Sentiment analysis

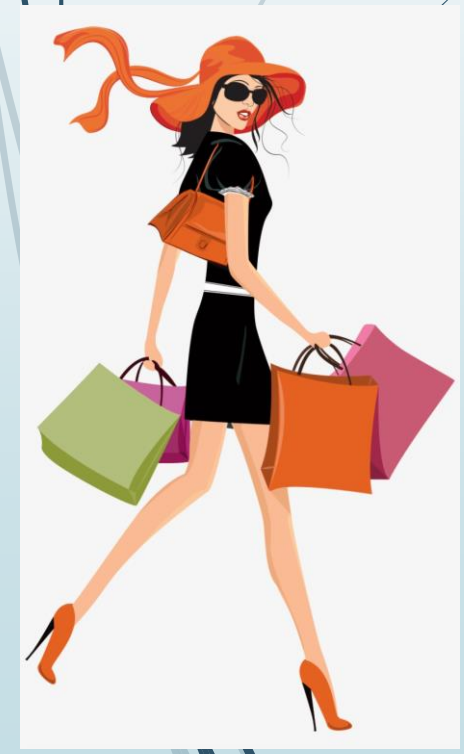
Based on Women E-commerce store reviews dataset

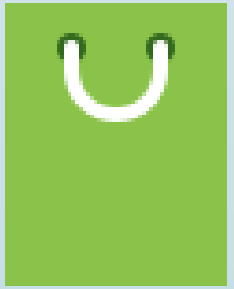
Author: Claudia Słaboń

Introduction

- Sentiment analysis based on review and rating given by customer.

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

- The dataset is real commercial data taken from anonymized women's clothing e-commerce store. It contains 10 variables (for example: age of customer, rating, text review, recommended, positive feedback).

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

- Variables needed for analysis:
  - Text Review
  - Rating

Positive Rating (4, 5)

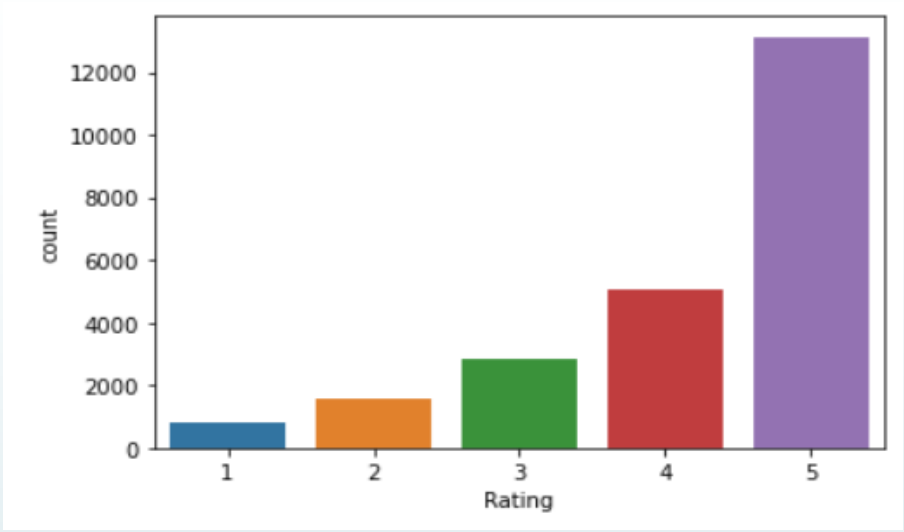
Neutral Rating (3)

Negative Rating (1, 2)

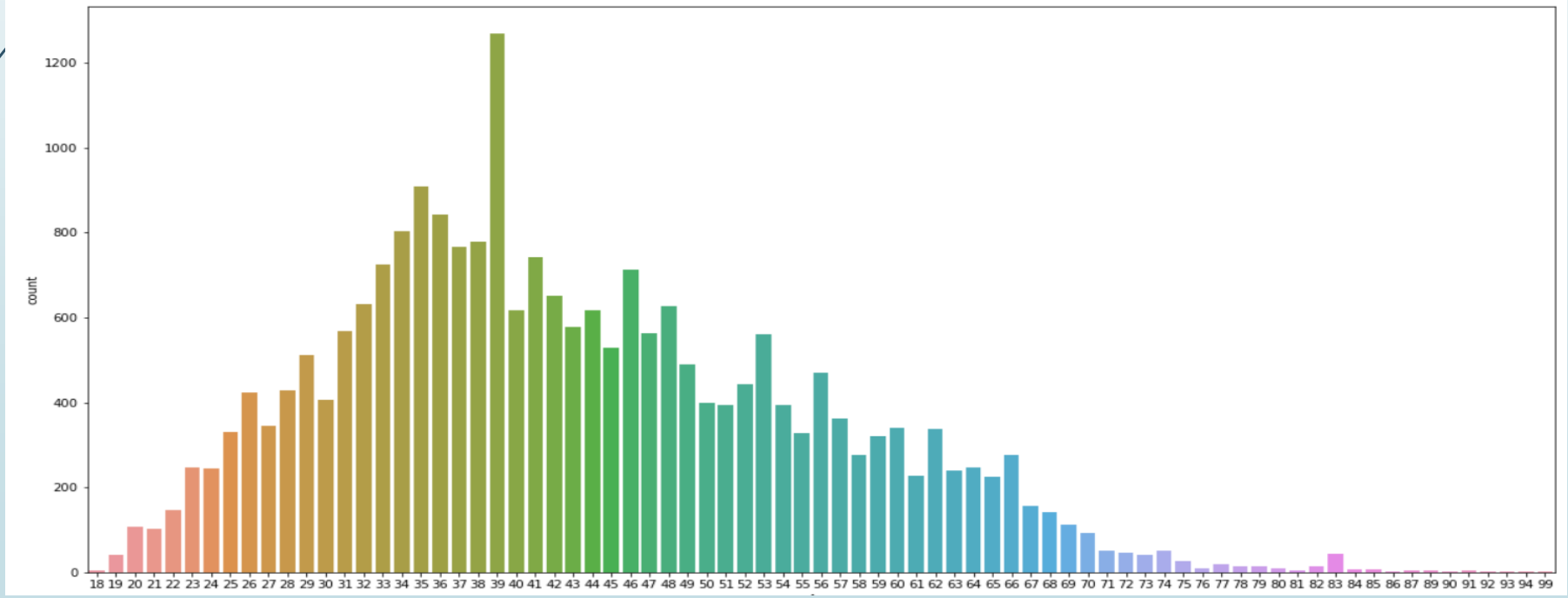
Positive review	Loved this top. great design. comfortable and unique. soft material
Neutral review	Dress is very pretty, but very short, almost tunic length.
Negative review	It looks like you are wearing cargo shorts. really unflattering. avoid buying this skirt

Dataset

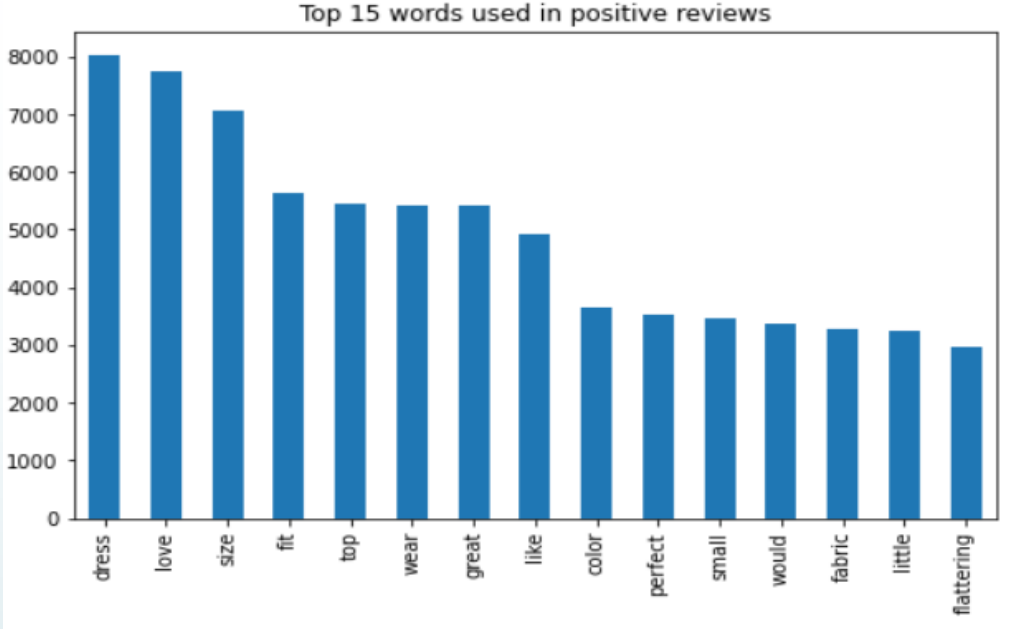
Ratings

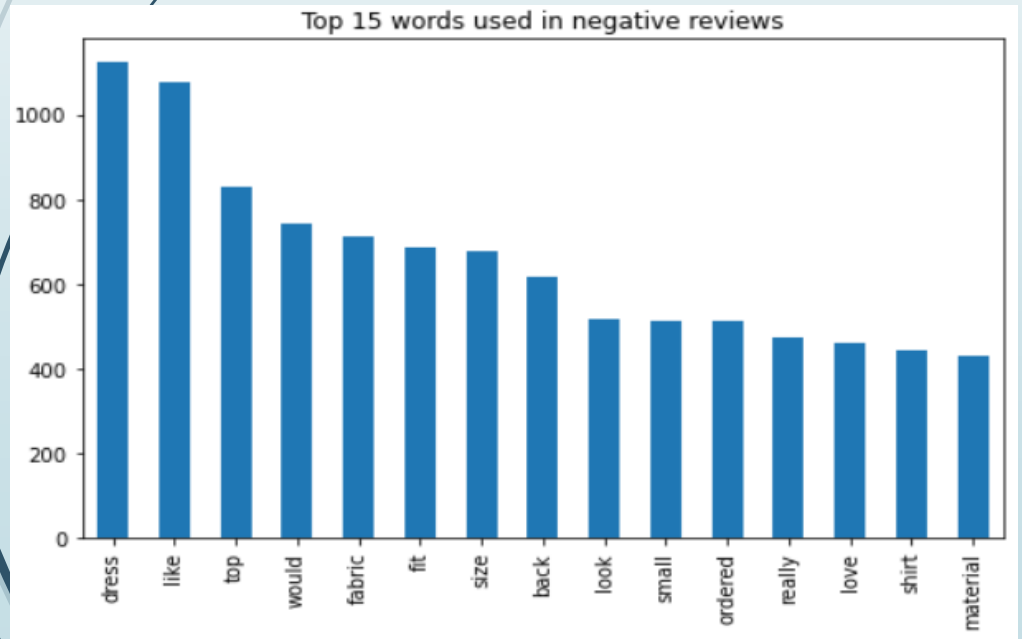


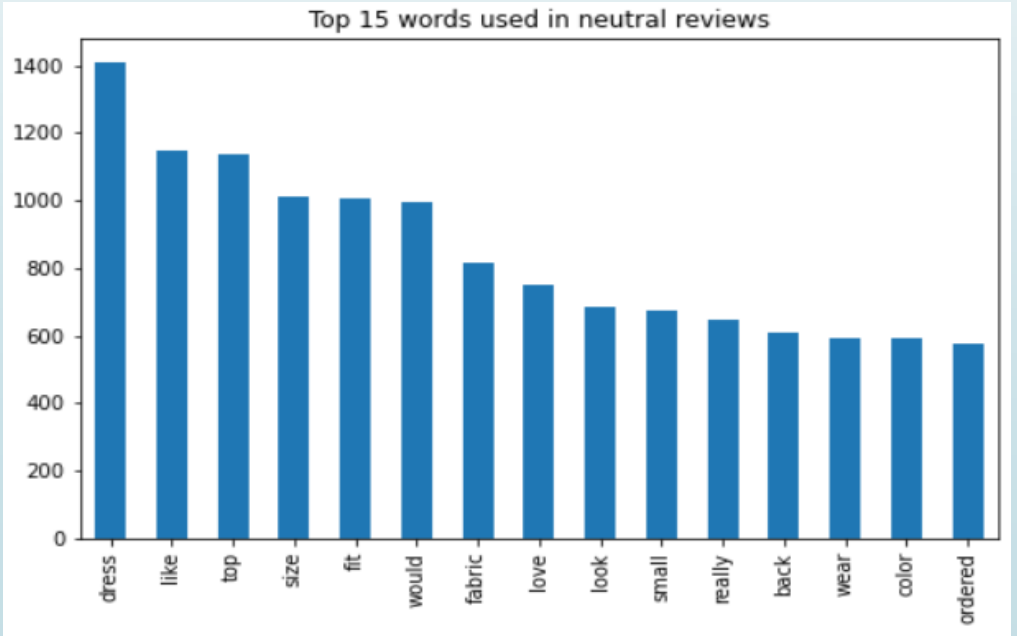
Age of customers




Total Reviews Word Cloud
skirt
shirt
comfortable
beautiful
see
pretty
tried
look
bottom
one
way
perfect
though
cut
make
cute
still
top
love
color
fabric
feel
wear
bought
long
style
sweater
nice
flattering
true
size
waist
great
design
jean
dress
think
even
sleeve
material
pant
got
really
much
short
work
well
little
jacket
back
fit


Top 15 words used in positive reviews
8000
7000
6000
5000
4000
3000
2000
1000
0
dress
love
size
fit
top
wear
great
like
color
perfect
small
would
fabric
little
flattering


Top 15 words used in negative reviews
1000
800
600
400
200
0
dress
like
top
would
fabric
fit
size
back
look
small
ordered
really
love
shirt
material


Top 15 words used in neutral reviews
1400
1200
1000
800
600
400
200
0
dress
like
top
size
fit
would
fabric
love
look
small
really
back
wear
color
ordered

Chosen models

- Benchmark: Multinomial Naïve Bayes classifier

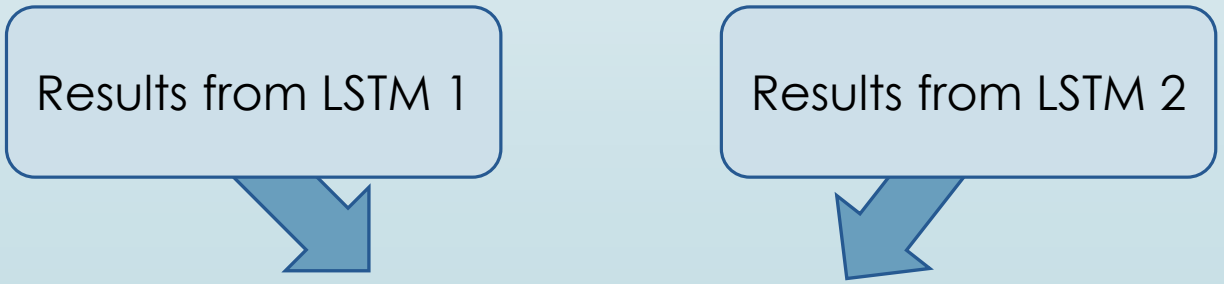
Firstly, text reviews transformed into count embedding.

Model uses multinomial distribution which describes the probability of observing counts among a number of categories

- Student's model – 2 approaches

LSTM 1st submodel (pre-trained GloVe method to create embedding matrix)

LSTM 2nd submodel (Word2Vec method to create embedding matrix)



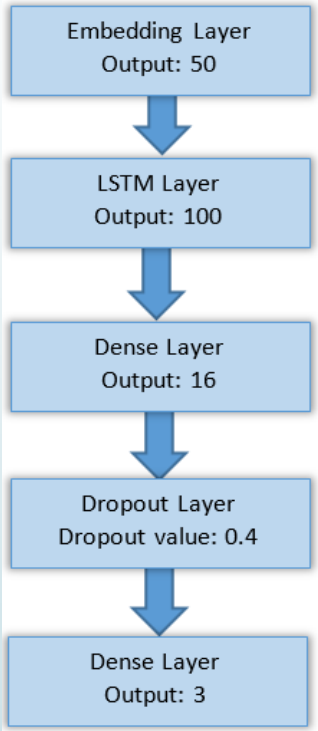
Optimized parameters

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

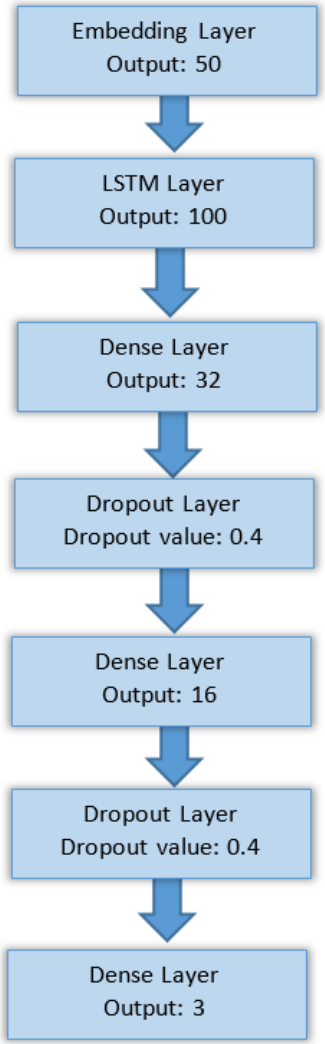
Additionally, if there was no improvement of validation accuracy after 3 more epochs the model stopped training. That way, also number of epochs for every LSTM model was chosen.

LSTM models

GloVe



Word2Vec



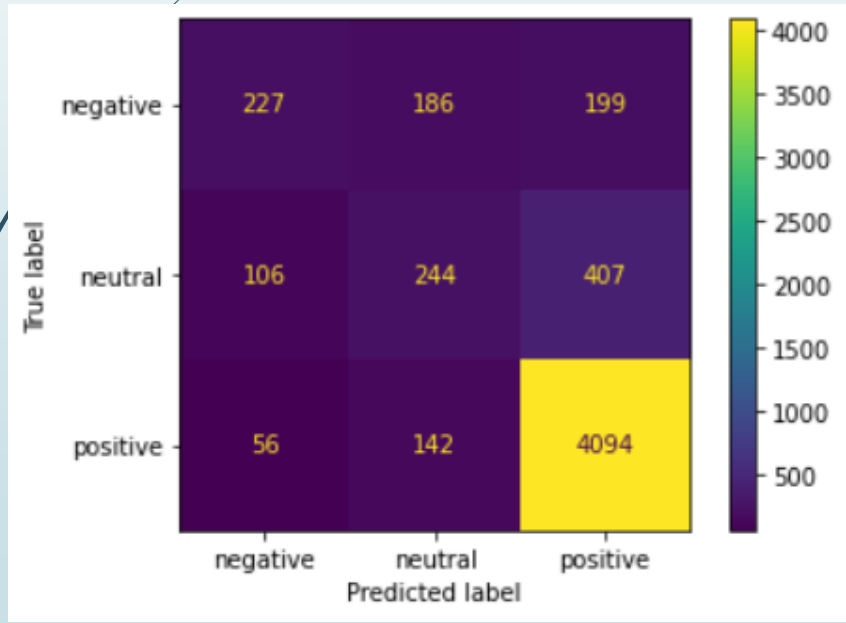
Chosen parameters' value

	LSTM1 (GloVe)	LSTM2 (Word2Vec)
Number of hidden layer	1	2
Dropout value	0.4	0.4
Learning rate	0.01	0.001
Number of epochs	5	13

Model comparison

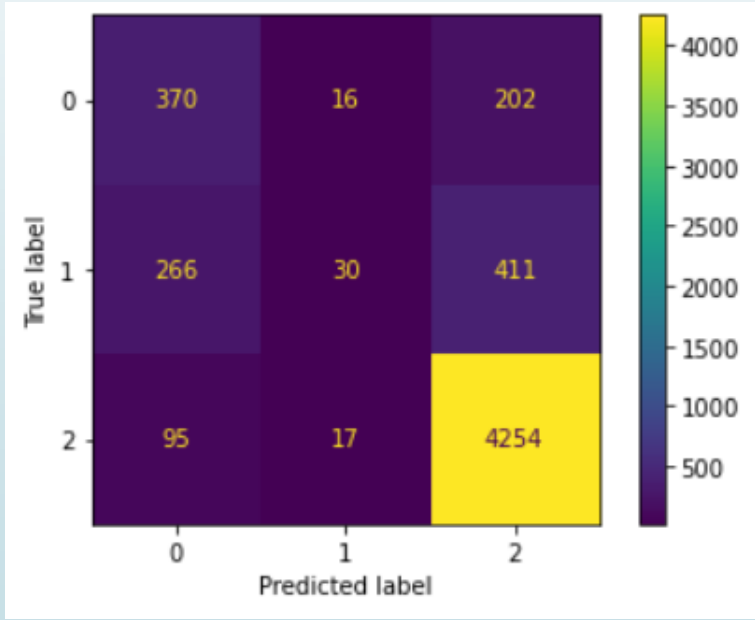
Benchmark model
Multinomial Naïve Bayes

Accuracy: 80,6%



Student's model
LSTM (composed from 2 submodels)

Accuracy: 82,2%



Conclusions

- Negative and neutral rating prediction difficulties

Further improvements:

- Further stop words removal
- Change of prediction target to negative and positive reviews

Thank you for your attention