Zahra Jalil Pour

Email: h19zahja@du.se

Final Project

Using NLP approach as an efficient way to analysis costumer reviews

Abstract__As the online marketing and prevalence of social media have been popular in recent decades, reviewing site become one of the vital and is a place for consumers to express their feelings about people, businesses, products, and services [1]. Online sellers want their purchasers to share their opinion about the quality of products or services. It has a powerful economic effect on both users and companies. These reviews and opinions are represented in two formats: structured data (numeric rating) and unstructured data (textual comments)[2]. Everyday thousands of reviews are added and make a big data that should be read and understand by decision makers. Hence opinion mining is a technique to analyzes these reviews. Trustpilot is one of the famous review sites and used daily by millions of users. But then, it still lacks functionality and capable of providing better insights if there is a better system. Here we want to focus on creating a better review system with the help of N-gram, Named Entity Recognition, Sentiment Analysis, and clustering techniques, which is implemented by using Python and its libraries. In this paper, we used web scraping to extract the Rating, Title, and Reviews from the Lufthansa and Nike pages in Truspilot.com. We used tools like Phyton, which was very easy and convenient.

1-Introduction

In recent decades, online market places have growing rapidly and there is a need that consumers share their opinions about the products that they have bought. Nowadays the huge number of reviews and opinions, are expressed by users as comments, reviews and tweets about the products and service. Although the number of available reviews grows, making a good decision to buy a product is becoming difficult too. Different opinions about the same product on one hand and ambiguous reviews on the other hand makes customers more confused to get the right decision[3]. Hence analyzing the content of these reviews has become crucial. Sentiment analysis, also called opinion mining, is attempts to solve this problem by analyzing opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes.

Trustpilot.com is a place where consumers express their feelings about business and their products. In this site, people can read thousands of reviews dropped by other customers about their desired products [4]. The website divided into many pages, and each page owned by a company where customers can write their reviews directly to the company itself. These reviews provide valuable opinions about a product such as its property, quality and recommendations which helps the purchasers to understand almost every detail of a product. This is not only beneficial for consumers but also helps sellers who are manufacturing their own products to understand the consumers and their needs better. There are a few disadvantages in this review system created by Trustpilot.com. Researchers note that the billions of publications left by people monthly, cannot be processed manually by holding public opinion polls.

The review system created by Trustpilot.com doesn't give full control to the users and businesses to explore the website when comparing to Amazon review section. The reason is that the users have to go through thousands of reviews to find a relevant content that is useful for them. This fact highlights the need for automated methods of intellectual analysis of text information, what allows in a short time to process large amounts of data and to understand the meaning of user messages. This understanding of the meaning of messages is the most important and complex element of the automated processing. We proposed a few solutions such as N-gram, Sentiment analysis and

clustering which will make the system works more efficient and provide better experiences for the users and the businesses.

The objective of this project is to build an efficient review system using natural language processing which occupied with purposed solution such as clustering, Named Entity Recognition, using N-gram and sentiment analysis system that can help users and businesses. We have used a dataset from trustpilot.com, an online review webpage, and selected Lufthansa and Nike page as our primary dataset for this experiment.

In this paper, we used web scraping to extract the Rating, Title, and Reviews from the Lufthansa and Nike pages in Truspilot.com. We used tools like Python and a few libraries in the Python programming language which was very easy and convenient.

2-Description of the Data

The first step for conducting the research includes data collection. In our work we use a two distinct annotated datasets obtained from reviews on different subjects: The datasets used in this paper were Lufthansa and Nike reviews in Truspilot.com. We implemented the Python code to scrape all the comments, titles, and ratings from these pages. The file was converted to the Comma Separated Values (CSV) format, as it is more convenient for python to handle this type of files. The data set of Lufthansa consists of 580 reviews and 1448 reviews for Nike. Each review includes four features as table1.

	Rating	Title	Comment	time
0	5	I don’t usually leave reviews but I…	I don’t usually leave reviews but I ...	2020-05-18
1	5	Good Customer Service	I ordered 3 pairs of shoes April 30. As ex...	2020-05-18
2	1	If there's the option to give 0 star	If there's the option to give 0 star, ...	2020-05-18
3	1	Do NOT order from Nike.com	Nike sent me the ugliest shoes I've ev...	2020-05-18
4	1	About 2 weeks ago	About 2 weeks ago, I've been trying to...	2020-05-18
...	...	...	...	...
1443	5	Exceptional service as standard	I ordered some items which unfortunately d...	2011-08-11
1444	5	Great Trainers	Recieved a fantastic pair of trainers here...	2010-11-13
1445	3	Great Product, But out of stock often	This site is the bomb! but quite often I f...	2010-11-11
1446	4	I bought nice hiking boots for my kids	I bought nice hiking boots for my kids, an...	2010-04-27
1447	5	Awesome	I love this website, the stuff is really g...	2009-12-03

1448 rows × 4 columns

table1- Describes features of each product

3-Method

In this project, we are using two famous NLP libraries for text processing, spaCy and NLTK.

SpaCy*:*

SpaCy is a library for advanced Natural Language Processing in Python and Cython. SpaCy comes with pretrained statistical models and word vectors, and currently support tokenization for 50+ languages. It features state-of-the-art speed, convolution neural network models for tagging, parsing and named entity recognition and easy deep learning integration. It's commercial open-source software, released under the MIT license[5].

NLTK:

Natural Language Toolkit (NLTK) is a well-known open-source library written in Python. NLTK developed at the University of Pennsylvania for natural language processing and is accessible for Windows, Mac, and Linux users. Natural Language Toolkit mainly used as a teaching or personal study tool due to its comprehensive API documentation alongside a hands-on guide for fundamental programming. Right now, more than 32 universities only in the United States, and approximately 25 countries are using Natural language Toolkit in their universities' syllabus [6] [7]. NLTK contains text processing libraries such as tokenization, POS tagging, stemming, and over 50 corpora and lexical resources, namely, Multilingual word net [7].

NLTK and SpaCy use different method to tokenize the sentences. For instance, NLTK considers the punctuations as a token; however, SpaCy does not. In SpaCy, if there is no space between the punctuation and the previous word, then it will consider both of them as a single token. However, if there is a space between the punctuation and the previous word, it will consider them as a separate token. The table below shows the

difference between NLTK and SpaCy tokenization.

```
: str = "This is a sample text for 522 project.
  nltk.word_tokenize(str)

: ['This',
   'is',
   'a',
   'sample',
   'text',
   'for',
   '522',
   'project',
and output; double click to hide output
   'Contrary',
   'to',
   '?',
   'popular',
   '!',
   'belief',
   ',',
   'Lorem',
   'Ipsum',
   'is',
   'not',
   'simply',
   'random',
   'text',
   '.']
```

Fig. 1. Tokenization using NLTK

```
from spacy.tokenizer import Tokenizer
from spacy.lang.en import English
nlp = English()
tokenizer = Tokenizer(nlp.vocab)
tokens = tokenizer(str)
for token in tokens:
    print(token)

This
is
a
sample
text
for
522
project.
Contrary
to?
popular
!
belief,
Lorem
Ipsum
is
not
simply
```

Fig. 2. Tokenization using spaCy

N-gram

N-grams are contiguous sequences of n-items in a sentence. N can be 1, 2, or any other positive integers, although usually, we do not consider very large N because those n-grams rarely appear in many different places [8]. When performing machine learning tasks related to natural language processing, we usually need to generate n-grams from input sentences. For example, in the text classification tasks, in addition to using each individual token found in the corpus, we may want to add bi-grams or tri-grams as features to represent our documents [8].

In this paper, we have used bigram and create 'mention' in review system. We will list down the most used two words together, which coming together in one section, such as "Customer Service," "Nike Product." When the users click on these, it will show the comments that have the words they choose. Although Trustpilot is one of the famous review websites, it does not have this mention system. This system can make users' life more comfortable, and they do not need to go through thousands of reviews. Bigram keywords can really help the users know what it is in before they go through the reviews. To demonstrate, we tokenize the comments from the selected website page, in this case, Nike. From here, we can see that the most frequent tokenize word that is used shown in this bar chart. There are also punctuation and stop words indicate the highest number in this bar chart.

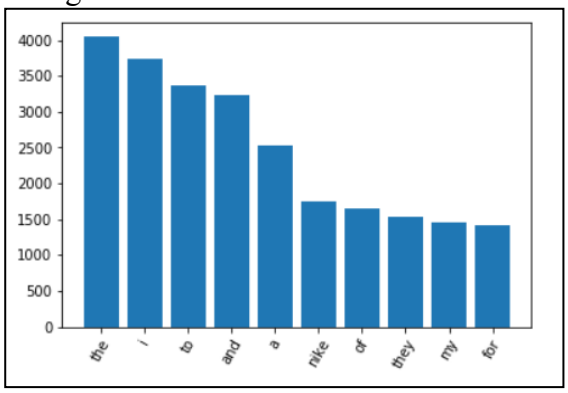

Fig. 3. Most frequent words after tokenization

We choose bigram as our primary tool for this experiment. In bigram, the tokenize words are put together, such as the table below. A bigram is a sequence of two adjacent elements from a string of tokens, which are typically letters or words. Not only that, but it also might contain punctuation too. In this paper, we can clearly see that the number of punctuations is higher in the above figure. To solve this problem, we created a data frame and

divided it into two columns that are bigram and punctDetect. The function of punctDetect is to detect whether there is any punctuation in the bigram, if yes, it will return TRUE.

```
('I', 'dont')
False
('I', 'dont')
False
('dont', 'usually')
False
('dont', 'usually')
False
('usually', 'leave')
False
('usually', 'leave')
False
('leave', 'reviews')
False
('leave', 'reviews')
False
('reviews', 'but')
False
('reviews', 'but')
False
```

Fig. 4. Bigrams that contain punctuation tagged in this figure

By using this method, it will be easy to identify the bigram, which contains punctuation, and we had dropped that bigrams.

Named Entity Recognition

In this section, we used Named Entity Recognition to help the business owner to understand more about their business and find important information about their companies way easier. Named Entity Recognition can detect different factors in a text such as an organization's name, time, money, and countries. Name Entity Recognition is part of the SpaCy library that designed to understand large volumes of text and highlight critical information. The dataset contains telephone conversations, newsgroups, broadcast conversation and weblogs to identify the different entity types that previously mentioned such as Countries, Cities, and States [9].

This will help the business to understand the review much faster by selecting specific keywords or entities they need. This will help them focus more on problem-solving and reduce the effort to go through all the reviews at one time.

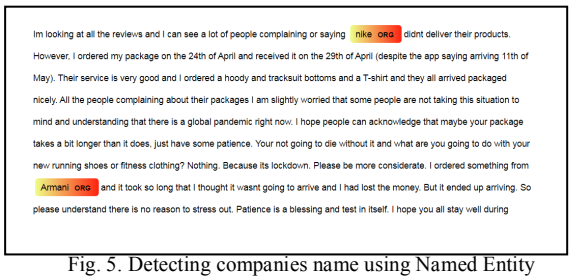

Fig. 5. Detecting companies name using Named Entity recognition

As we can see in the previous figure, we select to highlight ORG (Companies, agencies, institutions, etc.) Only by skimming, we are able to see that the user mentioned the Nike and Armani inside the comments (the comments belongs to Nike). In this case, they can go through that part of the comment and see why the user mentioned their competitors.

In the figure below, we can see another example using another type called "Date." In this case, companies can extract a lot of meaningful information. For instance, if the company collect all negative reviews and use date type to highlight them. They can easily see what time of the year or month customers mostly faced a problem.

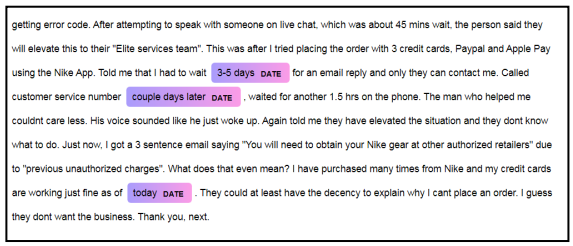

Fig. 6. Detecting time using Named Entity recognition

Sentiment Analysis

Sentiment Analysis is also a part of Natural Language Processing (NLP) that is used to identify and extract opinions within a given text or corpus. The objective of this analysis is to understand the attitudes and emotions of the users based on the computational treatment in the given text.

We are using sentiment analysis here for two reasons. Throughout this analysis, we have found out that there are good reviews, but the rating is one. This is because the Trustpilot rating system has a default rating that is one star. To solve this

problem, we have used sentiment analysis in this project. The analysis will detect whether the comments written by the users is positive or negative, then it will give a pop-out message to confirm the user rating. The second reason is for the business. Sentiment Analysis enables companies to make sense of data and provide insights that can improve their business performance. By applying this analysis, they are able to study insights from a vast unstructured dataset without having to indulge in it manually. In this case, we are using Varder Sentiment Analysis that is an open-source library belongs to MIT. For sentiment rating, they ask ten people to rate a word for instance, "Good." Then the raters can score the word based on their opinion the score should be between 4 and -4. 4 as a score means the word is completely positive and -4 it means completely negative. Then they will calculate the overall score for the word after collecting the scores. They asked the raters to give the score for different words and even emojis. The only thing that is important that the standard deviation should not be more 2.5. If the standard deviation is more than 2.5 it means some of the voters believe the word has a completely positive meaning however the rest think completely opposite [11].

278	j/j	2.0	1.34164	[4, 1, 1, 1, 1, 4, 4, 1, 2, 1]
279	j/k	1.6	1.2	[1, 2, 1, 3, 0, 0, 2, 2, 1, 4]
280	j/p	1.4	0.66332	[1, 1, 0, 2, 1, 2, 2, 2, 1, 2]
281	j/t	-0.2	1.46969	[1, -1, -1, -2, 1, 1, 2, -2, 1, -2]
282	j/w	1.0	1.0	[1, 1, 1, 3, 0, 0, 0, 2, 0, 2]
283	j4f	1.4	0.8	[2, 1, 1, 0, 3, 1, 1, 1, 2, 2]
284	j4g	1.7	1.18743	[1, 4, 1, 1, 3, 1, 3, 0, 2, 1]
285	jho	0.8	0.4	[1, 1, 1, 1, 0, 1, 1, 1, 0, 1]
286	jhomf	1.0	0.63246	[1, 1, 1, 0, 1, 0, 2, 2, 1, 1]
287	jj	1.0	0.63246	[1, 1, 1, 1, 2, 0, 2, 1, 1, 0]
288	jk	0.9	1.22066	[1, 0, 0, 1, 0, 0, 2, 1, 4, 0]
289	jn	0.8	0.74833	[1, 1, 1, 0, 2, 0, 2, 0, 1, 0]

Fig 7. Sentiment analysis scores [12]

Below is the bar chart; we extract the review data from the Lufthansa page to find out the number of ratings. Through this, we can conclude that Lufthansa has the highest number of rating 1, and this might be the users are not really happy with services that provide by the company.

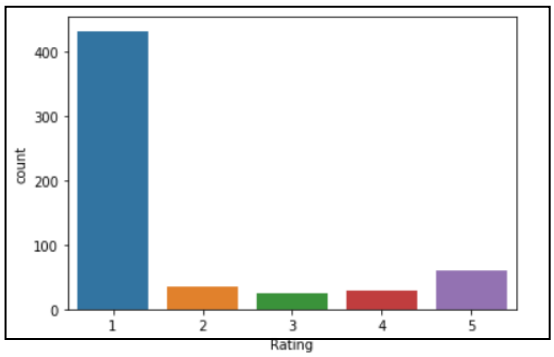

Fig. 8. Lufthansa rating

The bar chart below shows the lengths of the comments written by the user are mostly between 100 and 2000. It also shows the users write a longer review when they feel disappointed and angry. This is to express their frustration by writing a review.

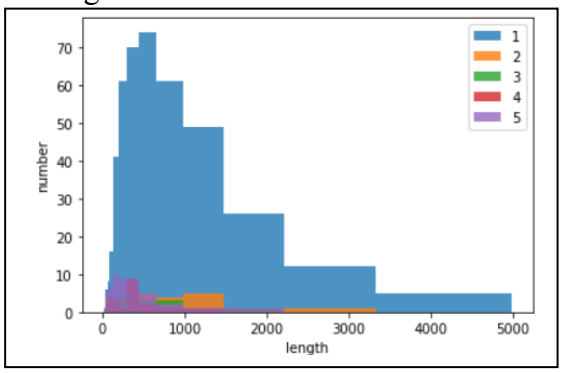

Fig. 9. Rating vs. length of the words

In this case, we applied sentiment analysis on the Lufthansa dataset. After applying the sentiment analysis on each comment using the NLTK library, it returns four values that measure the feelings in text. Negative, Positive, neutral, and compound that is the overall value. We extracted the compound and stored it in the separate column in our dataset. The value of the compound is between 1(Positive) and -1 (Negative).

```
In [21]: dataset.Comment[0]

Out[21]: '    Totally unhelpful woman on the phone. For the first time I have received that bad service on the
         ir site. She did not want at all listen my viewpoint.  Even in a period of crisis you need to be resp
         ectful. I do not know why those people do this kind of work if they can not offer the basic rules of
         communication. Too bad.'

In [22]: dataset.scores[0]

Out[22]: {'neg': 0.175, 'neu': 0.782, 'pos': 0.043, 'compound': -0.8527}
```

Fig. 10. Applying Sentiment Analysis on a comment

The joint plot below can clearly show the relationship between the ratings and the feelings using the Lufthansa dataset. In

this case, the company can easily see the amount of negative emotions. People get easily angry if something wrong happens for their flights. This is why we see the intense negative feelings.

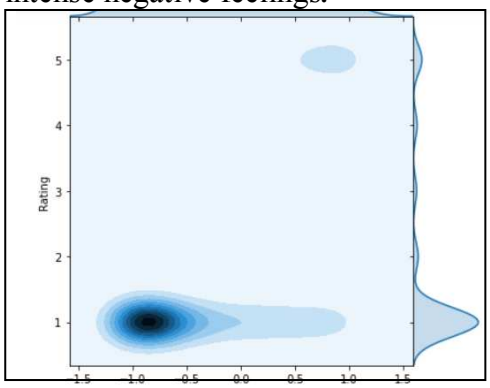

Fig.11. Using joint plot to demonstrate the relation between the feeling of the Lufthansa users and their rating

We came out with threshold 0.1 and -0.1 to change the compound to positive, negative, and neutral if the compound value is more than 0.1 (Positive), between 0.1 and -0.1 (neutral) and below -0.1 (Negative). We used wordCloud for both positive and negative comments; then, we dropped the Lufthansa, flight, and airline word to have a better picture.

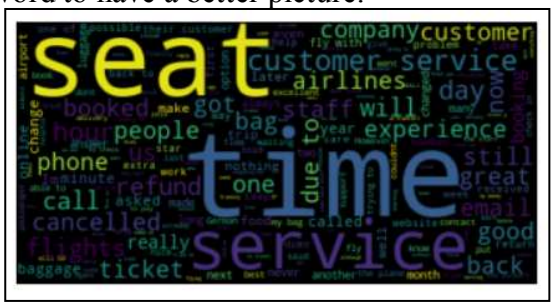

Fig. 12. Positive comments word cloud (Lufthansa dataset)

Fig. 13. Negative comments word cloud (Lufthansa dataset)

Clustering using LDA

Topic modeling is an unsupervised machine learning method that allows us to efficiently analyze significant volumes of text by clustering documents into topics. So the main goal to apply topic modeling to identify a cluster of similar comments grouped by topic.

In this case, we are using the Spacy toolkit and LDA (Latent Dirichlet Allocation) library that is based on probability distribution. So basically, LDA assumes that the comments with similar topics contain the same groups of words. Then, the LDA is looking for the group of the word that frequently mentioned together in a comment across the dataset. In this case, we can decide the number of clusters, and then the model can return the probability of the particular comment. For instance, the first comment has the highest chance to belong to the second cluster. After identifying the clusters, we can return the most frequent word in each cluster.

We used this technique to give businesses the ability to cluster their negative comments and see the problem mostly belongs to which section. This technique will work really well if the companies provide different services or products.

In this example, we used Nike reviews to apply clustering. We used the CountVectorizer to tokenize the data and remove the stop words that appear in 90% of comments. Then we used the LDA to fit our data. The n_components is equal to 2. It means we want two clusters, and then we returned the 15 most frequent words in each cluster.

Cluster1

```
The top 15 words for topic #1447
['received', 'email', 'ordered',
'time', 'just', 'said', 'refund',
'days', 'told', 'pair', 'customer',
'service', 'shoes', 'order', 'nike']
```

Cluster2

```
The top 15 words for topic #1447
['people', 'bought', 'flag',
'quality', 'product', 'customer',
'time', 'like', 'shoes', 'dont',
'products', 'just', 'buy',
'company', 'nike']
```

By clustering the negative comments, we can easily see the problems belong to two sections. The first cluster is mostly about

the order time and refunding. So most issues in this category are because of the problem in refunding and ordering time and maybe the customer service. However, the second one users mostly argued about the quality of the product. So, instead of going through the more 2500 negative comments, Nike can easily understand the source of the problem.

4-DISCUSSION/ANALYSIS/FINDINGS

A. N-gram

Our findings show that removing punctuation and stop words are essential to finding meaningful words for the user. Besides, changing the uppercase words to lowercase can show us a better result.

B. Named Entity Recognition

Name Entity Recognition can be helpful for companies to review the user's comments more efficient. However, sometimes, it cannot identify the specific name or part of the text, but it has an ability that a user can add new words on it and make it better.

C. Sentiment Analysis

Sentiment analysis is vital for the companies; however, as we mentioned before, it has its drawbacks. The companies cannot only rely on sentiment analysis based on its drawbacks, such as not detecting sarcasm, wrong spelling, and so on, but it can be a great tool to measure the overall feeling of the customers. As we saw earlier in sentiment analysis, lots of 1-star comments for Lufthansa shows that people express their negative feeling strongly about the Lufthansa airline. However, it does not mean if there are lots of bad 1-star comments; the negative feeling is intense. To demonstrate an example, we applied the same process for Nike because they also had lots of 1-star comments, but the result was different. We can see the less intense feeling in compression with the Lufthansa. Maybe a delay in receiving a Nike shoe is not pleasant for a user, but missing a flight to another country because of a mistake can make the user easily angry.

D. Clustering using LDA

However, the clustering can be useful for the business, but not all the time is informative. The other problem is the number of the cluster is tricky, and only people who know about the business can do and interpret that. If the business is providing a particular product and service, mostly the clusters will be similar to each other.

5-CONCLUSION

Review websites become one of the most important tools for users and businesses. Today people check on the reviews before making any purchasing decision. This review website allows the users to write a review regarding their experience on certain services or products. While a business can experience the feelings of their users and their needs through the review. An efficient review system can save a lot of time and provide the best experience to its users. The solutions which are suggested in this paper can improve the Trustpilot system and works well in providing the best experiences for the users.

Reference

[1]- Ravir, K," A survey on opinion mining and sentiment analysis: Tasks, approaches and applications", Institute for Development & Research in Banking Technology, June 2015hkjhkh

[2]-Hu, G," Sentiment Analysis of Surveys using both Numeric Ratings and Text Comments", Central Michigan University, January 2009

[3]. Paknejad, S," Sentiment classification on Amazon reviews using machine learning approaches", KTH University, Sweden, june2018

[4]. https://www.trustpilot.com/

[5]. https://github.com/explosion/spaCy

[6].https://en.wikipedia.org/wiki/Natural_Language_Toolkit

[7] https://www.nltk.org/

[8] http://www.albertauyeung.com/post/generating-ngrams-python/

[9] https://catalog.ldc.upenn.edu/LDC2013T19

[10] https://spacy.io/api/annotation#named-entities

[11] https://github.com/cjhutto/vaderSentiment

[12]https://github.com/cjhutto/vaderSentiment/blob/master/vaderSentiment/vader_lexicon.txt