TEXT ANALYTICS PROJECT

Alexa's Reviews

Group 03

Members

Naomi Esposito

- Data Science and Business Informatics

Silvia Cosmo

- Digital Humanities

Cristiano Ciaccio

- Digital Humanities

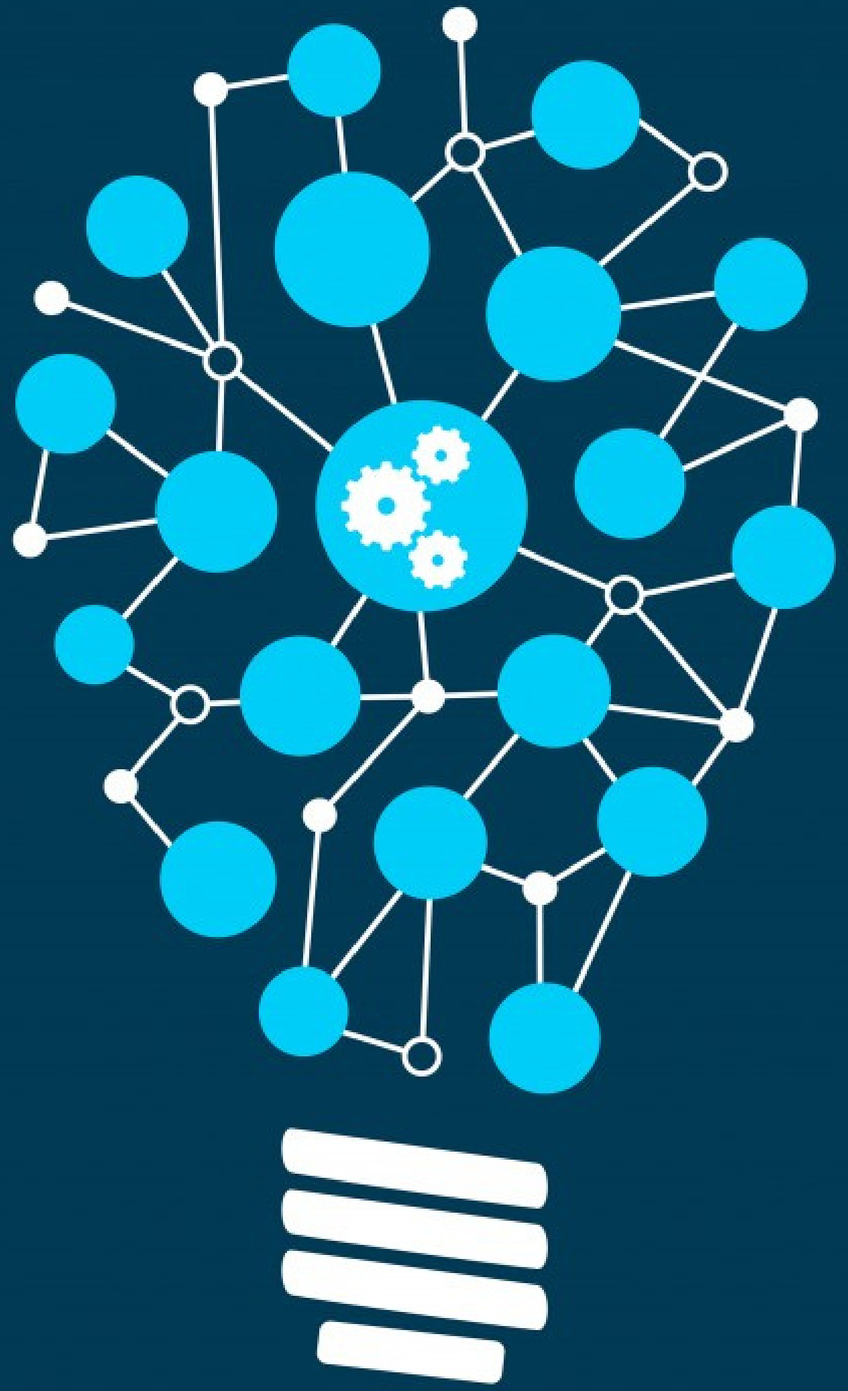
Alice Graziani

- Data Science and Business Informatics

Our goal:

Our goal is to use both supervised and unsupervised training techniques for sentiment analysis.

In particular, we want to quantify and evaluate the performance of the different methodologies and compare them with each other.



Supervised:

We will use BERT (Bidirectional Encoder Representations from Transformers), a deep neural network model that uses self-attention techniques and aims to capture as much semantic, syntactic and contextual knowledge as possible and use it to perform specific tasks related to language.

This machine learning model will be fine-tuned to correctly classify and predict the sentiment.

We will also test the performance of another classifier that weighs tokens through their probability, based on the functioning of the Naive Bayes association algorithm.

Unsupervised

Through sentiment and emotion Lexicons we will annotate the token of our dataset either with a polarity value or with an emotional category.
Then, we sum up the value of each token in order to obtain a score that will be compared with the result of the supervised method.

Moreover, through unsupervised methods such as clustering, we will try to extract interesting features for sentiment analysis and we will understand which are the most common features.

At the end, we will compare the obtained data with those of the classifier.

THE DATASET

The Alexa's Reviews dataset contains 3150 reviews referring to Alexa products extracted from Amazon's website.

Relevant features for each review are:

- feedback: goes from 0 to 1, that is negative and positive.
- rating: the score given by a user together with the review, it goes from 1 to 5.

Reviews with a feedback of 0 have a rating of 1 or 2. Meanwhile, reviews with a feedback of 1 have a rating between 3 and 5.

THE DATASET

The dataset is structured like this:

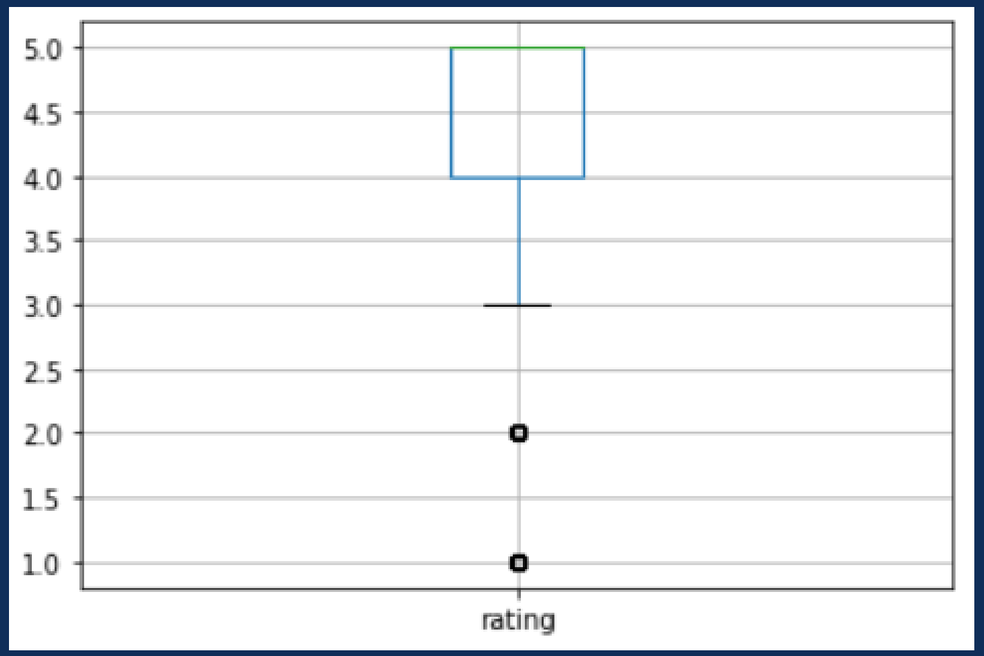
	rating	variation	verified_reviews	feedback
0	5	Charcoal Fabric	Love my Echo!	1
1	5	Charcoal Fabric	Loved it!	1
2	4	Walnut Finish	Sometimes while playing a game, you can answer...	1
3	5	Charcoal Fabric	I have had a lot of fun with this thing. My 4 ...	1
4	5	Charcoal Fabric	Music	1

The variation column describes the type of Alexa products.
The dataset contains no NaN values.

The following slides will go through a bit of data understanding:

THE DATASET

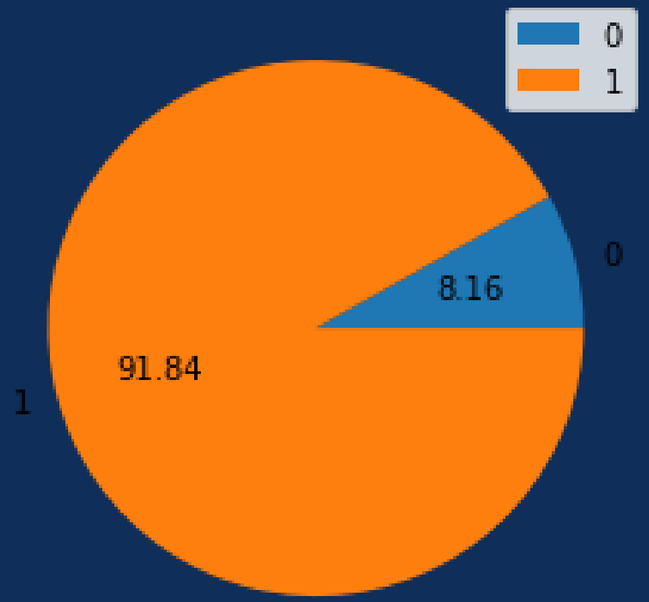
Ratings are distributed like so:



The majority of the ratings is distributed between 4 and 5. Ratings of 1 and 2 are seen as outliers...

THE DATASET

In fact, looking at the distribution of the feedback variable, we see that our data is really skewed towards positive reviews.



We expect our model, then, to have some problems in classifiyng negative reviews. Problems we will try to deal with.

THE DATASET

Looking at the dataset through the feedback variable we have:

- 2893 positive reviews with an average number of chars of 11.65
- 257 negative reviews with an average number of chars of 19.08, almost double!

And after some basic text cleaning (removing stopwords, punct., lemmatizing) we see that:

- For the positive reviews there are 33710 tokens and 3075 types (TTR: 0.091)
- For the negative ones there are 4904 tokens and 1318 (TTR: 0.268, probably due to the small size of the negative reviews corpus)

This confirms that, even after counting the number of tokens for each category, the data is really skewed towards positive reviews.

THE DATASET

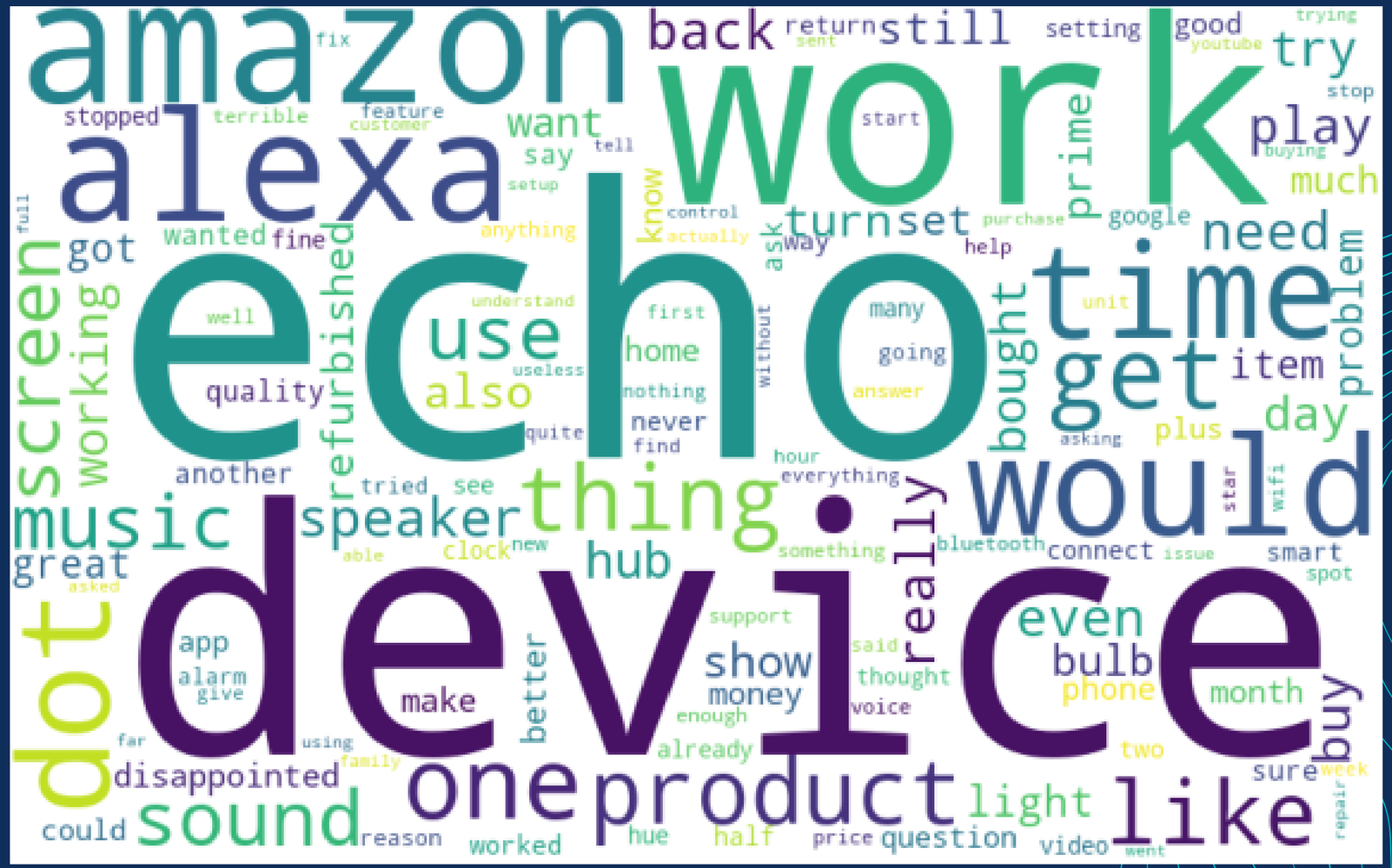
A wordcloud for the entire collection of reviews:

THE DATASET

A wordcloud for only the positive reviews (obviously similar to the total tokens one):

THE DATASET

A wordcloud for only the negative reviews:



ISSUES

Major issues are:

- The size of the dataset:
  - The dataset seems really short in comparison to other datasets, the lack of information could be a problem to train a classifier.

- The imbalance of the annotated sentiment:
  - The dataset is really skewed towards positive reviews, taking 92% of the entire feedbacks. Even if negative reviews are usually longer the data still remain extremely skewed

Can we overcome these issues?

Thank you for your attention