from 1 (the worst), to 5 (the best).
6. Recommended IND: contains 2 values (1 if the product is recommended, 0 if not)
7. Positive Feedback Count: the number of other customers who found this review helpful.
8. Division Name: name of the product high level division.
9. Department Name: name of the product department name.
10. Class Name: name of the product class name.

However, for the purpose of this analysis only 2 variables are used (Rating and Review Text).

The aim of the model is to predict the customer's satisfaction based on review. Additionally, the target variable has been transform in the following way:

- If rating is 1 or 2, then it's negative review,
- If rating is 3, then it's neutral review,
- If rating is 4 or 5, then it's positive review.

Description of the models:

1. Benchmark model

As a benchmark model Multinomial Naïve Bayes classifier has been chosen. It is a probabilistic machine learning model, which is widely used in case for document classification problem. Firstly, the data for the benchmark model were transformed using count embedding.

2. Student's model

As a student's model LSTM model has been chosen. In order to predict the final rating 2 models' results were combined using average of their predictions.

First LSTM model (GloVe approach)

First LSTM model contained words' weight obtained from GloVe (Global Vectors for Word Representation) pre-trained model which can be downloaded from the following link (https://nlp.stanford.edu/projects/glove/). GloVe method is an unsupervised algorithm which aims to obtain vector representations for words.

First LSTM model (Word2Vec approach)

Second LSTM model contained words' weight obtained from Word2Vec model which was trained on the dataset. This model combines 2 approaches: continuous bag-of-words (CBOW) and skip-gram embedding.

Optimized parameters:

- Number of hidden layers (model contains 1, 2 or 3 hidden layers)
- Dropout value (parameter was equal to: 0.2, 0.3, 0.4 or 0.5)
- Number of hidden neurons (depending on the layer parameter was equal to: 64, 32 or 16)
- Learning rate for Adam optimizer (parameter was equal to: 0.01, 0.001 or 0.0001)

Above values gave 36 LSTM models using GloVe approach and 36 LSTM models using Word2Vec approach. The final 2 models were chosen based on the highest validation accuracy.

Additionally, if there was no improvement of validation accuracy after 3 more epochs the model stopped training. That way, also number of epochs for every LSTM model was chosen.

Structure of the 2 LSTM models:

The best LSTM model with GloVe embedding layer had following parameters:

Number of hidden layers: 1

Dropout value: 0.4

Learning rate for Adam optimizer: 0.01

Additionally, the best number of epochs was equal to 5.

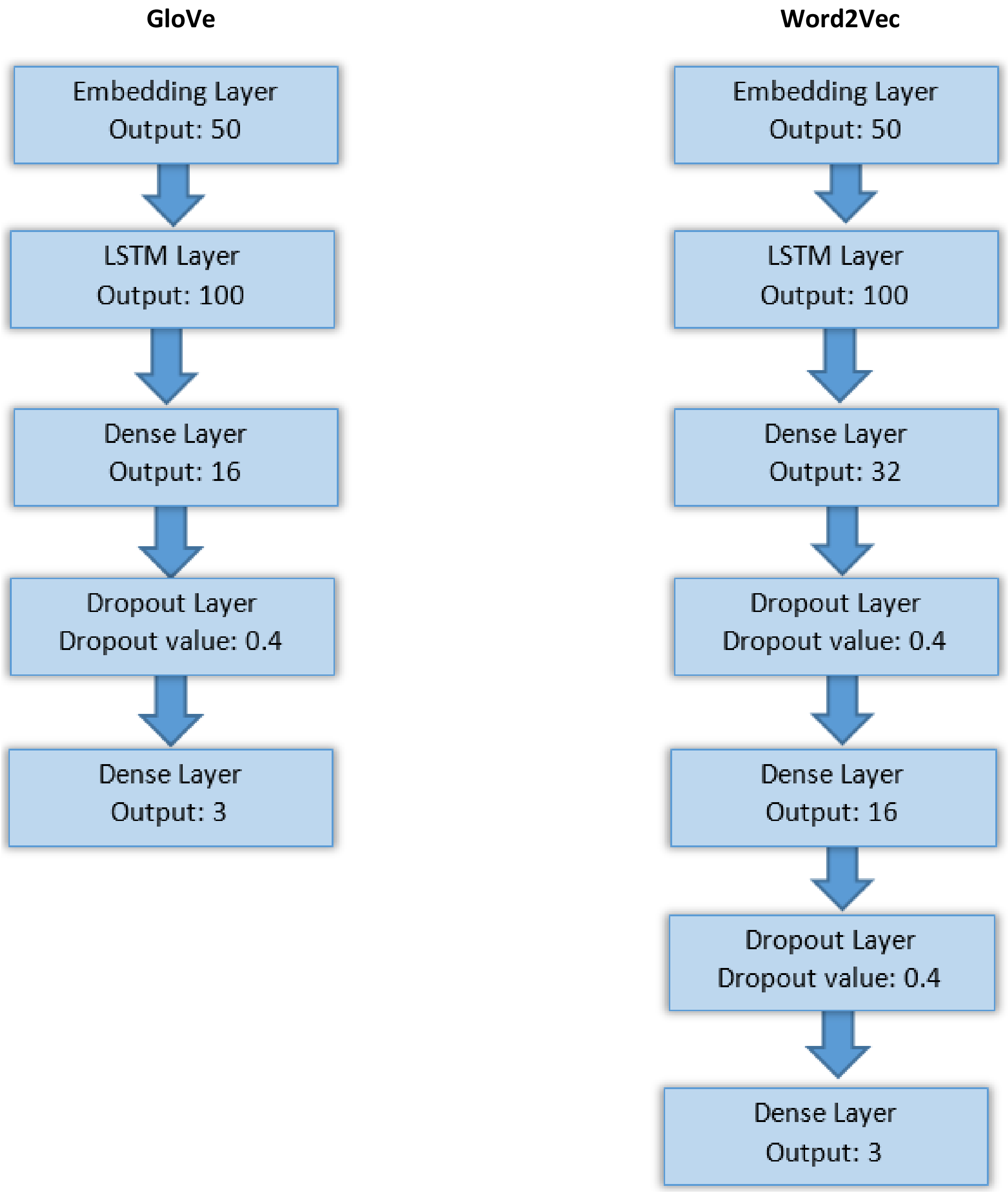
The best LSTM model with Word2Vec embedding layer had following parameters:

Number of hidden layers: 2

Dropout value: 0.4

Learning rate for Adam optimizer: 0.001

Additionally, the best number of epochs was equal to 13.



GloVe

Word2Vec

Conclusions:

The final values of evaluation metrics, which is accuracy in this case, are equal:

- Benchmark model: accuracy: 80,6%
- Student's model: accuracy 82,2%

Additionally, based on confusion matrix it can be observed that models have difficulties to predict neutral rating. Negative reviews' prediction results are also unsatisfied. It is in fact difficult to classify review as neutral as it contains negative, as well as, positive words. From the dataset review one can observe that in case of negative rating, reviews contain often positive words, which can be a problem for the model.

In order to improve results one can introduce more stop words and also focus on predicting 2 classes (negative and positive) rather than 3 (negative, neutral, positive).

References:

https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews