NLP

SENTIMENT ANALYSIS & TOPIC MODELLING

Kwan Pei Ni

TABLE OF CONTENTS

Amazon

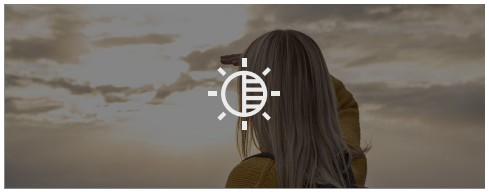
The North Face

American outdoor recreation products company was founded by the climber and his wife in 1966 and based in San Francisco. The North Face produces outdoor clothing, footwear, and related equipment.


OBJECTIVES
&
APPROACH

OBJECTIVES

- To seek higher peaks and push limits
- To create a stronger connection with consumers



Physical Stores + Online Stores = Breakthrough

Thousands of physical stores worldwide, The North Face **caters for the customers' needs** based on feedback obtained from sales assistants

Reviews from the online buyers are one of the **precious possessions** that The North Face values the most in order **to fulfill their demands thus improving the business**

A humble shop not only grown to become one of the most reputable manufacturers of adventure gear, but has also managed to break into the fashion markets

THE APPROACH

Sentiment Analysis (Supervised Learning)

Topic Modelling (Unsupervised Learning)

VADER
(Valence Aware Dictionary and sEntiment Reasoner)

ROBERTA
(Robustly Optimized BERT Approach)

LSA & LDA
(Latent Semantic Analysis) & (Latent Dirichlet Allocation)

BERTOPIC
(Bidirectional Encoder Representations from Transformers)


WEB
SCRAPING
(DATASET)

DATA PREPROCESSING & EXPLORATION

(I) There are 7 categories in Men's sections:
- Jackets & Vests
- Sweatshirts & Fleece
- Tops
- Bottoms
- Footwear
- Accessories

(II) Data Understanding:
- Fill in missing values
- Drop duplicated data
- Change to datatime
- Remove repeated words

Link:
https://www.amazon.com/stores/page/4CB8EA89-2A36-48BC-A2C2-E23C816F9339?ingress=0&visitId=28e24b97-79ad-4fc4-995d-93bc59e60e8d

```
1 df.head(10)
```

	name	price	reviewers	dates	ratings	titles	contents
0	THE NORTH FACE Men's Venture 2 Waterproof Hood... 109.85 – \n	109.95	Abraham	2023-06-04	4.6	Worth every penny	This is a must have if you are in a rainy area...
1	THE NORTH FACE Men's Venture 2 Waterproof Hood... 109.85 – \n	109.95	The Review	2023-04-23	4.6	comfortable and completely water proof	very wet rainy season this year, so had ample ...
2	THE NORTH FACE Men's Venture 2 Waterproof Hood... 109.85 – \n	109.95	Abraham	2023-06-04	4.6	Muy útil	Cómoda , resiste el frio y el agua
3	THE NORTH FACE Men's Venture 2 Waterproof Hood... 109.85 – \n	109.95	Amazon Customer	2023-05-17	4.6	Very Nice	Fits well and great materials and construction
4	THE NORTH FACE Men's Venture 2 Waterproof Hood... 109.85 – \n	109.95	Nelson Junior Jara Montiel	2023-05-28	4.6	Different fit, cheaper design	I am replacing the original venture jacket. Th...
5	THE NORTH FACE Men's Venture 2 Waterproof Hood... 109.85 – \n	109.95	JRandyMyrick	2023-05-01	4.6	Garment	Color inside. Pockets
6	THE NORTH FACE Men's Venture 2 Waterproof Hood... 109.85 – \n	109.95	The Review	2023-04-23	4.6	Keeps me dry.	The jacket looks to be well made and did a gre...
7	THE NORTH FACE Men's Venture 2 Waterproof Hood... 109.85 – \n	109.95	harv	2023-06-03	4.6	Outstanding Quality	Exactly what I ordered! The product is outstan...
8	THE NORTH FACE Men's Venture 2 Waterproof Hood... 109.85 – \n	109.95	Sam	2023-02-27	4.6	as expected	very lightweight weatherproof shell. the mater...
9	THE NORTH FACE Men's Venture 2 Waterproof Hood... 109.85 – \n	109.95	Amy	2023-03-18	4.6	It's just a basic windbreaker	You're definitly only paying for the name. The...

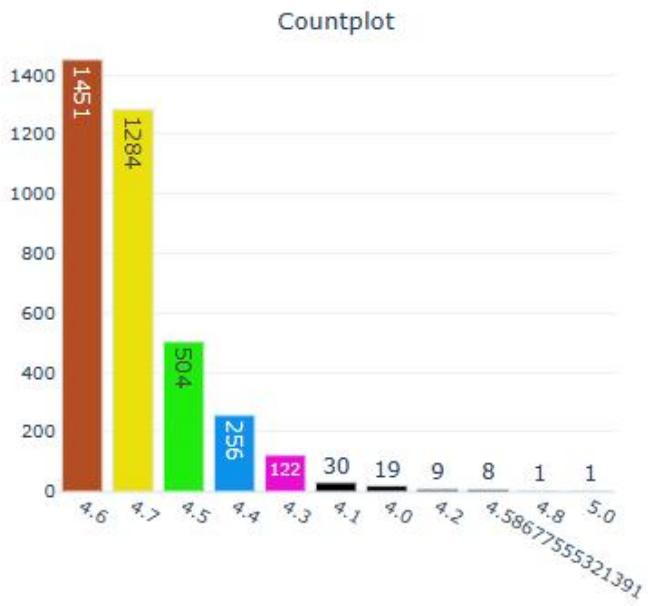

EDA
ANALYSIS

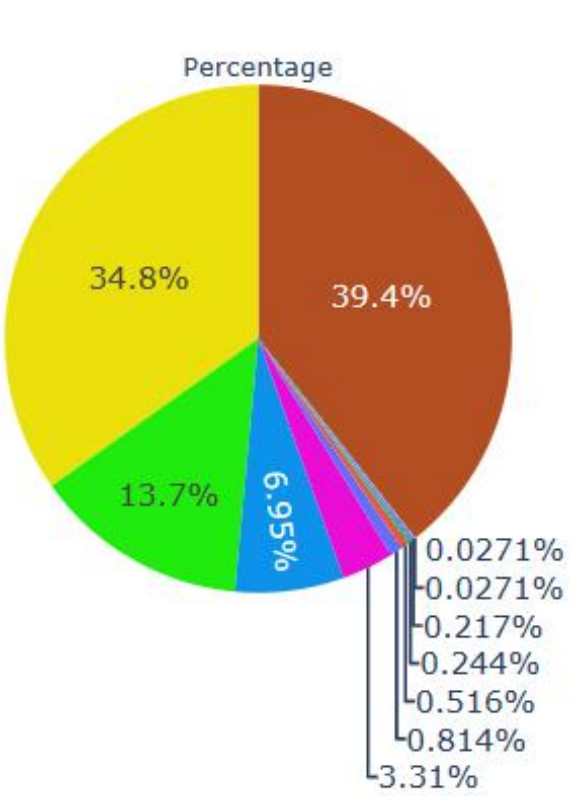
EDA


```
1 print(df['ratings'].value_counts())
```

```
4.600000    1451
4.700000    1284
4.500000     504
4.400000     256
4.300000     122
4.100000      30
4.000000      19
4.200000       9
4.586776       8
4.800000       1
5.000000       1
Name: ratings, dtype: int64
```

EDA





Summary:
- Ratings of 4.6 has the highest countplot and percentage of 1451 and 39.4% respectively


NLP
Natural Language Processing

To analyze sentiment in textual data

VADER

(Valence Aware Dictionary and sEntiment Reasoner)

VADER

Bag of Words Approach

(I) By using NLTK's **Sentiment Intensity Analyzer** to get the neg/neu/pos scores of the text.

- This uses a "bag of words" approach:
1) Stop words are removed
2) Each word is scored and combined to a total score

(II) The compound score typically ranges from -1 to 1, where:
- A compound score of 1 indicates highly positive sentiment.
- A compound score of 0 indicates neutral sentiment.
- A compound score of -1 indicates highly negative sentiment.

```
1 sia.polarity_scores('My son loves it and wears it to scout camp')
```

```
{'neg': 0.0, 'neu': 0.709, 'pos': 0.291, 'compound': 0.5719}
```

Observation 1: The result shows that the sentence does not have any negative information (neg=0). It has higher neutral tones compared to positive (neu=0.709 and pos=0.291). However, the overall sentiment is positive because compound > 0.05

```
1 sia.polarity_scores('I really like it, great quality except it was too small. I had to return it.')
```

```
{'neg': 0.0, 'neu': 0.626, 'pos': 0.374, 'compound': 0.7996}
```

Observation 2: From here, the compound jumped to 0.7996, which makes the sentence more positive then the one from Observation 1

```
1 sia.polarity_scores('Rubbery bill! Low profile cap! Not a hunter cap!')
```

```
{'neg': 0.298, 'neu': 0.702, 'pos': 0.0, 'compound': -0.4545}
```

Observation 3: The above sentence does not have any positive information (pos=0). It has some neutral and negative tones (neu=0.702 and neg=0.298). It has an extreme negative sentiment due to the compound score which is close to -1.

```
1 sia.polarity_scores('What a HORRIBLE hat. Fits horrible! Try it on and you'll be returning it like me')
```

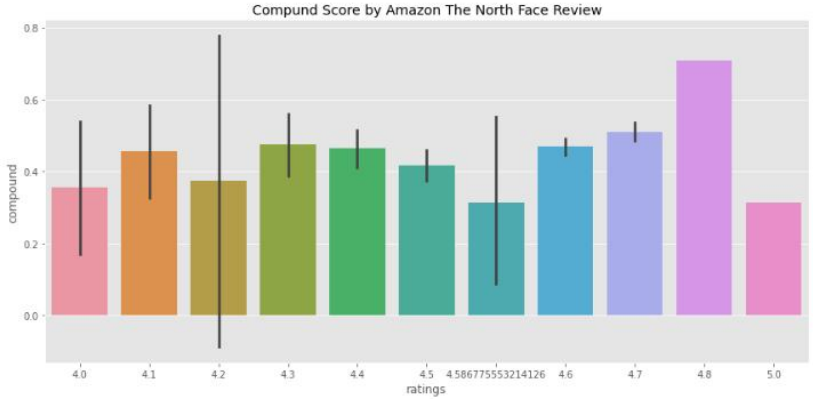
```
{'neg': 0.356, 'neu': 0.533, 'pos': 0.111, 'compound': -0.7597}
```

```
1 sia.polarity_scores('If you have small head it will fit perfectly 👍')
```

```
{'neg': 0.0, 'neu': 0.511, 'pos': 0.489, 'compound': 0.7717}
```

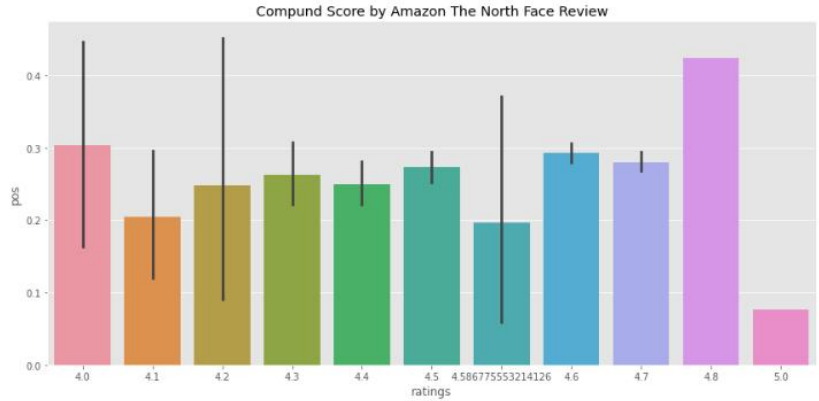
VADER

```
plt.figure(figsize=(12, 6))
ax = sns.barplot(data=vaders_new, x='ratings', y='compound')
ax.set_title('Compund Score by Amazon The North Face Review')
plt.tight_layout()
plt.show()
```



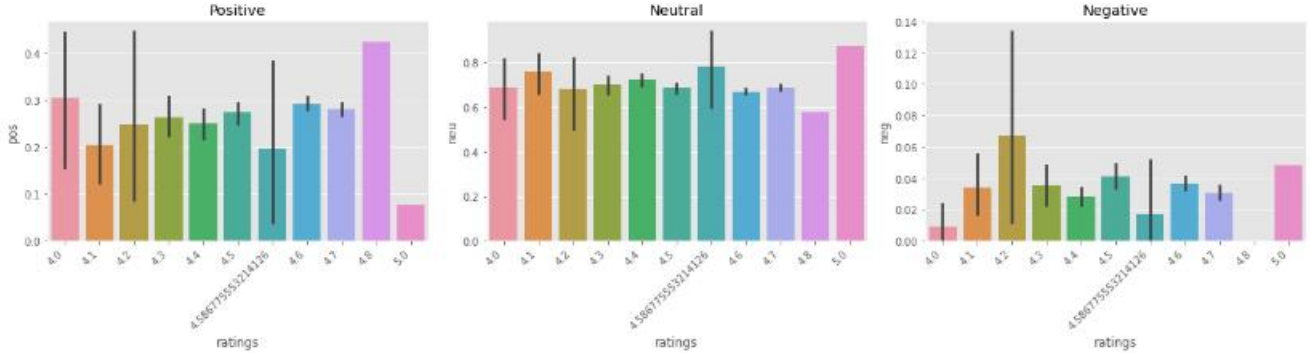
Plot VADER Results

```
plt.figure(figsize=(12, 6))
ax = sns.barplot(data=vaders_new, x='ratings', y='pos')
ax.set_title('Compund Score by Amazon The North Face Review')
plt.tight_layout()
plt.show()
```



VADER

```
fig, axs = plt.subplots(1, 3, figsize=(18, 5))
sns.barplot(data=vaders_new, x='ratings', y='pos', ax=axs[0])
sns.barplot(data=vaders_new, x='ratings', y='neu', ax=axs[1])
sns.barplot(data=vaders_new, x='ratings', y='neg', ax=axs[2])
axs[0].set_title('Positive')
axs[1].set_title('Neutral')
axs[2].set_title('Negative')
axs[0].set_xticklabels(axs[0].get_xticklabels(), rotation=45, ha='right')
axs[1].set_xticklabels(axs[1].get_xticklabels(), rotation=45, ha='right')
axs[2].set_xticklabels(axs[2].get_xticklabels(), rotation=45, ha='right')
plt.tight_layout()
plt.show()
```



I) From the above graph, it is not directly proportional whereby the higher the ratings, the more positive the review are.

II) Upon checking the dataset it is observed that for some of the reviews, it has a **high rating but with a negative comment.**

III) Hence, the ratings is **not directly proportional** to the contents.

VADER

To generate the performance (precision, recall and f1 score)

Accuracy: 0.3603799185888738

	precision	recall	f1-score	support
negative	0.01	0.09	0.02	58
neutral	0.62	0.14	0.23	2341
positive	0.36	0.78	0.49	1286
accuracy			0.36	3685
macro avg	0.33	0.33	0.25	3685
weighted avg	0.52	0.36	0.32	3685

I) The **accuracy is low** which is **0.36**, meaning the polarity is not accurate as some of the bad reviews have high ratings.

II) This might due to the reason that the reviewers want their comments to be at the top of the reviews, so that it is more noticeable.

III) In order to remove the "faulty" data, reviews with high ratings but low compounds score are required to be filtered out.

VADER

Data with high ratings but low compound score:

	index	neg	neu	pos	compound	sentiment	name		price	reviewers	dates	ratings	titles	contents
2	2	0.000	1.000	0.000	0.0000	neu	THE NORTH FACE Men's Venture 2 Waterproof Hood...	109.85 – \n	109.95	Abraham	2023-06-04	4.6	Muy útil	Cómoda , resiste el frio y el agua
5	5	0.000	1.000	0.000	0.0000	neu	THE NORTH FACE Men's Venture 2 Waterproof Hood...	109.85 – \n	109.95	JRandyMyrick	2023-05-01	4.6	Garment	Color inside. Pockets
9	9	0.111	0.785	0.104	0.1882	neu	THE NORTH FACE Men's Venture 2 Waterproof Hood...	109.85 – \n	109.95	Amy	2023-03-18	4.6	It's just a basic windbreaker	You're definitly only paying for the name. The...
14	15	0.000	1.000	0.000	0.0000	neu	THE NORTH FACE Men's Venture 2 Waterproof Hood...	109.85 – \n	109.95	Amazon Customer	2023-05-17	4.6	Garment	Color inside. Pockets
18	19	0.111	0.785	0.104	0.1882	neu	THE NORTH FACE Men's Venture 2 Waterproof Hood...	109.85 – \n	109.95	harv	2023-06-03	4.6	It's just a basic windbreaker	You're definitly only paying for the name. The...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3619	3773	0.356	0.533	0.111	-0.7597	neu	THE NORTH FACE Mudder Trucker Mens Cap		$31.95	Wade K.	2023-04-19	4.6	What a HORRIBLE hat. Fits horrible! Try it on ...	What a HORRIBLE hat. Fits horrible! Try it on ...

```
1 total_rows = 3685
2 faulty_rows = 700
3
4 proportion_faulty = (faulty_rows / total_rows) * 100
5 proportion_faulty
```

```
18.99592944369064
```

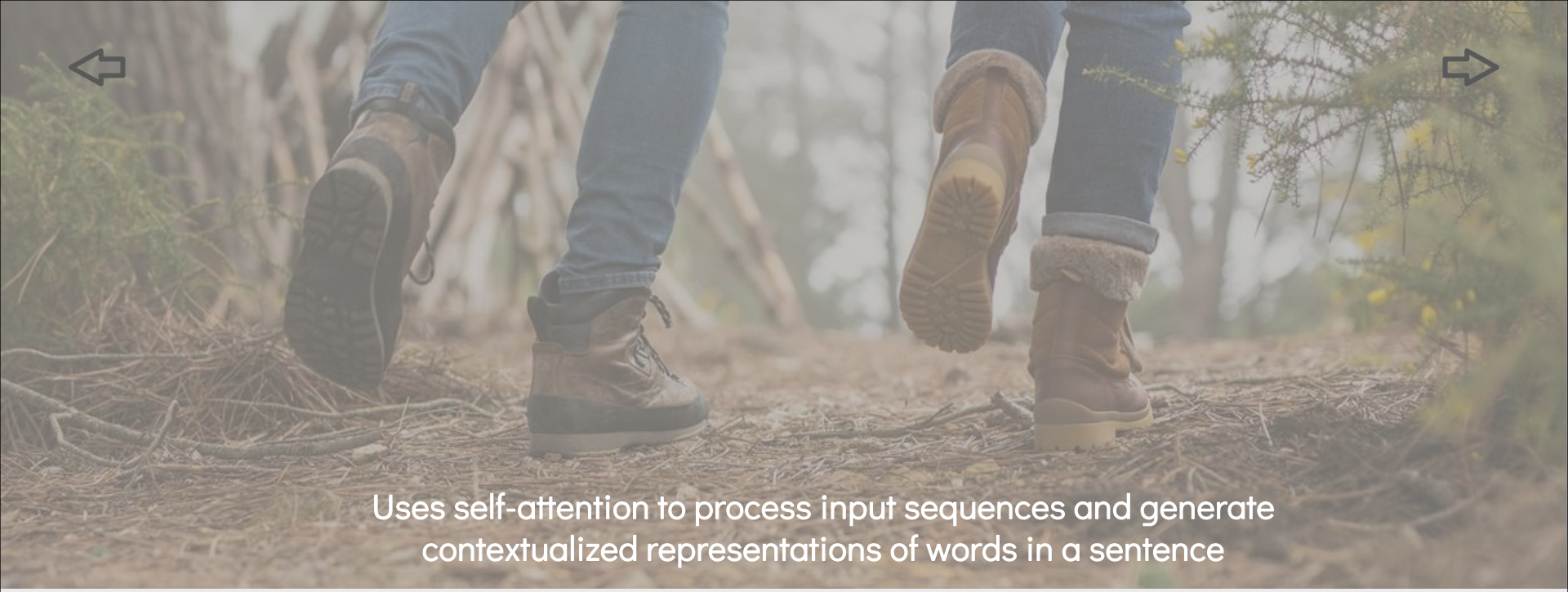
I) It is about 19% of faulty data

VADER

Accuracy: 0.37152428810720267

	precision	recall	f1-score	support
negative	0.04	0.09	0.06	58
neutral	0.89	0.07	0.12	1950
positive	0.36	1.00	0.53	977
accuracy			0.37	2985
macro avg	0.43	0.38	0.24	2985
weighted avg	0.70	0.37	0.25	2985

I) After removing the 19% of faulty data, the accuracy has a slight increase to 0.37



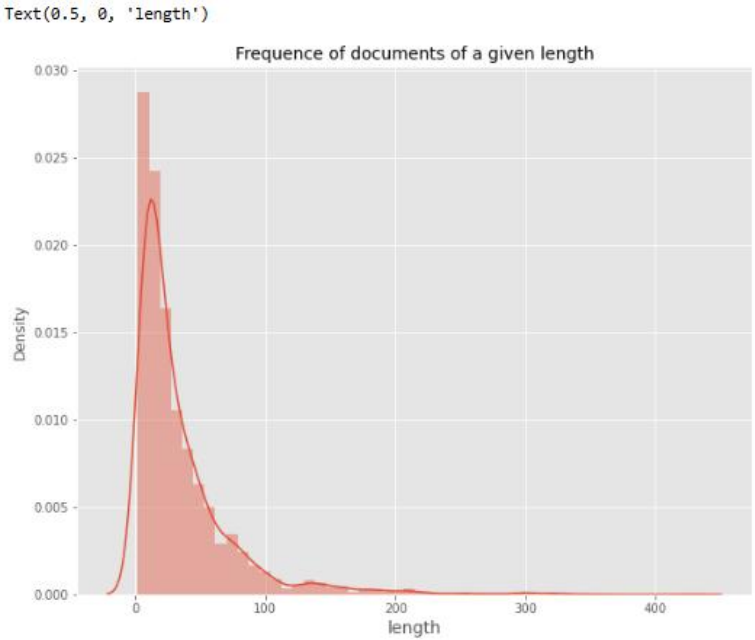
Uses self-attention to process input sequences and generate contextualized representations of words in a sentence

ROBERTA

(Robustly Optimized BERT Approach)

RoBERTa

(I) Use a model trained of a large corpus of data
(II)Transfomer model accounts for the words but also the context related to other words



(I) The histogram depicts the text length distribution

(II) Length is the combination of titles and contents

(III) New data of titles_contents is then stored in a new excel file, named ‘prep_excel_file_final.xlsx

Vader & RoBERTa

Model Training & Fine Tuning

index	vader_neg	vader_neu	vader_pos	vader_compound	roberta_neg	roberta_neu	roberta_pos	name	price	reviewers	dates	ratings	titles	contents	converted_ratings	titles_contents	length
0	0.029	0.737	0.234	0.8473	0.002333	0.034382	0.963286	THE NORTH FACE Men's Venture 2 Waterproof Hood...	109.85 – \n 109.95	Abraham	2023-06-04	4.6	Worth every penny	This is a must have if you are in a rainy area...	1	Worth every penny. This is a must have if you ...	48
1	0.051	0.677	0.272	0.7810	0.002992	0.048708	0.948300	THE NORTH FACE Men's Venture 2 Waterproof Hood...	109.85 – \n 109.95	The Review	2023-04-23	4.6	comfortable and completely water proof	very wet rainy season this year, so had ample ...	1	comfortable and completely water proof. very w...	30
2	0.000	1.000	0.000	0.0000	0.097650	0.820784	0.081566	THE NORTH FACE Men's Venture 2 Waterproof Hood...	109.85 – \n 109.95	Abraham	2023-06-04	4.6	Muy útil	Cómoda , resiste el frío y el agua	1	Muy útil. Cómoda , resiste el frío y el agua	10
3	0.000	0.446	0.554	0.7351	0.003950	0.071658	0.924392	THE NORTH FACE Men's Venture 2 Waterproof Hood...	109.85 – \n 109.95	Amazon Customer	2023-05-17	4.6	Very Nice	Fits well and great materials and construction	1	Very Nice. Fits well and great materials and c...	9
4	0.060	0.843	0.096	0.3716	0.564688	0.366071	0.069241	THE NORTH FACE Men's Venture 2 Waterproof Hood...	109.85 – \n 109.95	Nelson Junior Jara Montiel	2023-05-28	4.6	Different fit, cheaper design	I am replacing the original venture jacket. Th...	1	Different fit, cheaper design. I am replacing ...	155
5	0.000	1.000	0.000	0.0000	0.224240	0.711936	0.063824	THE NORTH FACE Men's Venture 2 Waterproof Hood...	109.85 – \n 109.95	JRandyMyrick	2023-05-01	4.6	Garment	Color inside. Pockets	1	Garment. Color inside. Pockets	4

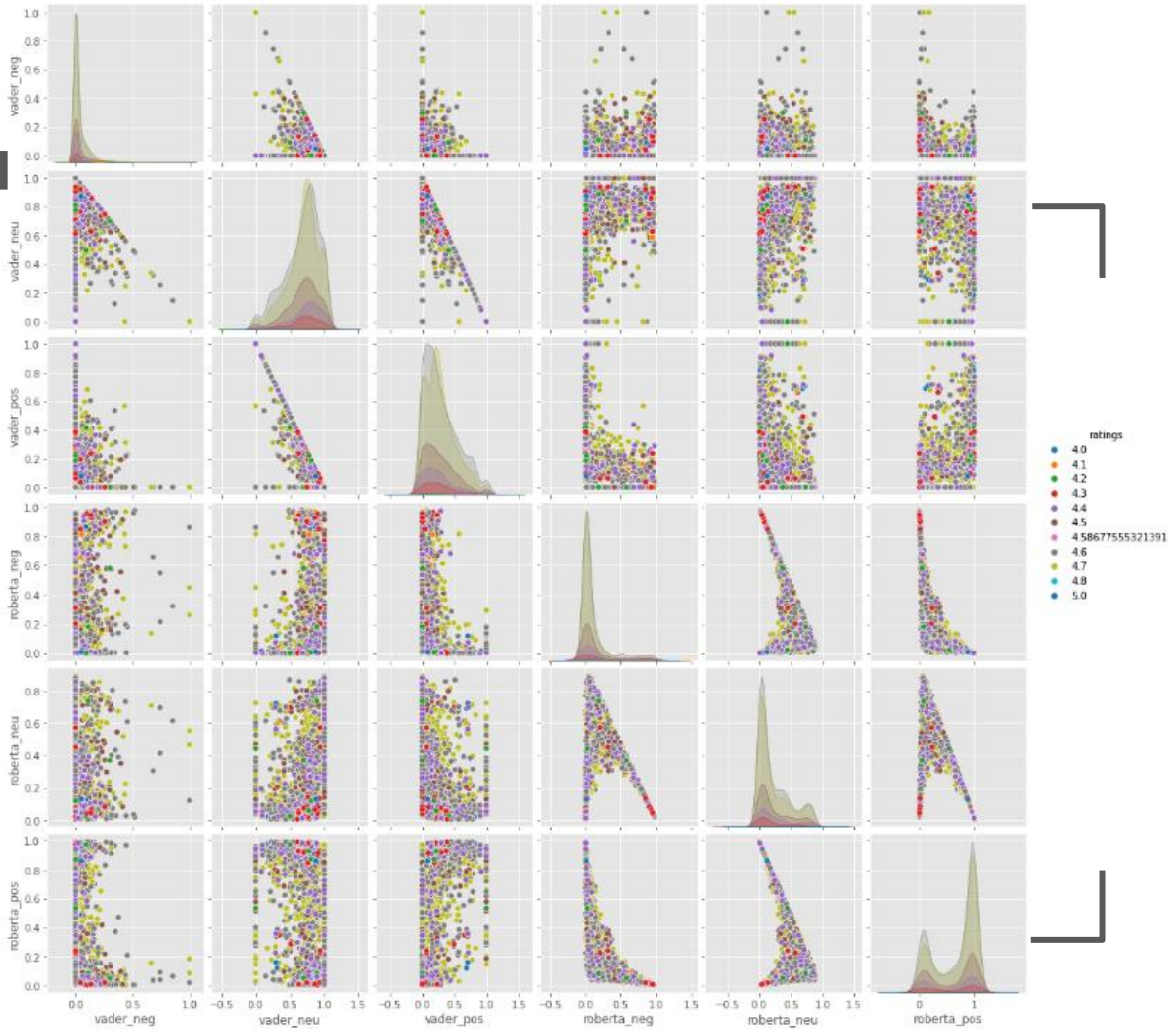
Vader & RoBERTa

Combine & Compare

(I) Blue represents the ratings of 4.8 and 5 while yellowish-green and green represents the ratings of 4.6 and 4.7

(II) Positive review for RoBERTa is more on the right compared to VADER

(III) Vader shows that it is less confident in prediction compared to RoBERTa



LSA & LDA

(Latent Semantic Analysis)

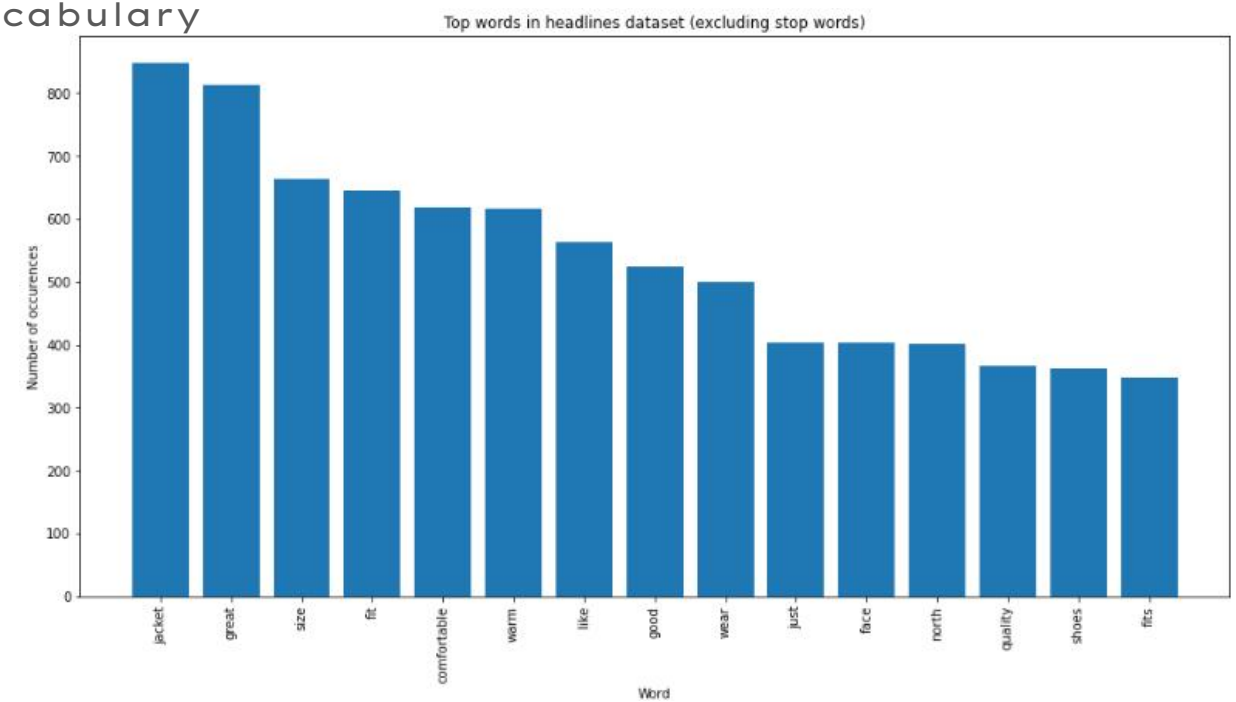
LSA captures semantic relationships and similarities between terms and documents

(Latent Dirichlet Allocation)

LDA model the underlying topic structure within the documents

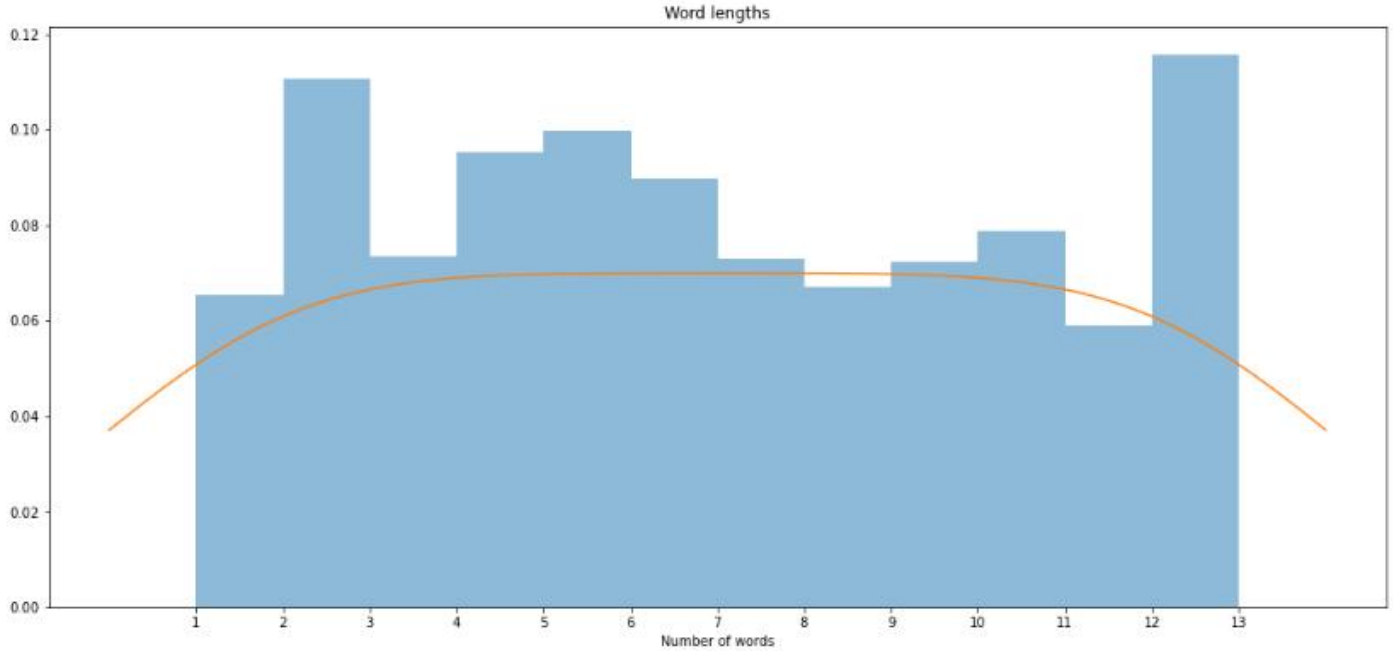
LSA & LDA

EDA - To develop top words to get a glimpse into the core vocabulary



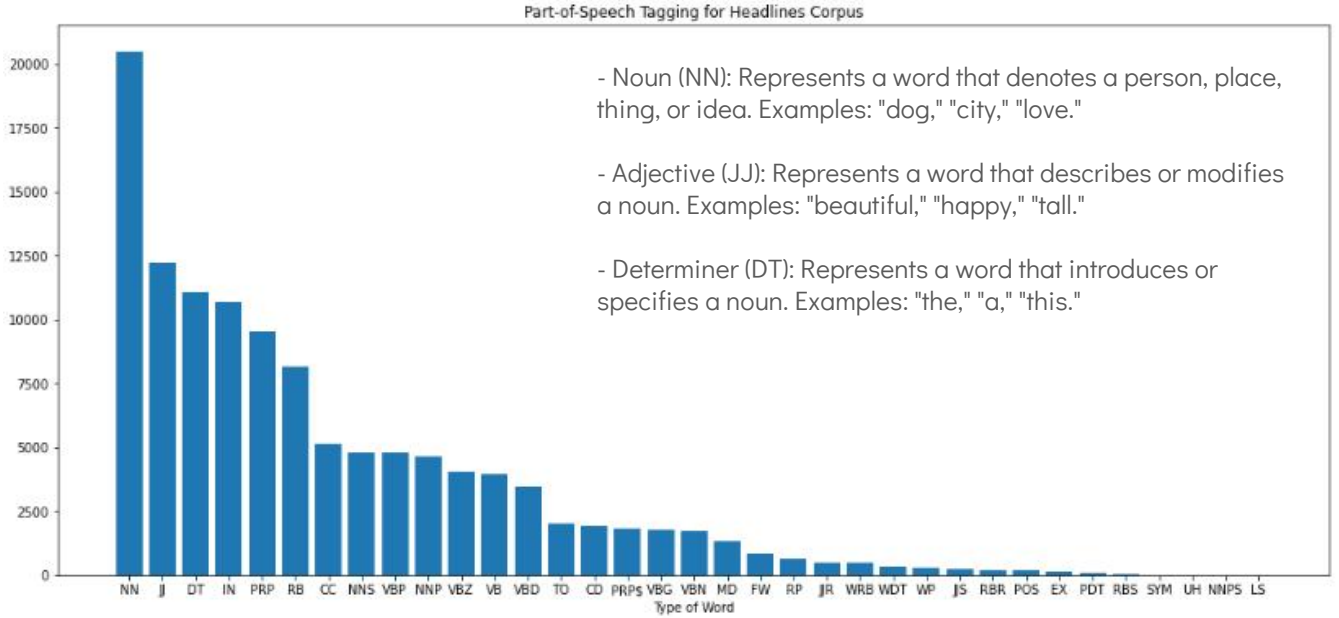
A histogram of headline word lengths is then generated and use part-of-speech tagging to understand the types of words used across the corpus.

LSA & LDA



A combined histogram is generated and KDE plot to provide insights into the distribution of word counts. The histogram represents the observed frequencies or densities of word counts, while the KDE estimates the underlying continuous distribution.

LSA & LDA



- Noun (NN): Represents a word that denotes a person, place, thing, or idea. Examples: "dog," "city," "love."

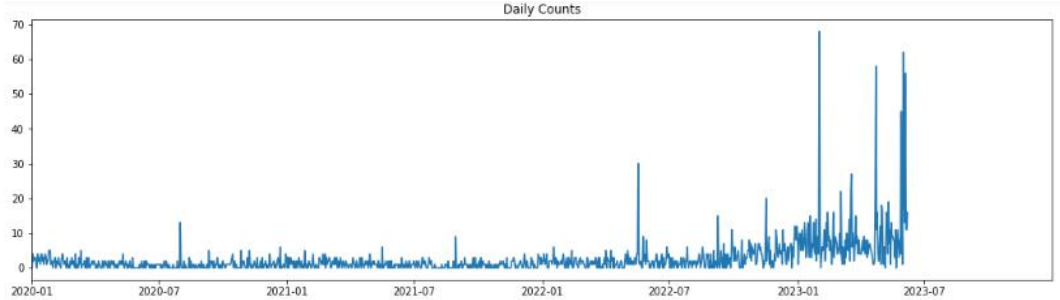
- Adjective (JJ): Represents a word that describes or modifies a noun. Examples: "beautiful," "happy," "tall."

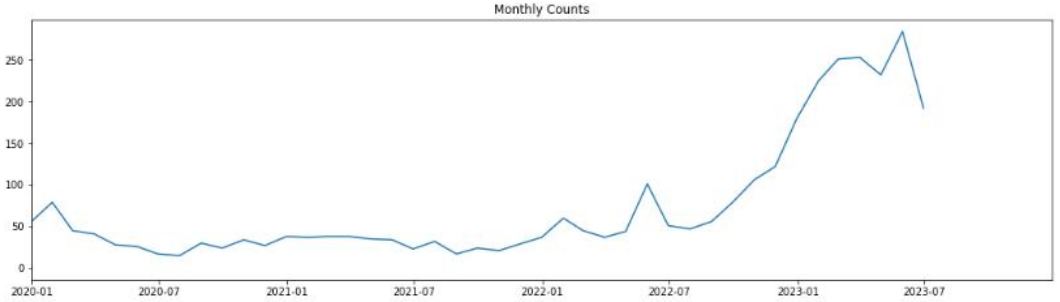
- Determiner (DT): Represents a word that introduces or specifies a noun. Examples: "the," "a," "this."

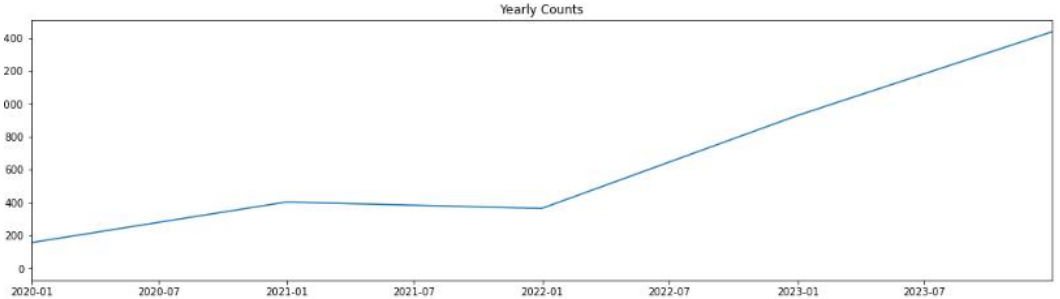
A bar plot is generated where each bar represents the frequency or count of different part-of-speech tags in a corpus of headlines. The x-axis shows the types of words (part-of-speech tags), and the height of each bar represents the count of that particular tag. This plot provides a visual representation of the distribution of part-of-speech tags in the headlines corpus.

LSA & LDA

Above graphs produce three line plots representing the counts of a variable over different time intervals: daily, monthly, and yearly. It provides a visual representation of how the counts vary over time at different levels of granularity.







LSA & LDA

Clustering algorithm was applied to the corpus in order to study the topic focus, as well as how it has evolved through time.

Preprocessing

```
small_count_vectorizer = CountVectorizer(stop_words='english', max_features=40000)
small_text_sample = reindexed_df.sample(n=1000, random_state=0).values

print('Before vectorization: {}'.format(small_text_sample[123]))

small_document_term_matrix = small_count_vectorizer.fit_transform(small_text_sample)

print('After vectorization: \n{}'.format(small_document_term_matrix[123]))
```

```
Before vectorization: Love this hat. Great quality.
After vectorization:
  (0, 1324)     1
  (0, 1386)     1
  (0, 2331)     1
  (0, 1751)     1
```

To convert each string to a numerical vector using CountVectorizer.

LSA vs LDA

Top N words are extracted for each topic generated by LSA and then prints the topics along with their associated top words. This is to understand the main themes or concepts represented by each topic in the LSA analysis.

```
top_n_words_lsa = get_top_n_words(10, lsa_keys, small_document_term_matrix, small_count_vectorizer)

for i in range(len(top_n_words_lsa)):
    print("Topic {}: ".format(i+1), top_n_words_lsa[i])

```

```
Topic 1:  jacket great comfortable warm fit size good like wear just
Topic 2:  shoes shoe pair comfortable purchased cold size weather wore durable
Topic 3:  hat north face head great like quality super die high
Topic 4:  la son el que es en muy si para il
Topic 5:  size small grey color tnf hat large dark black fits
Topic 6:  logo tnf did like choice beanie liked just don lining
Topic 7:  venture original rain release hood jackets spring charging resolve adjustable
Topic 8:  north face shoe bought jacket mountain hardwear wear ultra weather
```

LSA

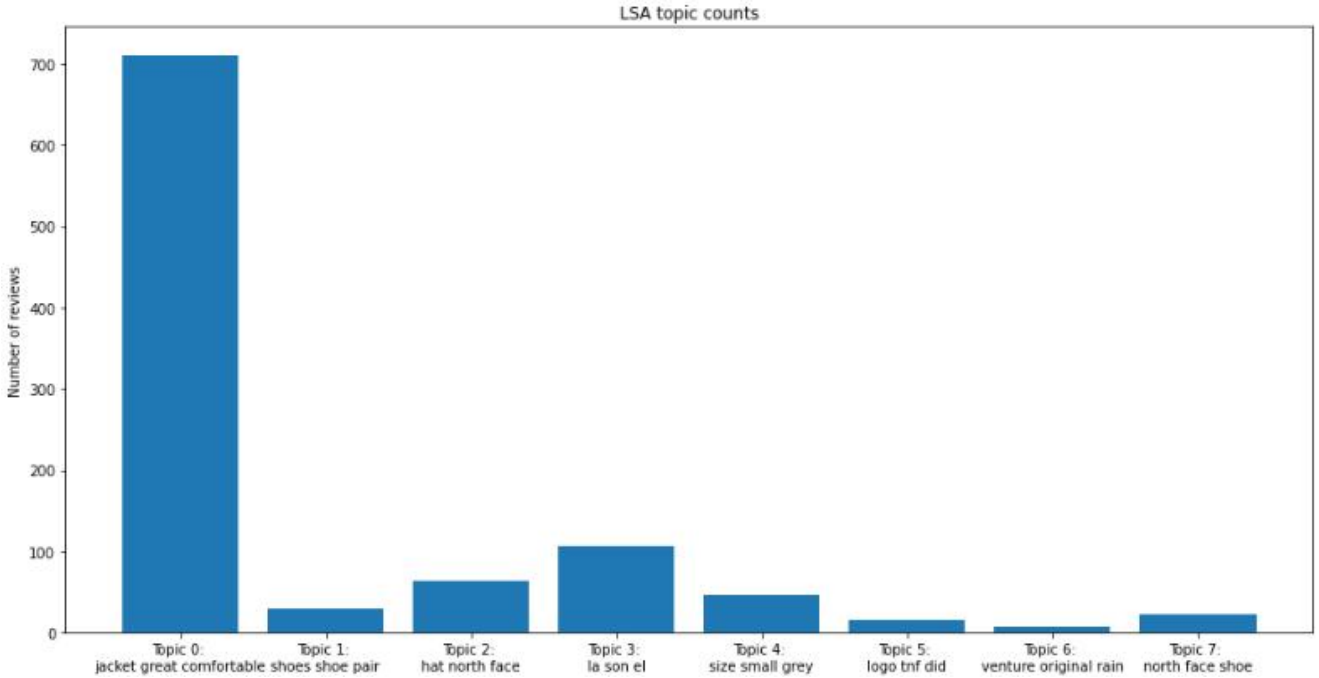
```
top_n_words_lda = get_top_n_words(10, lda_keys, small_document_term_matrix, small_count_vectorizer)

for i in range(len(top_n_words_lda)):
    print("Topic {}: ".format(i+1), top_n_words_lda[i])
```

```
Topic 1:  size fit warm comfortable like wear great jacket small good
Topic 2:  la il non di logo molto si sono prodotto tessuto
Topic 3:  jacket great fit warm hat comfortable good like north face
Topic 4:  comfortable jacket love great feet face north wear warm loves
Topic 5:  la el es en que muy son para se calidad
Topic 6:  die sie und jacke sich der ich hält ein das
Topic 7:  expected face north exactly hardwear mountain cómodas est bought le
Topic 8:  super shirt caldo ich comfortable soft ma il 30 stretto
```

LDA

LSA



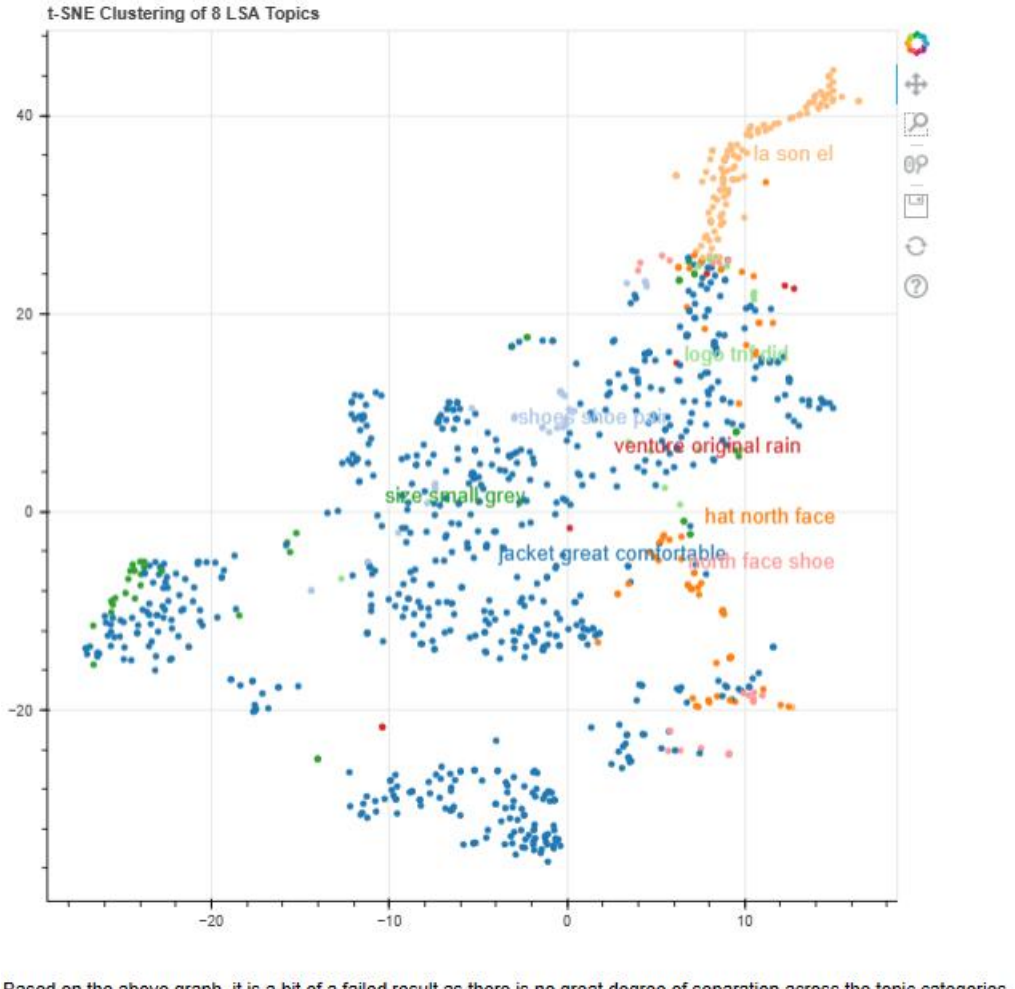
A bar chart showing the topic counts for the LSA topics is generated, with each bar labeled by the topic index and its corresponding top 3 words. The height of each bar represents the number of headlines associated with that topic. Topic 0 has the highest numbers for LSA topic counts which corresponds to the scatter plot

LSA

(I) A scatter plot is generated using t-SNE clustering for the LSA topics. Each data point represents a headline, and its position in the plot is determined by the t-SNE vectors. The color of each data point corresponds to its associated LSA topic. Additionally, labels are added to the plot, displaying the top 3 words for each LSA topic at the mean coordinates of the topic in the t-SNE space.

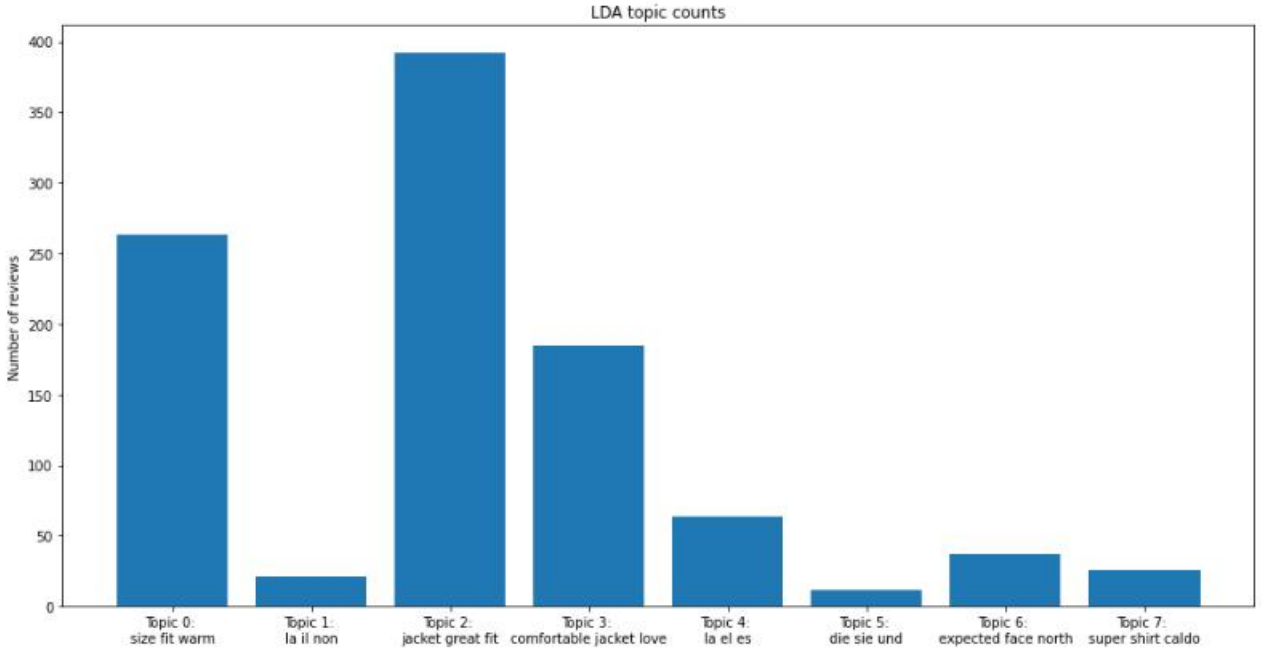
(II) From here, we could observe that the data for 'jacket great comfortable' has a **spread out data points thus do not exhibit a clear pattern or trend**.

(III) Since the data points are scattered randomly without any discernible pattern, it **suggests that there is no significant relationship between the variables**.



Based on the above graph, it is a bit of a failed result as there is no great degree of separation across the topic categories.

LDA



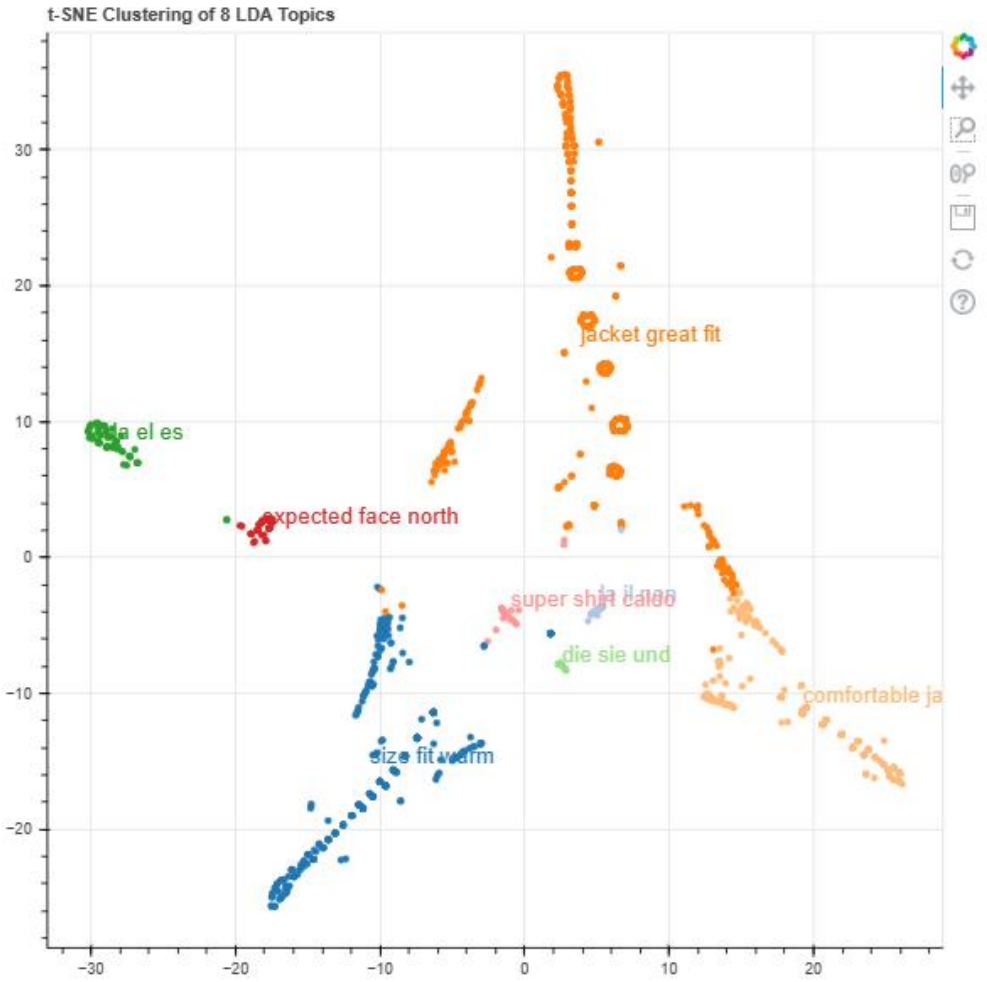
Topic 2, 0 and 3 are ranked first, second and third in terms of the number of reviews as per LDA topic counts respectively, this results corresponds to the scatter plot obtained, which depicts that distinct linear line pattern of datapoints are observed.

LDA

(I) **Focused or clustered** data points are observed, it exhibits a **distinct** pattern or concentration while indicating a **stronger correlation or relationship between the more structured variables.**

(II) It forms a **tight** cluster or follows a **distinct trend**. The values of one variable provide useful information about the values of the other variable, the clustered data points forms a **linear line** pattern.

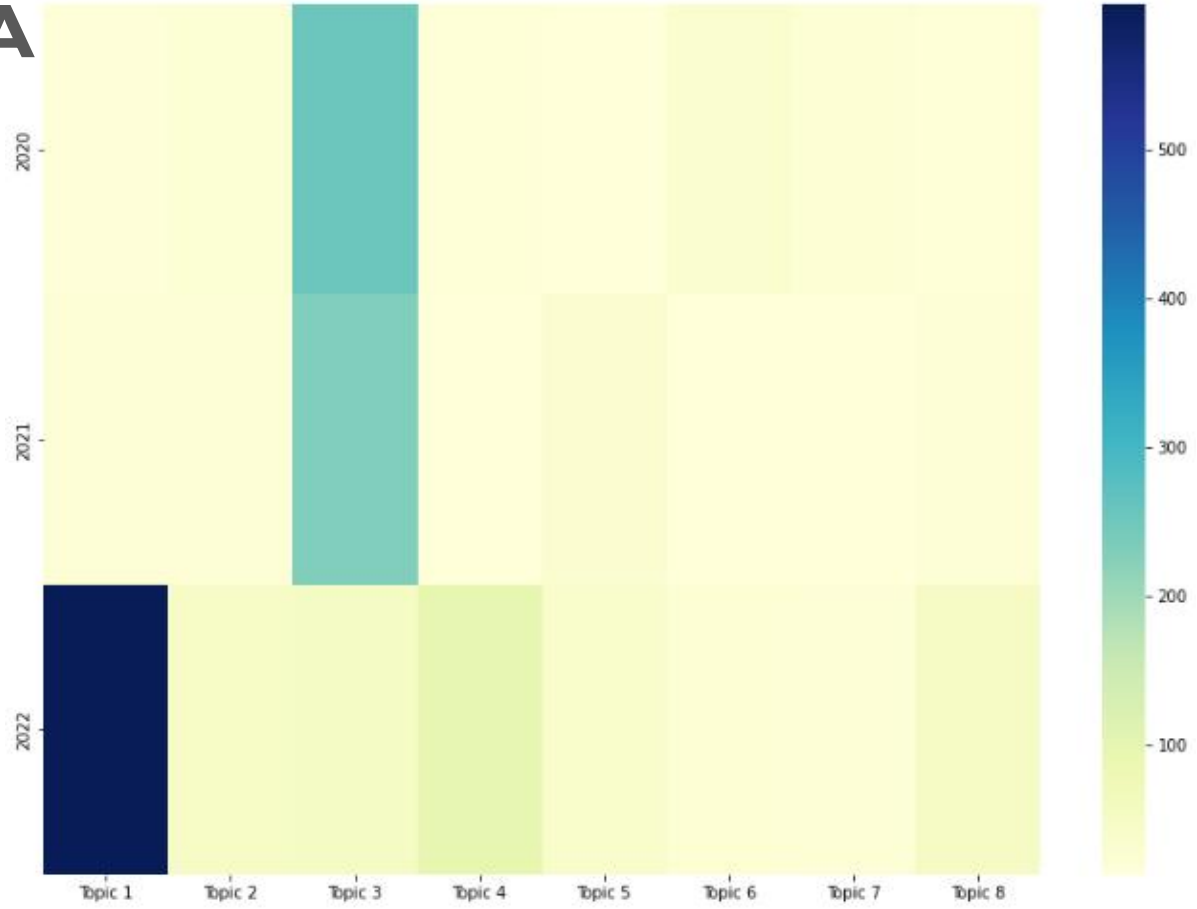
(III) This is a much better result. It would seem that **LDA is more effective than LSA in separating out the topic categories**.



LSA & LDA

(I) The heatmap helps to analyze the distribution or patterns of topics over the years, allowing us to identify trends or changes in topic prevalence.

(II) The evolution across time are shown using a heatmap. It shows that Topic 1 is the most popular topic.



LSA & LDA

Evaluation Metrics

(I) For LSA, **cosine similarity metric** is used to evaluate the performance of the model:
-**Dimensionality reduction** are performed to calculate the **cosine similarity** which provides a measure of the average similarity between pairs of vectors in the reduced matrix. It indicates similarity between the data points in the reduced space.

```
Mean Cosine Similarity: 0.33433140354372787
```

- The mean cosine similarity of 0.33 shows a **low mean cosine similarity indicates more diversity** or dissimilarity.

(II) For LDA, perplexity and coherence score for LDA is obtained via Gensim
- Perplexity measures how **well the model predicts the observed data**
- Coherence measures the **semantic coherence of the topics generated**

```
Perplexity: -7.0130244085139
Coherence Score: 0.5165183261215247
```

- Perplexity of -7.01 shows a **low perplexity score indicates better predictive performance**
- Coherence Score of 0.52 shows a **moderate coherence score** means that the topics generated by the LDA model **are not perfectly clear and distinct**, but they are not entirely random or unrelated either, a higher coherence score would be better as it indicates more interpretable and meaningful topics

BERTOPIC

(Bidirectional Encoder Representations from Transformers)

To extract meaningful topics from a collection of documents

BERTopic

(I) Columns of 'titles' and 'contents' are extracted and form a new dataframe named df_new

```
columns_to_keep = ['titles', 'contents']
```

```
df_new = df.drop(df.columns.difference(columns_to_keep), axis=1)
df_new
```

	titles	contents
0	Worth every penny	This is a must have if you are in a rainy area...
1	comfortable and completely water proof	very wet rainy season this year, so had ample ...
2	Muy útil	Cómoda , resiste el frío y el agua
3	Very Nice	Fits well and great materials and construction
4	Different fit, cheaper design	I am replacing the original venture jacket. Th...
...	...	...
3680	Finally a hat that looks decent on me	so small. should be for children. no full grow...
3681	The North Face Logo is easily removed!!	Perfect fit!
3682	this hat was way too small	I love this hat! I have a shaved head and glas...
3683	Finally a hat that looks decent on me	Great Hat!\nThe North Face logo label on the f...
3684	Loved it	Bucket hat with SPF and a pocket! The pocket w...

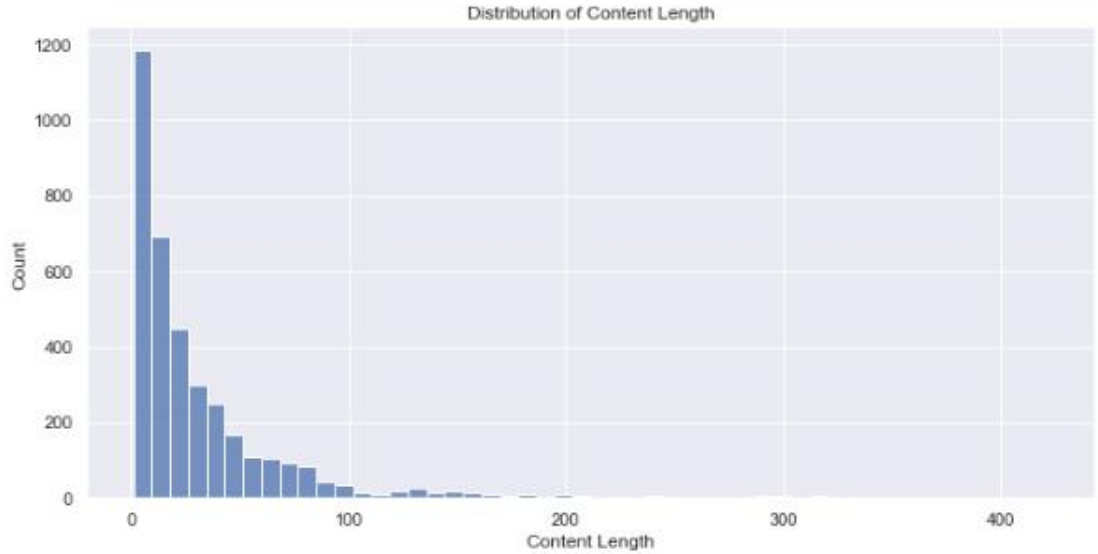
3685 rows × 2 columns

```
# Create a new column containing the length each headline text
df_new["contents_len"] = df_new["contents"].apply(lambda x : len(x.split()))
df_new["contents_len"]
```

```
0        45
1        25
2         8
3         7
4       151
       ...
3680     19
3681      2
3682     81
3683     66
3684     45
Name: contents_len, Length: 3685, dtype: int64
```

BERTopic

<Figure size 2000x1200 with 0 Axes>



A histogram is generated to visualize the distribution of content lengths in the DataFrame df_new. It provides insights into the frequency of different content length ranges and allows for the analysis of patterns or outliers in the data.

BERTopic

```
print("The longest contents has: {} words".format(df_new.contents_len.max()))
```

```
The longest contents has: 422 words
```

```
# Let have a look at some reviews and the longest one
for idx in df_new.sample(3).index:
    reviews = df_new.iloc[idx]
    print("Reviews #{}:".format(idx))
    print("Titles: {}".format(reviews.titles))
    print("Contents: {}\n".format(reviews.contents))

```

```
Reviews #1200:
Titles: Image incorrect
Contents: Oh, North Face, Excellent clothing ranges, Jackets Tops Everything is the Best Quality.

This Latest range is So Soft and comfortable and Warm to wear, Prefer the zip top but these Buttons are like a Polo shirt in wi
nter.

I got this Red one for Christmas time, got in July before the Winter rush, but am waiting for my blue & lime colored ones to ar
rive now

Reviews #1997:
Titles: Best house shoe!
Contents: I can't say enough about how great these are. Wear around the house and you can walk outside and grab the mail in the
m. I wear them when we have people over for game night and I get asked where did you get those all the time. I wear anywhere fr
om an 11 to a 12 in shoes and I got a 12 and they fit perfect.. I highly recommend!

Reviews #491:
Titles: Nice but runs large
Contents: Fits well and great materials and construction
```

BERTopic

UMAP - Uniform Manifold Approximation and Projection

Dimensionality reduction technique that is commonly used for visualizing high-dimensional data

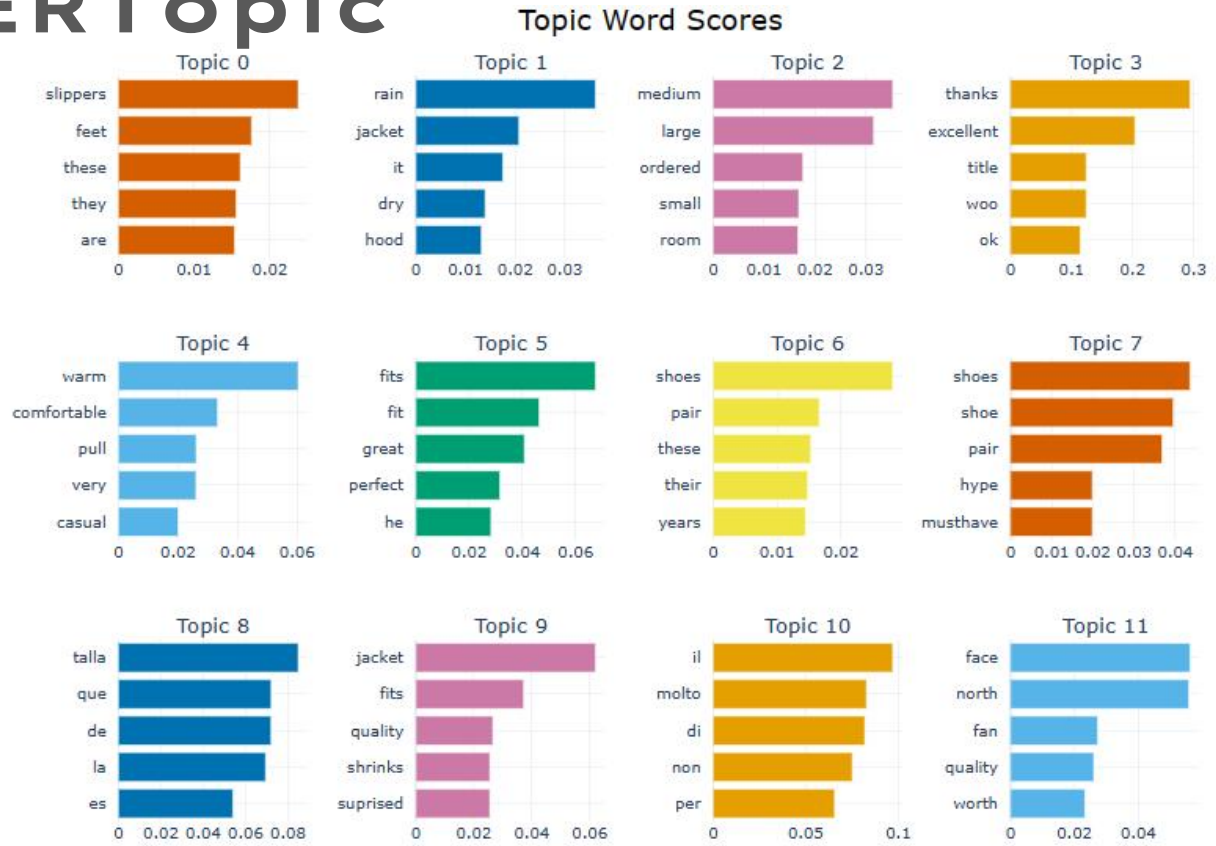
Extract Topics From Topic Modelling

```
# Get the list of topics
topic_model.get_topic_info()
```

	Topic	Count	Name	Representation	Representative_Docs
0	-1	927	-1_the_and_to_it	[the, and, to, it, is, this, in, for, my, are]	[These shoes came early, and fit right out of ...
1	0	235	0_slippers_feet_these_they	[slippers, feet, these, they, are, them, warm,...	[I have only had these she for a couple weeks ...
2	1	159	1_rain_jacket_it_dry	[rain, jacket, it, dry, hood, pocket, lightwei...	[I was looking for a lightweight rain jacket t...
3	2	87	2_medium_large_ordered_small	[medium, large, ordered, small, room, arms, hi...	[Love this coat and the color. I normally wear...
4	3	86	3_thanks_excellent_title_woo	[thanks, excellent, title, woo, ok, good, na, ...	[Thanks, Thanks, Thanks]
...	...	...	...	...	...
90	89	11	89_quality_excelent_originals_approved	[quality, excelent, originals, approved, total...	[Totally approved great price. Good quality! O...
91	90	11	90_parent_vocal_woke_bud	[parent, vocal, woke, bud, participate, charad...	[I am returning it. Will truly miss its comfor...
92	91	11	91_adhesive_logo_hat_off	[adhesive, logo, hat, off, north, letter, tool...	[Great Hat!\nThe North Face logo label on the ...
93	92	11	92_jacket_patagonia_chile_apex	[jacket, patagonia, chile, apex, bionic, idea,...	[I purchased this jacket prior to a trip to so...
94	93	10	93_price_high_little_cheap	[price, high, little, cheap, penny, value, eac...	[Price was a little👉 high, Price was a little👉 ...

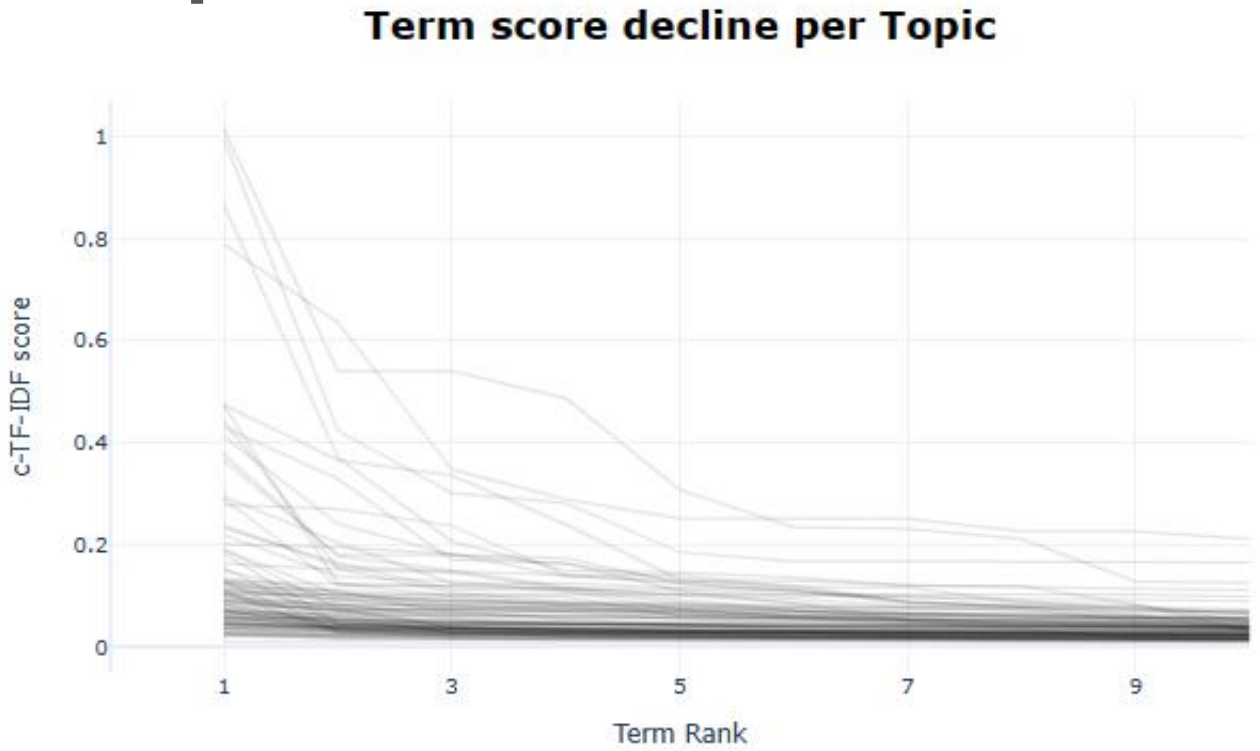
95 rows × 5 columns

BERTopic



(I) A bar chart visualization of the top N topics is generated
(II) It shows the importance or prevalence of each topic

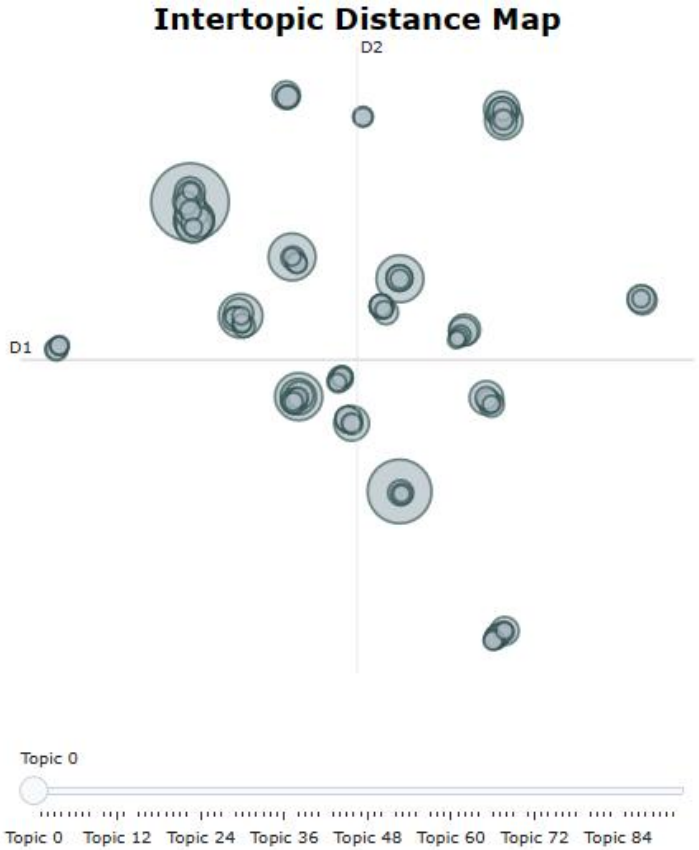
BERTopic



(I) A visualization of the term ranks within each topic in the model is generated
(II) It provides insights into the importance or relevance of individual terms within each topic by showing their ranks.

BERTopic

Topic Similarities

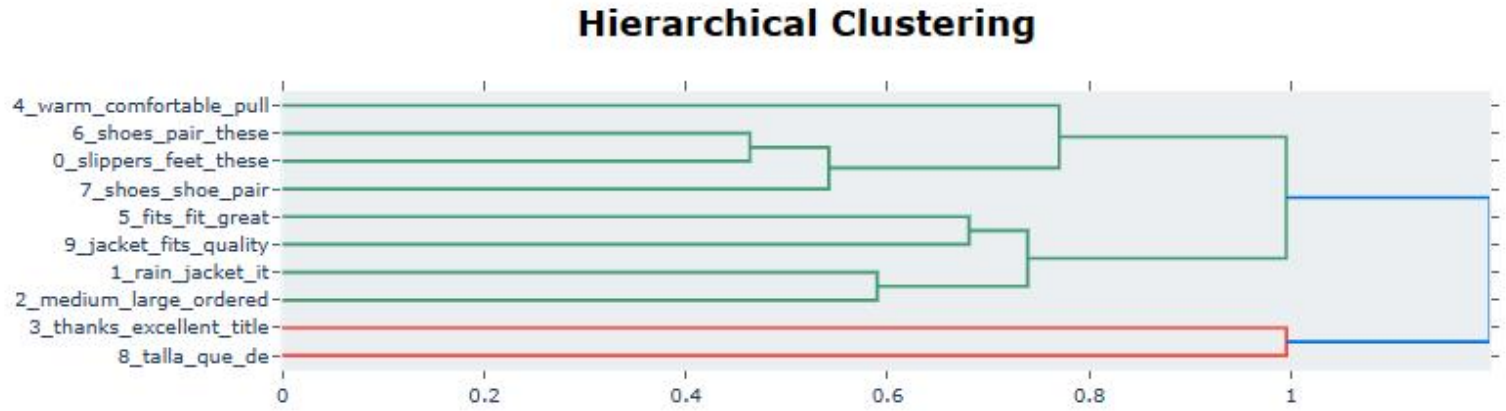


(I) **Intertopic distance map** measures the **distance between topics**. Similar topics are closer to each other, and very different topics are far from each other. From the visualization, there are **17 distinct topics**. Topics with **similar semantic meanings** are in the same topic group.

(II) The size of the circle represents the **number of documents** in the topics, and larger circles mean that **more reviews belong to the topic.**

BERTopic

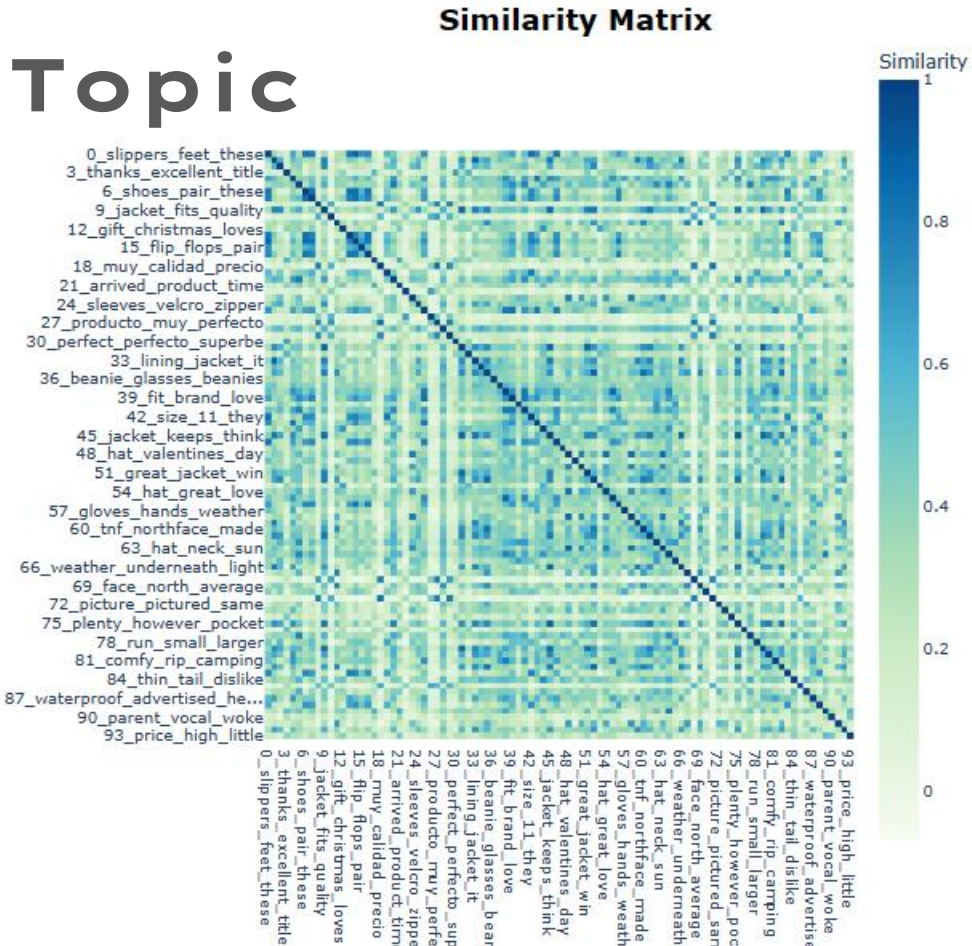
Topic Similarities



(I) Hierarchical clustering graph above shows how the topics are connected.

(II) Top 10 topics are included in the hierarchical graph. It shows that (4)warm and comfortable topic is closely connected to (7)shoes and shoe topics, which the latter is connected to the (6)shoes and pair, together with (0)slippers and feet topics as well.

BERTopic



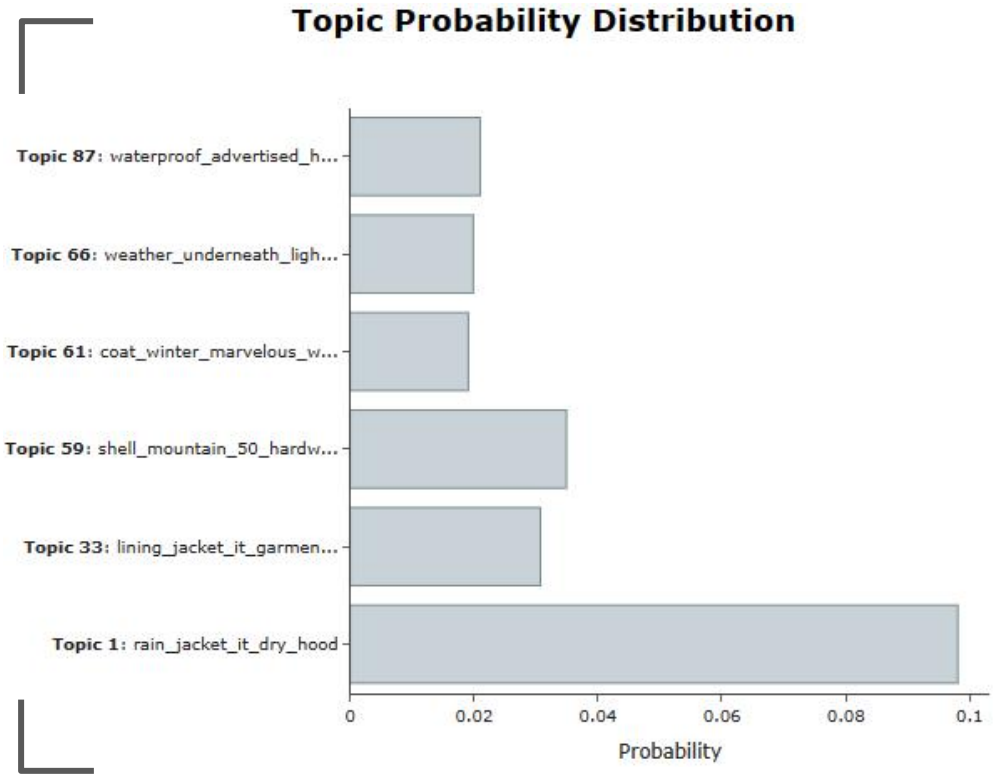
Topic Similarities

(I) Heatmap can also be used to analyze the similarities between topics.

(II) The similarity score ranges from 0 to 1. A value close to 1 represents a higher similarity between the two topics, which is represented by darker blue color.

BERTopic

Topic Model Predicted Probabilities

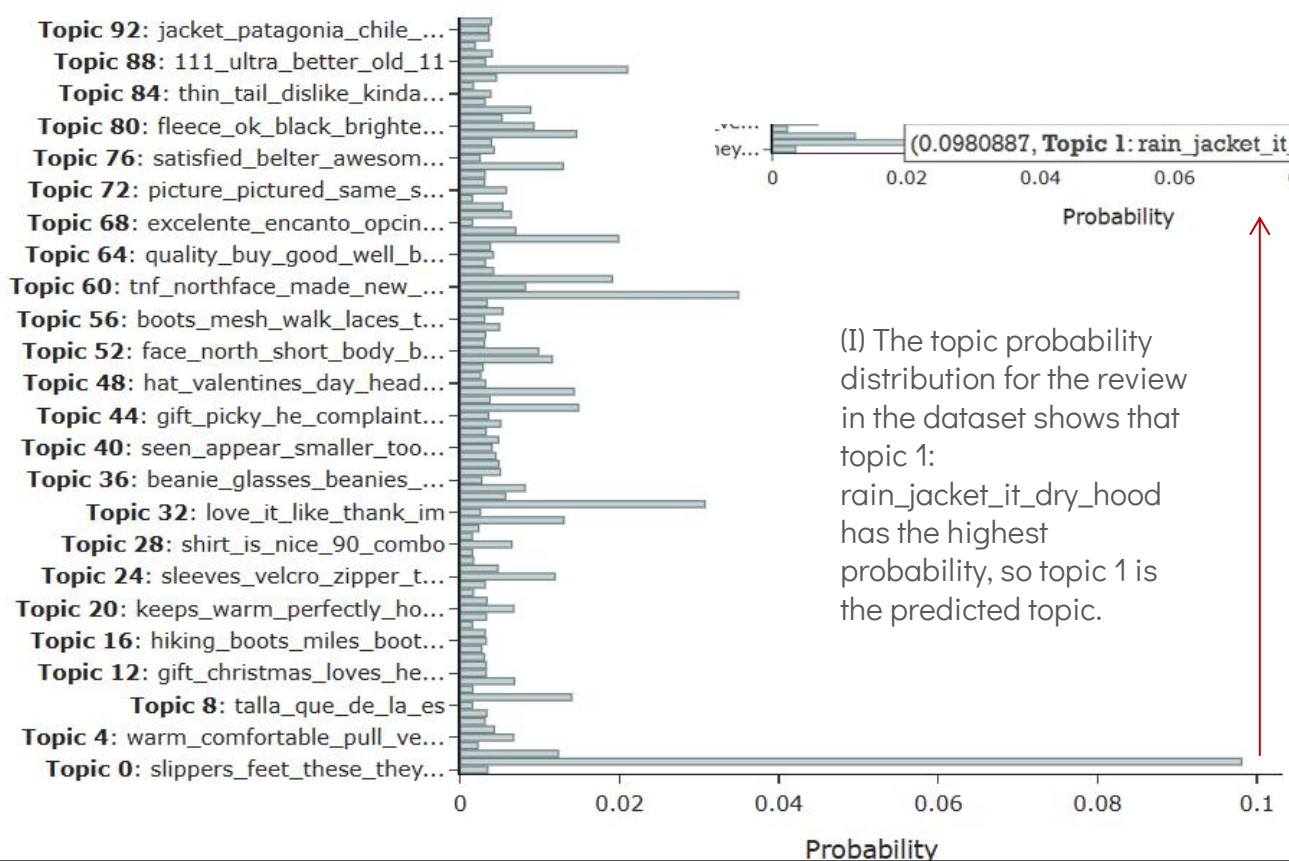


(I) The topic prediction is based on the **predicted probabilities** of the document. The topic with **the highest probability is the predicted topic**.

(II) It has the default probability threshold of 0.015, so only the topic with a **probability greater than 0.015 will be included.**

BERTopic

Topic Probability Distribution



(I) The topic probability distribution for the review in the dataset shows that topic 1: rain_jacket_it_dry_hood has the highest probability, so topic 1 is the predicted topic.

BERTopic

```
1 # Check the content for the first review
2 df_new['contents'][0]
```

```
'This is a must have if you are in a rainy area. I will make sure to mention that if you want a more tighter fit to go down a s
ize. I have already used it 3 times under heavy rain and it works great.'
```

- The first review is 'This is a must have if you are in a rainy area. I will make sure to mention that if you want a more tighter fit to go down a size. I have already used it 3 times under heavy rain and it works great.', and the topic of 'rainy' is pretty relevant.

(I) From the Topic Probability Distribution, we noticed that **topic 1: rain_jacket_it_dry_hood** has the highest probability, this corresponds to the first review, which is in the **topic of 'rain'**

BERTopic

```
1 max_index = np.argmax(topic_model.probabilities_[0])
2
3 print("Index of the highest value:", max_index)
4 print("Highest value:", probabilities[max_index])
```

```
Index of the highest value: 1
Highest value: [1.49343874e-03 7.48123057e-01 5.46317462e-03 1.01385018e-03
 2.81311008e-03 1.93093271e-03 1.35324551e-03 1.45353906e-03
 7.13825019e-04 6.16212246e-03 7.26837770e-04 2.76311622e-03
 1.48247077e-03 1.41881974e-03 1.32790758e-03 1.20538234e-03
 1.41875266e-03 1.41904213e-03 7.09806599e-04 1.43473082e-03
 2.92764031e-03 1.52797570e-03 7.63498057e-04 1.49289621e-03
 5.46707564e-03 2.04172981e-03 7.67050102e-04 7.13562897e-04
 2.97734979e-03 6.98267650e-04 1.06094552e-03 5.47479090e-03
 1.15657734e-03 1.80901837e-02 2.68614309e-03 3.37885890e-03
 1.27144693e-03 2.25147283e-03 2.11806646e-03 1.95924187e-03
 1.76856323e-03 2.10222339e-03 1.41658207e-03 2.17935501e-03
 1.62310282e-03 6.47658930e-03 1.62905211e-03 5.51494542e-03
 1.51998670e-03 1.14270039e-03 1.36844744e-03 4.78104665e-03
 4.19625046e-03 1.38801561e-03 1.51675576e-03 2.14181354e-03
 1.33588726e-03 2.46064823e-03 1.62479162e-03 1.80983187e-02
 3.41243974e-03 9.34233695e-03 1.85517355e-03 1.49996407e-03
 1.86382075e-03 1.66961108e-03 8.45237403e-03 2.83974201e-03
 7.10000511e-04 2.95060779e-03 2.45738103e-03 7.06806125e-04
 2.62818252e-03 1.45948030e-03 1.41493540e-03 5.51304136e-03
 1.13758428e-03 1.92134767e-03 1.72152039e-03 6.40211188e-03
 4.24867095e-03 2.26813321e-03 3.68897778e-03 1.36545761e-03
 1.73496892e-03 7.64988548e-04 2.02656585e-03 7.22549151e-03
 1.38203053e-03 1.80531776e-03 8.66349839e-04 1.75122150e-03
 1.61517936e-03 1.76117415e-03]
```

- There are 94 probability values, one for each topic. The index 1 has the highest value of 7.48123057e-01, indicating that topic 1 is the predicted topic with the most dominant or relevant topic based on the input.
- The overall probability distribution shows variations in the probabilities across different topics. Some topics have higher probabilities, while others have lower probabilities, indicating the varying importance or relevance of different topics within the dataset or model.
- The sum of all probability values is approximately 1, indicating that the probabilities represent a valid probability distribution.

BERTopic

Scatter Plot

(I) Topics visualized by reducing sentenced embeddings to 2-dimensional space.
(II) Outliers are shown as small dots in a neutral color, while clustered points are colored based on their assigned cluster labels
(III) It is difficult to visualize the individual clusters due to the number of topics generated. However, it shows that even in 2-dimensional space some local structure is kept..

BERTopic

	Topic	Size
0	-1	1361
2	1	422
14	13	284
10	9	223
39	38	88
1	0	86
19	18	70
48	47	55
44	43	54
12	11	51

(I) To calculate the sizes of each topic in the DataFrame that contains document-topic assignments.

(II) It shows the topic index and the number of documents in each topic.

Topic Representation

```
1 top_n_words[1][:10]
```

```
[('la', 0.08969868011589092),
 ('que', 0.06774281843846568),
 ('es', 0.05861986528903239),
 ('el', 0.058484329485807965),
 ('en', 0.0566251256334775),
 ('muy', 0.05140561416968078),
 ('los', 0.03978664242414283),
 ('para', 0.03692900869375973),
 ('se', 0.03598444626515059),
 ('calidad', 0.03442878426958464)]
```

```
1 top_n_words[13][:10]
```

```
[('loves', 0.08269942271690822),
 ('husband', 0.08247596111198593),
 ('son', 0.06412805051929543),
 ('bought', 0.04527389307223432),
 ('great', 0.032535175893653866),
 ('christmas', 0.032500272545866335),
 ('gift', 0.029375466053784736),
 ('warm', 0.02936713375803458),
 ('wears', 0.02881328641813684),
 ('loved', 0.026437919448406264)]
```

```
1 top_n_words[9][:10]
```

```
[('small', 0.07839082867553959),
 ('tight', 0.06532899700045654),
 ('smaller', 0.04951529547311524),
 ('little', 0.04826716056701254),
 ('nice', 0.04681681283722384),
 ('like', 0.04260454351388943),
 ('quality', 0.039934976937918934),
 ('jacket', 0.034745854962684196),
 ('waist', 0.03465950576125692),
 ('really', 0.032335646440638736)]
```

BERTopic

Model Testing & Evaluation

1	coherence

0.47526544564719103

(I) Coherence refers to the measure of semantic similarity between the top words within each topic.

(II) A coherence score of 0.48 suggests that the generated topics have a moderate level of semantic coherence, meaning that the top words within each topic share some degree of semantic relatedness.


STREAMLIT

Amazon The North Face Sentiment Analysis & Topic Modelling NLP App

Streamlit Projects

Home

Enter Text Here

I love it!

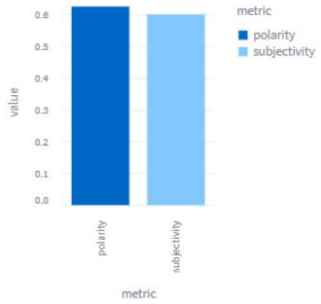
Analyze

Results


```
{
  "polarity" : 0.625
  "subjectivity" : 0.6
}
```

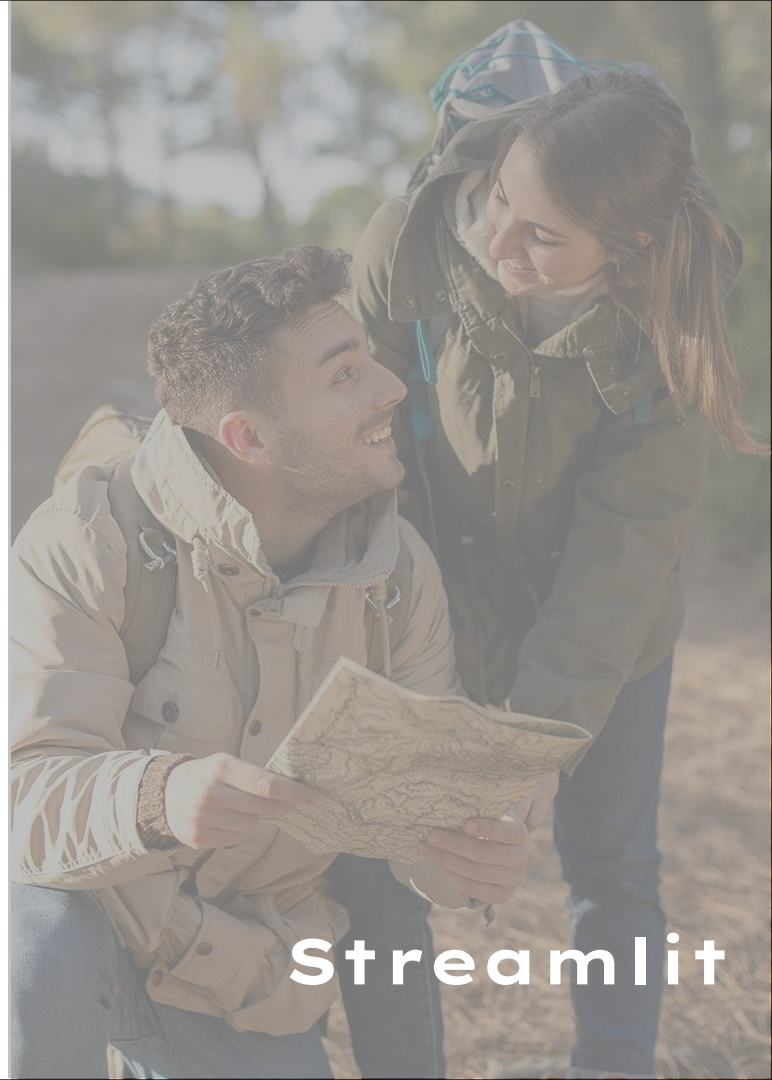

Sentiment: Positive 😀

	metric	value
0	polarity	0.625
1	subjectivity	0.6



Token Sentiment


```
{
  "positives" : [
    0 : "love"
    1 : 0.6369
  ]
  "negatives" : []
  "neutral" : [
    0 : "I"
    1 : "it!"
  ]
}
```


CONCLUSION

Sentiment Analysis:
- RoBERTa is a better model compared to Vader

Topic Modelling:
- BERTopic ranks first in the effectiveness for prediction, followed by LDA, then LSA

NEVER STOP EXPLORING

THANK YOU!