Women's E-Commerce Clothing Reviews

Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky
Obor Data Analytics

nám. W. Churchilla 4
130 67 Praha 3
Česká republika

e-mail: mini02@vse.cz

4IZ172 – Text analytics I

Iana Minibaeva

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

Data Preprocessing

- I have chosen 3000 text reviews randomly as a sample to work with
- Dropping NA values
- Converting reviews to lowercase
- Removal of special characters and numbers
- Tokenization
- Removal of stopwords
- Lemmatization
- Merging the tokens back into sentences

Classification

Data Split: The sample is divided into subsets with 80% used for training and 20% for testing.

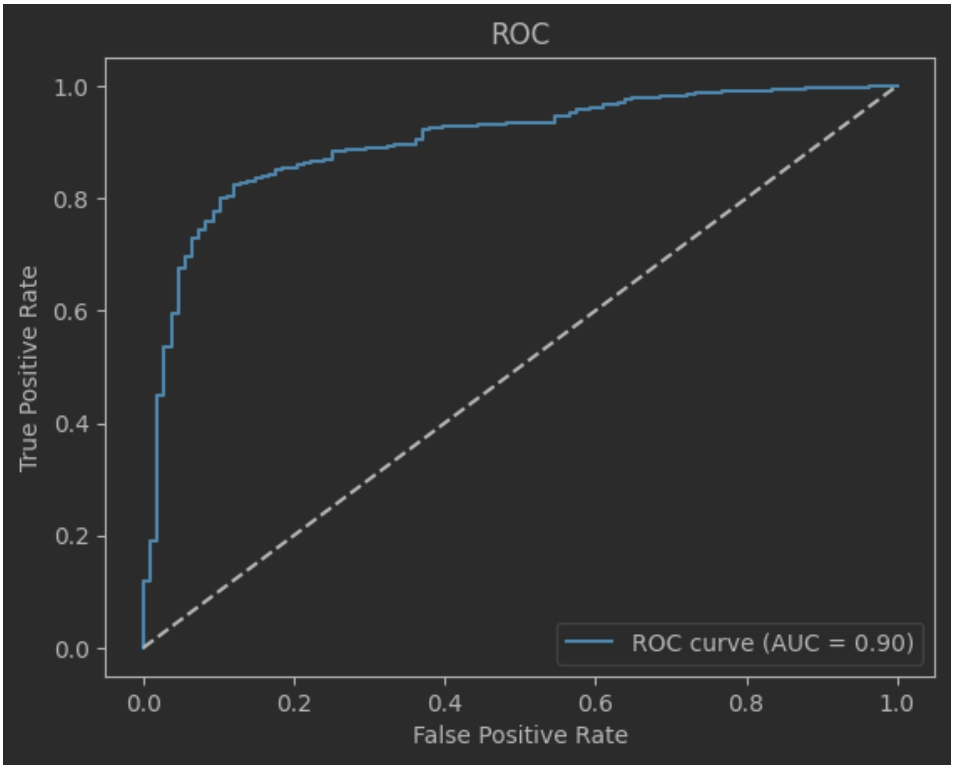
Classifier: Logistic Regression

The classifier is evaluated based on the following metrics:

- **Accuracy**: 0.8456
- **Precision**: 0.8468
- **Recall**: 0.9894
- **F1-score**: 0.9126
- **Confusion matrix**:

23	**85**
5	**470**

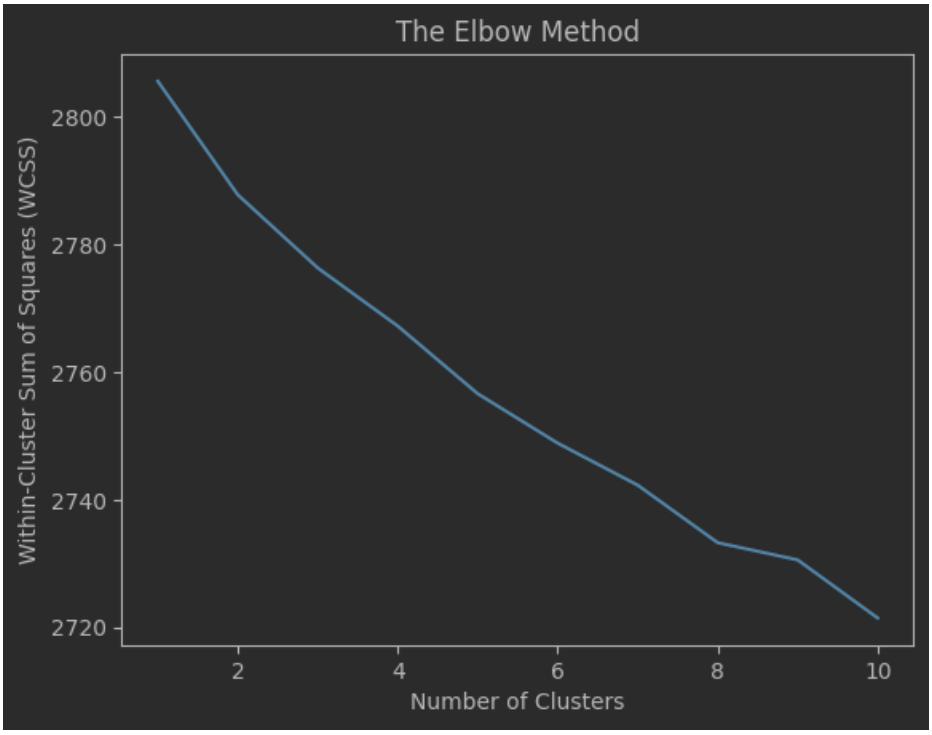
- **ROC curve**:



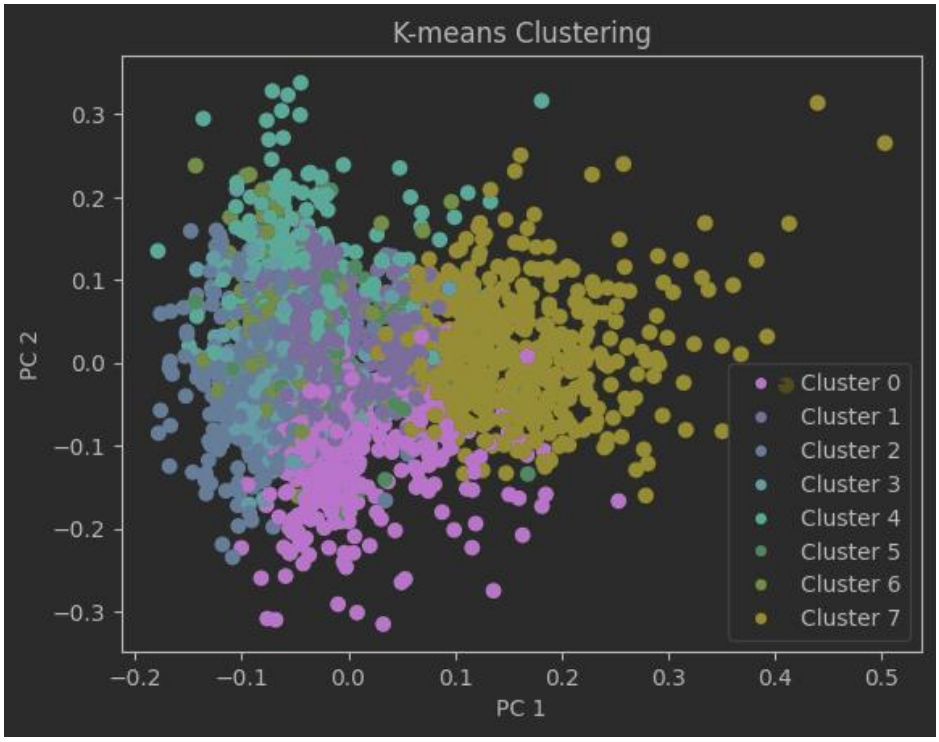
The AUC Score for the ROC curve is 0.9004.

Clustering

To figure out the number of clusters to use I performed the Elbow Method.

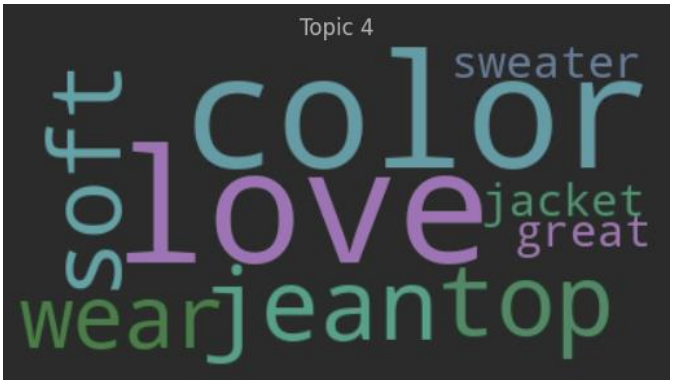


I have chosen 6 clusters for K-means clustering.

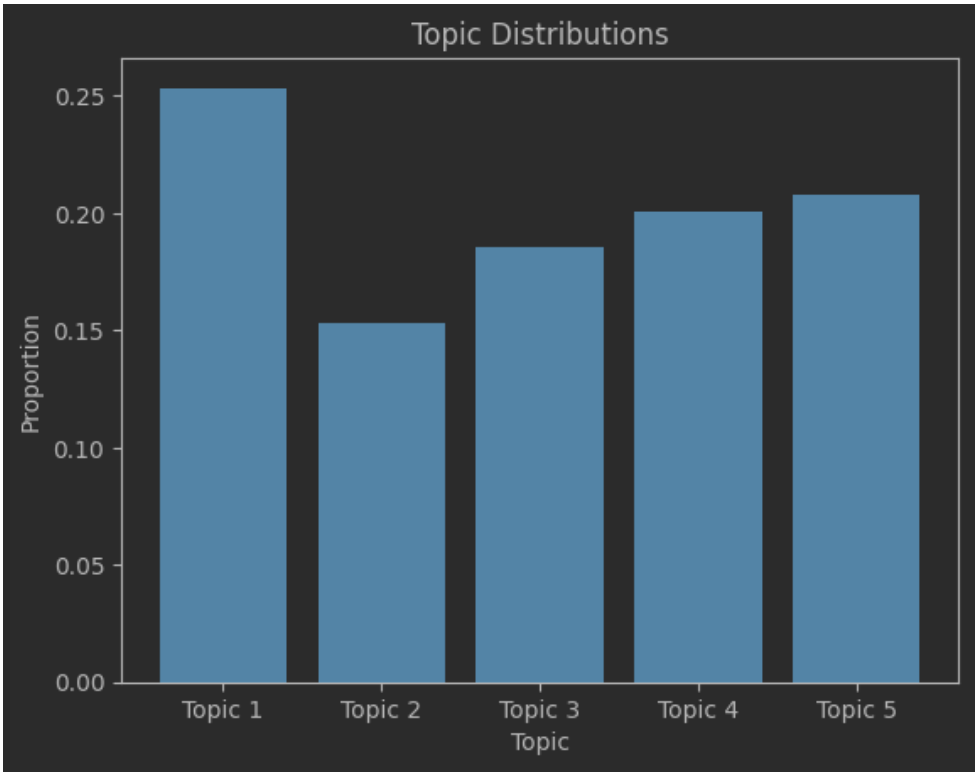


Topic modelling

I performed Topic Modeling using Latent Dirichlet Allocation (LDA) and visualized them with word clouds. Here are the results:




Topic 5
love
top
im
size
medium
look
fit
small
wear


Topic Distributions
Proportion
0.25
0.20
0.15
0.10
0.05
0.00
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic

Collocation Analysis

I transformed the reviews using TFIDF vectorizer to capture bigrams and calculated the average TF-IDF scores for each. Here are the top 15 bigrams:

	Bigram	TF-IDF Score
46881	true size	0.006034
14950	fit perfectly	0.004896
24726	look great	0.004425
25663	love top	0.003782
25427	love dress	0.003722
24754	look like	0.003575
36483	run large	0.003556
15056	fit true	0.003340
14866	fit great	0.003157
15077	fit well	0.003073
48189	usually wear	0.003063
20226	im lb	0.002945
39549	size small	0.002835
39355	size fit	0.002657
50291	well made	0.002643