*Sentiment analysis of Amazon Alexa Review**
Machine Learning for Natural Language Processing
ENSAE 2021

Cheick Mohamed KABA
ENSAE 3A Data science and business decision
cheickmohamed.kaba@ensae.fr
Emmanuel KAMLA FOTSING
ENSAE 3A Data science and statistical learning
emmanuel.kamlafotsing@ensae.fr

Objective

The development of new information and communication technologies has given considerable impetus to the market of voice assistants led by web giants such as Apple with Siri, Google and its Google assistant as well, Cortana from Microsoft, etc. Amazon also designed in November 2014 a similar product, Alexa. As the voice assistant market is very competitive, it is a key marketing issue to evaluate consumer satisfaction following the use of these products in order to define or redirect strategies and thus gain market share. By using customer comments on the purchase pages of these products, it is humanly possible to analyze and evaluate the overall feeling. The goal of this project will be to implement NLP techniques to perform this task in an automated and unassisted way. For this, we use Alexal's review database which contains nearly 3150 Amazon consumer reviews(input text), scores, review date, variant and reviews of various Amazon Alexa products like Alexa Echo, Echo dots, Alexa Firesticks etc. to learn how to train a machine for sentiment analysis.

Thus, in order to obtain a better prediction of the feelings that emerge from the comments, several Machine Leanrning (SVM, RF and Regression Logistic) and Deep Learning(LSTM) algorithms are trained and compared.

This report is divided into three sections: the first section presents the descriptive analysis of the database and the data augmentation, the second section focuses on the preprocessing of the textual data and finally the modeling and the results are presented in the last section.

1 Descriptive Statistics and Data augmentation

In this section we will analyze the different variables of the dataset and then, if possible, perform some treatments.

1.1 Descriptive Statistics

The data used in this analysis comes from kaggle on Amazon's voice assistant product evaluation. This dataset contains five columns namely: rating, date, variations, verifiedreviews et feedback.

1. rating: is the score on a scale of 1 to 5 given by the consumer to the product,
2. date: corresponds to the date of the comment,
3. variation: different types of alexa products. There are 16 types of variation.
4. verifiedreviews: comment made by each consumer,
5. feedback: feeling of satisfaction. 1 if the consumer is satisfied and o otherwise.

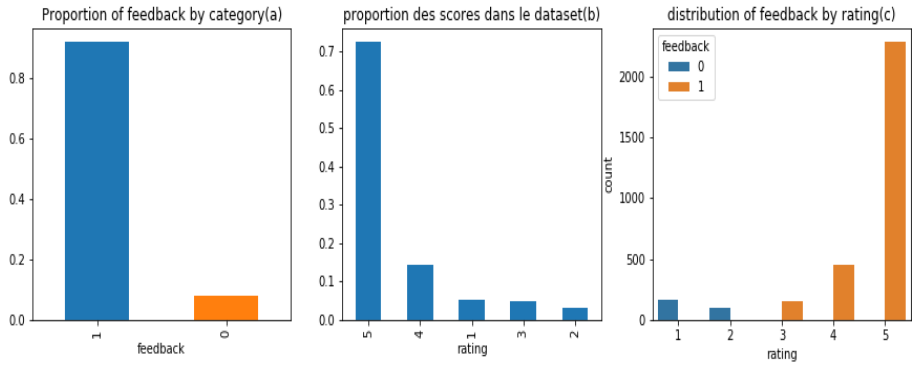


Figure 1:

The graph (a) represents the proportion of feelings by category. We notice an imbalance between positive feelings (represents about 93% of the dataset) and negative feelings (represents about 7% of the dataset). This imbalance of class clearly increases the difficulty of learning by the classification algorithm. Indeed, the algorithm has only a few examples of the minority class (negative feelings) to learn from. It is therefore biased towards the negative population and produces potentially less robust predictions than in the absence of imbalance.

The graph (c) shows a certain correlation between the target(feedback) and the rating variable since on the one hand, when the feeling is negative the rating is either 1 or 2 and either 3 or 4 or 5 when the feeling is positive on the other hand. Therefore, this variable will not be considered for the modeling.

1.2 Data augmentation

Several data augmentation techniques in the literature have been developed to address the problem of class imbalance in a dataset.

Remplacement de synonyme (SR): le remplacement de synonyme est une technique dans laquelle nous remplaçons un mot par l'un de ses synonymes.

Synonym Replacement (SR): Synonym replacement is a technique in which we replace a word with one of its synonyms.

Random swapping (RS): In random swap, we randomly swap the order of two words in a sentence.

Random insertion (RI): Finally, in random insertion, we randomly insert the synonyms of a word at a random position.

2 Preprocessing

2.1 First step : Train and test

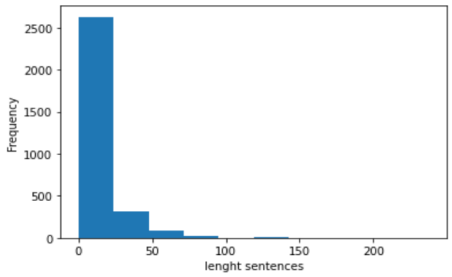
We start by splitting our data in train and test. Then we augment train that in order to balance the dataset. We use 70% of the data to train our models and the rest of the 30% to evaluate them.

2.2 Second step : Cleaning and tokenizing

The cleaning process consists in removing all non-informational word in the corpus(punctuation, numbers, special characters, stopword etc.). Then We use part of speech tagging and WordNetLemmatizer to tokenize.

2.3 Third step : Embedding with word2vec

Rather than using a pre-trained model of Word2vec we choose to fit it on our corpus. We use 100 as length for each vector. Then we use 2 types of transformation. One type for the sequential lstm and another type for non sequential machine learning models that we perform. For the sequential transformation we fix the maximal number of the sequence to 30 based on the distribution of sequence length in our dataset. For the non-sequential transformation the vector obtained is computed as the mean of the vectors of the sequence.



3 Results

This section presents the results obtained from the different analyses conducted. The four models listed in the previous section were trained on 70% of the data and tested on 30% of the data.

3.1 Model performance

The performance of the different models was evaluated based on the different predictions made by the models for comment. Table 1 shows the precision, recall, F1-score and accuracy of each model.

Table 1: Performance des modèles

Models	precision	accuracy	recall	F1-score	AUC
Logistic Regression	0.97	0.9	0.9	0.94	0.92
Support vector machine	0.98	0.87	0.87	0.92	0.91
RandomForest	0.98	0.95	0.97	0.97	0.96
LSTM	0.98	0.92	0.94	0.96	0.97

We can see from table 1 that all models perform very well in terms of accuracy and F1-score, and in terms of accuracy the LTSM and the Random Forest perform the best as well as in terms of recall. However, in terms of AUC the LSTM model outperforms all other models.The following curve shows the ROC of each model.

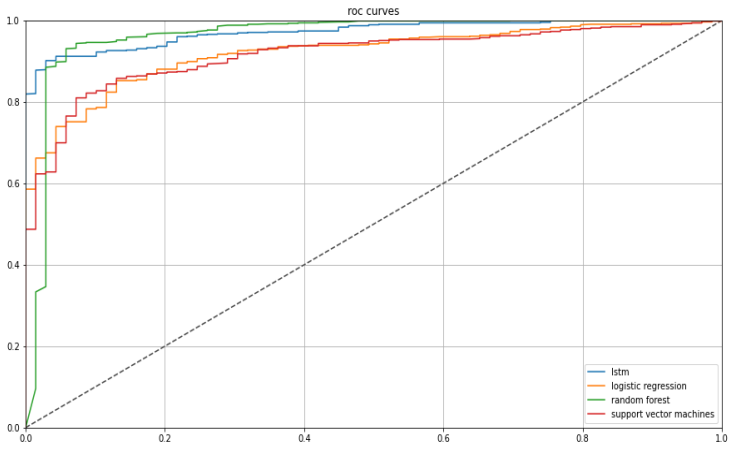


Figure 2: ROC curves of the models

3.2 Qualitative Evaluation

For this qualitative validation we gave examples of reviews and then evaluated by our different models the probability of being qualified as good.

	test_review	lstm	svc	random forest	logistic regression
0	I love it	0.999988	1.000000	1.000000	0.999999
1	The best i have ever use	0.992426	0.999987	0.843333	0.998276
2	I hate this product	0.531311	0.137467	0.646667	0.055492
3	It absolutely change my life	0.991650	0.997113	0.800000	0.997979
4	The product does not work very well	0.968885	0.743152	0.826667	0.691677

Figure 3: qualitative evaluate

We notice that for the sentence ”I hate this product”, all the models give a rather low probability compared to the other probabilities. The prediction of the logistic regression is rather good comparing to the other models.