



UNIVERSITÀ DI PISA

Business and Project Management

Analysis of Online Reviews

Project developed by

Cantini Irene,

De Filomeno Elisa

INDEX

1. Introduction	3
2. Dataset	4
2.1 Dataset description	4
2.2 Dataset cleaning	5
3. First Question	7
3.1 Training set building	7
3.2 Text analysis using python	8
3.3 Code	10
3.4 Conclusion	10
4 Second Question	11
4.1 Samsung	12
4.2 BLU	14
4.3 Apple	16
4.4 LG	18
4.5 Nokia	20
4.6 Conclusion	21

Introduction

The project discussed in this paper deals with the topic of online reviews that customers write about products on the internet. The goal is to analyze customer sentiments related to the reviews to give additional tools to companies using Natural Language Procession (NLP) techniques.

The project mainly focuses on understanding two aspects, review helpfulness and customer satisfaction, thanks to text mining.

In the following sections, the paper is divided into two principal parts to answer two business-relevant questions:

1) Is Sentiment analysis useful to classify Online reviews and verify customer satisfaction with different products?

As specified by the question text, the goal of this first phase, is to use the Sentiment Analysis technique on Online reviews to provide instant feedback about the customers' opinion on purchased products. Thanks to this process companies can obtain a first approach to customer satisfaction or dissatisfaction with their products using text mining classification strategies.

2) After the development of sentiment analysis, is it possible to identify strengths and weaknesses of brands using keyword extraction in order to make better future business decisions?

The goal of this second phase is to explore a different NLP technique called Keyword Extraction to make the companies able to approach more deeply with online reviews. Thanks to this process companies can extract from reviews the most important words to understand better the product features more or less appreciated. With this analysis the companies might take in future decision strategies to accommodate customers' needs.

All the codes and files used to develop this project are available at the following GitHub link: [BPM Project](#).

Dataset

Dataset description

The dataset used in this paper, to answer both questions and develop the analysis, could be downloaded by *data.world* an internet site at this link: [Dataset](#).

This dataset contains more than 400 thousand reviews collected by amazon about some of the most important mobile brand and their products.

The following fields are available in this dataset:

- Product Title
- Brand
- Price
- Rating (from 1 to 5)
- Review text
- Number of people who found the review helpful

The following picture shows a sample of dataset rows:




	 product_name ▼	 brand_name ▼	# price ▼	# rating ▼	 reviews ▼	# review_votes ▼
106	"Nokia Asha 302 Unlocked	Nokia	299	3	a pretty capable, c	2
107	"Nokia Asha 302 Unlocked	Nokia	299	5	Excellent phone ha	0
108	"Nokia Asha 302 Unlocked	Nokia	299	1	Do not buy it!. It	0
109	"Nokia Asha 302 Unlocked	Nokia	299	5	Excellent phone, fo	1

Figure 1: Dataset sample

Dataset cleaning

The dataset downloaded from the internet was found to be very "dirty" and therefore subjected to a thorough cleaning procedure to be able to make a more accurate analysis of the text left in the comments.

Four scripts are realized to clean the dataset and better organize the work. Each script has its own task.

- 0_dataset_cleaning.py → this script is used to remove all rows having at least one null field end and drop the price column useless for the paper purpose.

```
# remove rows with at least one null field
df = df.dropna(how='any')

# remove price column
df = df.drop(columns=['Price'])
```

- 1_comments_cleaning → this script is used to clean the text removing all the useless characters like links, emojis and multiple spaces.

```
emoji_pattern = re.compile("[
    u\"\\U0001F600-\\U0001F64F\" # emoticons
    u\"\\U0001F300-\\U0001F5FF\" # symbols & pictographs
    u\"\\U0001F680-\\U0001F6FF\" # transport & map symbols
    u\"\\U0001F1E0-\\U0001F1FF\" # flags (iOS)
    u\"\\U00002500-\\U00002BEF\" # chinese char
    u\"\\U00002702-\\U000027B0\"
    u\"\\U00002702-\\U000027B0\"
    u\"\\U000024C2-\\U0001F251\"
    u\"\\U0001f926-\\U0001f937\"
    u\"\\U00010000-\\U0010ffff\"
    u\"\\u2640-\\u2642\"
    u\"\\u2600-\\u2B55\"
    u\"\\u200d\"
    u\"\\u23cf\"
    u\"\\u23e9\"
    u\"\\u231a\"
    u\"\\ufe0f\" # dingbats
    u\"\\u3030\"
    \"]+", re.UNICODE)

def modifyStr(x):
    x = re.sub('http[^\s]+', '', x)
    x = emoji_pattern.sub(r'', x)
    x = x.replace('\n', '')
    x = re.sub(' +', ' ', x)
    return x

df['Reviews'] = df['Reviews'].apply(modifyStr)
```

- 2_language_detection.py → this script is used to delete all the non-English comments.

```
def detect_language(x):
    try:
        return detect(x)
    except:
        return np.nan

print(df['Reviews'].shape)

df['language']=df.Reviews.apply(detect_language)
df=df[df.language.eq('en')]
dfdf.drop(columns=['language'])
```

- 3_contraction → this script is used to expand the normal English contraction like *don't* is converted into *do not* and then we left only the letters by eliminating all other characters. The last step is to make the text lowercase because if the text is in the same case, it is easy for a machine to interpret the words because the lower case and upper case are treated differently.

```
contractions_dict = {
    "ain't": "am not / are not / is not / has not / have not",
    "aren't": "are not / am not",
    "can't": "cannot",
    "can't've": "cannot have",
    "'cause": "because",
    "could've": "could have",
    "couldn't": "could not",
    "couldn't've": "could not have",
    "didn't": "did not",
    "doesn't": "does not",
    "don't": "do not",... }

# Regular expression for finding contractions
contractions_re=re.compile('%s' % '|'.join(contractions_dict.keys()))

def expand_contractions(text,contractions_dict=contractions_dict):
    def replace(match):
        return contractions_dict[match.group(0)]
    return contractions_re.sub(replace, text)

# Expanding Contractions in the reviews
df['Reviews'] = df['Reviews'].apply(lambda x:expand_contractions(str(x)))
df['Reviews'] = df.Reviews.str.replace('[^a-zA-Z ]', '')
df['Reviews'] = df['Reviews'].apply(lambda x: x.lower())
```

First Question

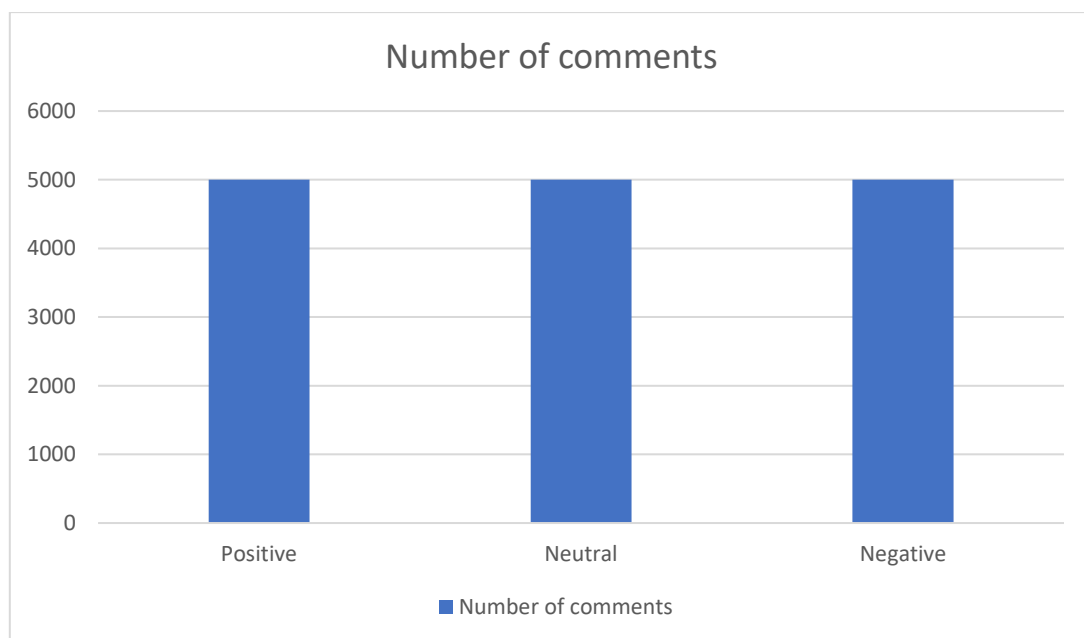
Training set building

In this section the first question description process is reported. The object is to build a classifier able to classify online reviews with the predicted related sentiment. To do that a training set is extracted from the entire initial dataset. The number of comments is very high and for this reason we decided to keep just a representation of them.

For the training set we built a dataset composed of 15000 comments labeled with three different classes using a ground truth:

- Positive comments: their rating is 5.
- Neutral comments: their rating is 3.
- Negative comments: their rating is 1.

We decided to have a balanced training set composed of 5000 comments from each class. The distribution of the training set is shown below:



Text analysis using python

We performed four pre-processing steps to transform our comments into BOW (bag of words) representation in Python. To carry out them, we used the CountVectorizer function from sklearn.

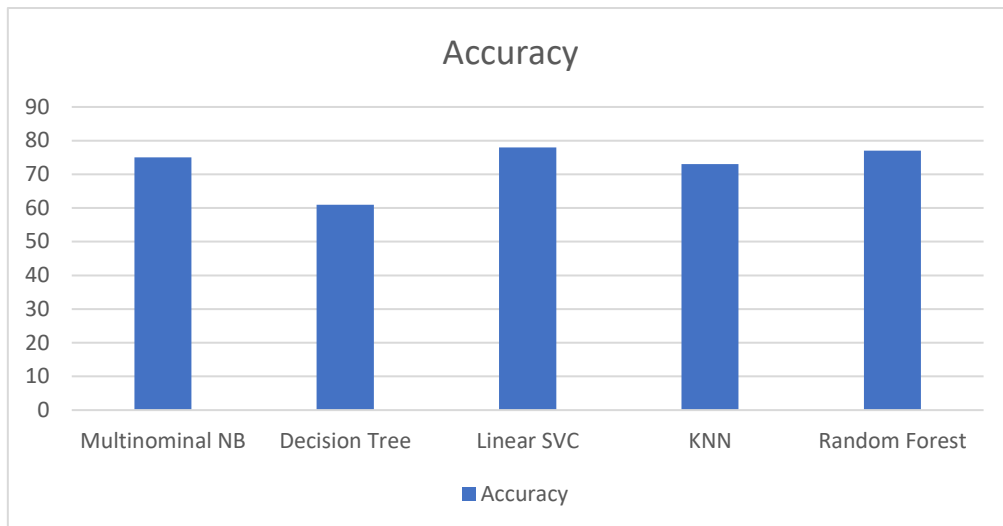
- **Tokenization:** this step transform a stream of characters into a stream of processing units called tokens. In this way each text is represented as a set of words. To do this we set some parameters of the CountVectorizer function. We set *strip_accents='ascii'* which removes accents and performs other character normalization during the preprocessing step working on characters that have a direct ASCII mapping. the *token_pattern* parameter is set to *r"(?u)\b[^\d\W][^\d\W]+\b"* to define what constitutes a "token". Using this regular expression a token is composed of at least two characters.
- **Stop Words Filtering:** with this step the stop words are removed (which provide little or no useful information to the text analysis).
- **Lemmatization:** This step is used to reduce inflectional and derivationally related forms to a common base because the text contains different forms of a word or derivationally related words with similar meanings.
- **Token Filtering:** this step is used to reduce the number of tokens maintaining only the most relevant ones. Setting *max_features=3000*, as CountVectorizer's parameter, only the top *max_features* ordered by term frequency across the corpus are considered.

The tokenization, stop word process and lemmatization are computed all together thanks to the function 'LemmaTokenizer' passed as a parameter to the CountVectorizer.

At the end of the pre-processing steps other two phases are processed:

- a supervised learning stage to assign at each feature of each text the Tf-IDF value using the TfidfTransformer function.
- some tests with different classifiers and for each one we made a "5 fold Cross-Validation".

These steps are assigned to a pipeline with five different possible classifiers to select the best one. The obtained results are shown below:



After observing these results a paired t-test is performed to compare the best two classifier: Random Forest and Linear SVC. We assumed a significance threshold of $\alpha=0.05$ for rejecting the null hypothesis that both algorithms perform equally well on the dataset and conducted the 5-fold cross-validated t-test. To make this test we used a python script, the code is shown below:

```
1. t, p = paired_ttest_kfold_cv(estimator1=text_clf3,
2.                             estimator2=text_clf5,
3.                             X=X, y=y, cv=5, random_seed=1)
4.
5. print('p-value: %.3f' % p)
```

The *p-value* is equal to 0.003. Since $p < \alpha$ the null hypothesis is rejected and for this reason the Linear SVC is selected as the best solution because it is the model with a lower error rate.

The table below shows the confusion matrix of the Linear SVC classifier:

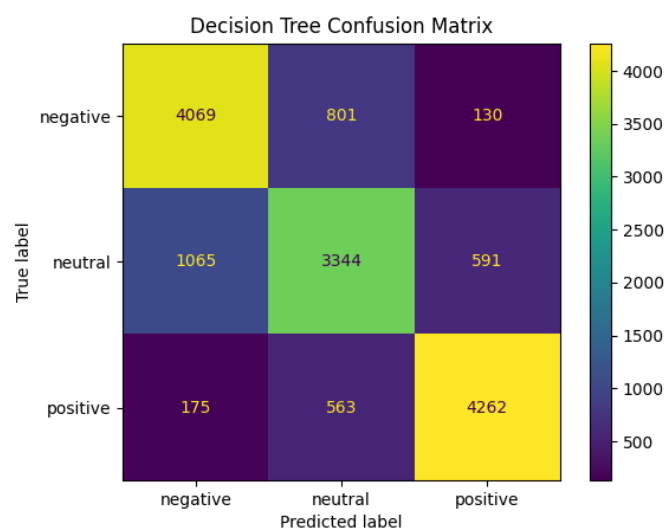


Figure 2- Confusion matrix Linear SVC

Code to build the best classifier

```
class LemmaTokenizer:
    def __init__(self):
        self.wnl = WordNetLemmatizer()
    def __call__(self, doc):
        from nltk.corpus import stopwords
        stopwords = stopwords.words('english')
        return [self.wnl.lemmatize(t) for t in word_tokenize(doc) if t is not stopwords]

text_clf = Pipeline([
    ('vect', CountVectorizer(tokenizer=LemmaTokenizer(), strip_accents='ascii',
max_features=3000)),
    ('tfidf', TfidfTransformer()),
    ('clf', svm.LinearSVC(C=0.1)),
])

text_clf.fit(x_train,y_train)

#calculating accuracy in cross-validation
scores = cross_val_score(text_clf3, X, y, cv=5)
print("Accuracy SVM : %0.2f (+/- %0.2f)" % (scores3.mean(), scores3.std() * 2))
# prediction in cross validation
predicted = cross_val_predict(text_clf3, X, y, cv=5)

print(metrics.classification_report(y, predicted,
                                   target_names=target_names)) # metrics extraction
print(metrics.confusion_matrix(y, predicted))
ConfusionMatrixDisplay.from_predictions(y,predicted)
plt.title("Linear SVC Confusion Matrix")
plt.show()
```

Conclusion

The question from which the paper starts was: **“Is Sentiment analysis useful to classify Online reviews and verify customer satisfaction with different products?”**

After performing this analysis there is sufficient evidence that sentiment analysis works well with online reviews and can classify, with good accuracy, the sentiments associated with the reviews.

In conclusion, companies can use this NLP technique to obtain a first impact on their product goodness and customer satisfaction.

Second Question

“After the development of sentiment analysis, is it possible to identify strengths and weaknesses of brands using keyword extraction in order to make better future business decisions?”

In the second part of our study, we analyzed customers’ reviews using the keyword extraction technique to elaborate better future business strategies. For this analysis we used the entire dataset containing reviews, which were already classified as positive, neutral or negative sentiment by a linear svc classifier.

We discovered that studying the positive and negative reviews, separately, by extracting the most frequent word pairs led to more interesting results than analyzing all the reviews both positive and negative together.

We were able to capture which aspects of the brands were the most and less appreciated by consumers.

We focused our study on the 5 brands with the most reviews:

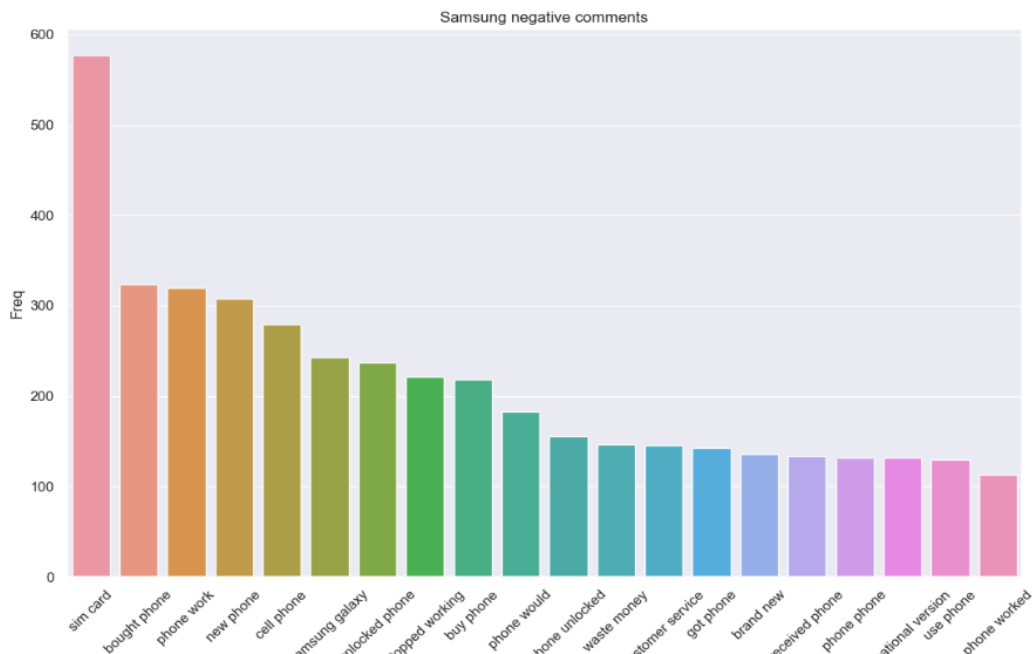
BRAND	NUMBER OF COMMENTS
Samsung	2633
BLU	2570
Apple	2441
LG	929
Nokia	732

The dataset contains only unlocked phones. Unlocked phones aren’t tied to a specific carrier or contract. A locked phone, on the other hand, only works with the network of the carrier that the device was bought from and the phone will block access to other carrier plans, even with a different SIM card. In general, unlocked phones are preferred by customers, their flexibility can bring economic benefits for example when traveling overseas because sim cards from a local carrier are usually less expensive.

Next, we’ll report the results and reflections made by studying the following graphs.

Let’s start with the most commented brand: Samsung

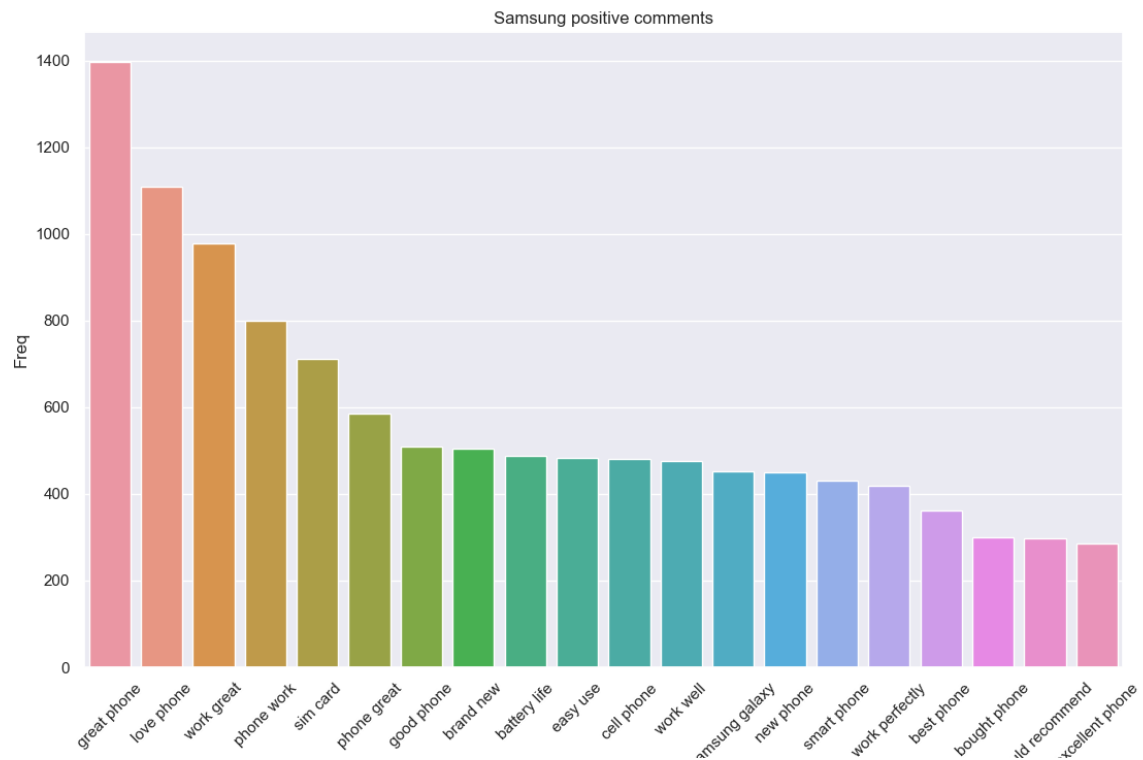
Samsung



Let's analyze the most frequent pair words contained in negative comments. To further understand some of these bi-grams we checked the comments in the dataset containing them.

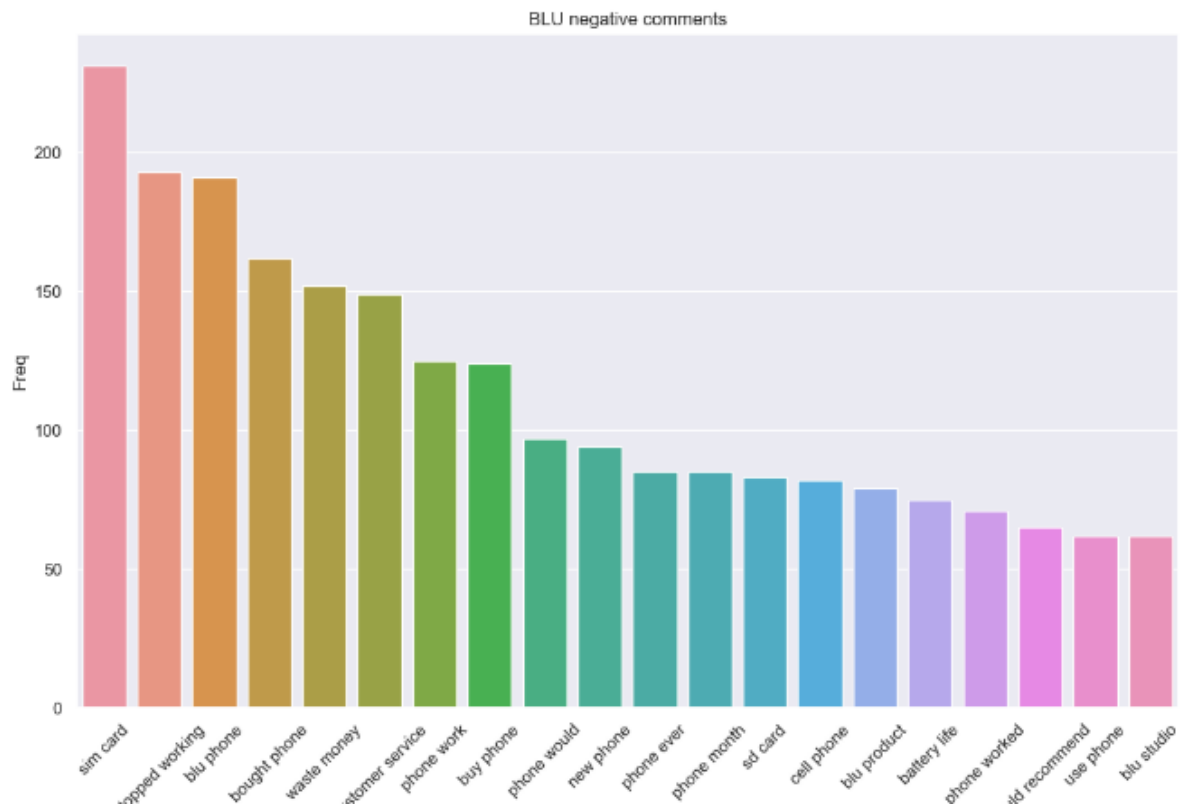
After the previous introduction about unlocked phones, we can have a better understanding of customers' complaints about their new phones and why the main customers' problems concern the 'sim card'. Some customers had their old sim not recognized in the new phone, many of them encountered compatibility issues caused by passing from a GSM phone to a CDMA phone, some had some problems with the dual sim version of their product, often not working correctly or received erroneously the single sim version one. To summarize many customers received the locked version of their product, which is why "unlocked phone" and "phone unlocked" are very frequently used words in negative comments.

Next, we looked at some comments to understand why bi-grams like "phone work", "stopped working", and "phone worked" compared in the bar plot. Consumers have received a broken or malfunctioning product, due to the seller or due to an intrinsic defect of the product. For example they had problems related to the touch screen and camera which stopped working just after a few weeks, which is why they decided to call Amazon or Samsung "customer service". Some customers were satisfied with the service obtained but the relative comment was classified as negative because of the malfunctioned product, others were unsatisfied with both the product and customer service.

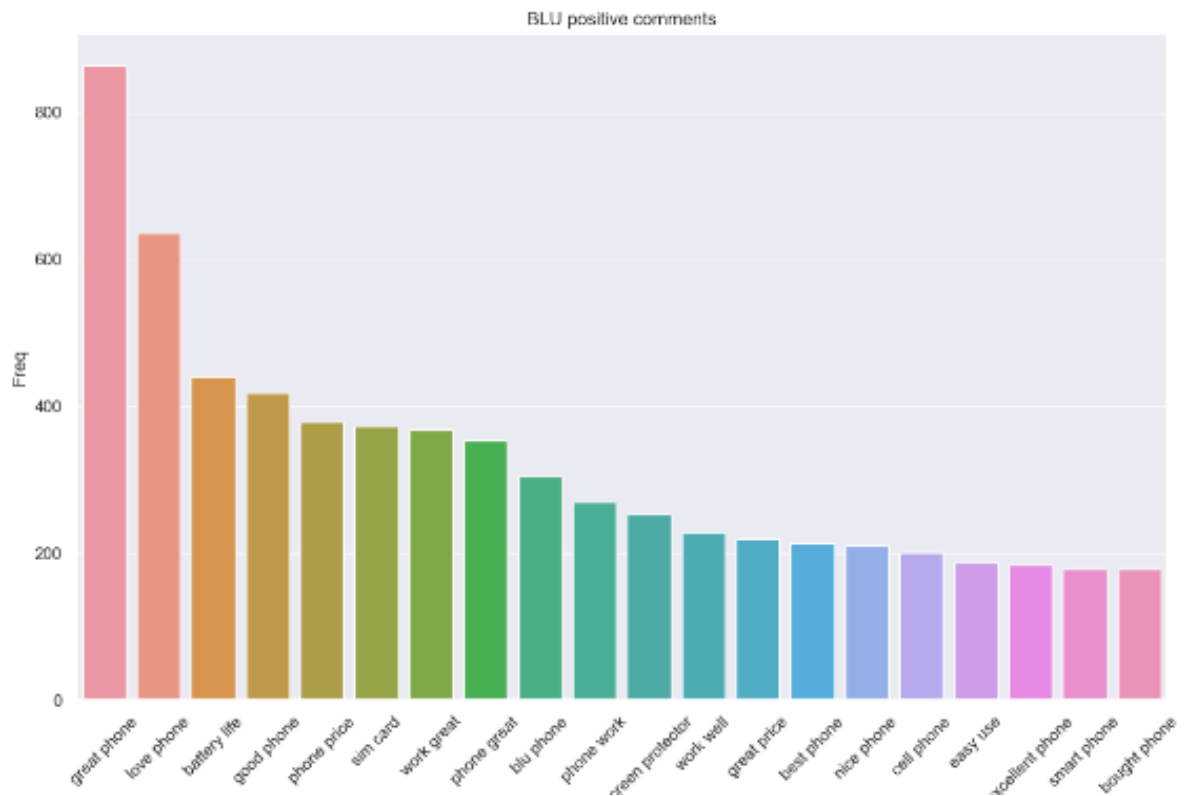


Looking at positive comments we can notice “brand new”, “battery life” and “easy use” as Samsung’s strengths. Customers who bought used devices often were pleased with the very good conditions of their products, many said it was like brand new. We can speculate that, statistically speaking, Samsung’s phones pass pretty well the test of time and many of them “would recommend” them.

BLU



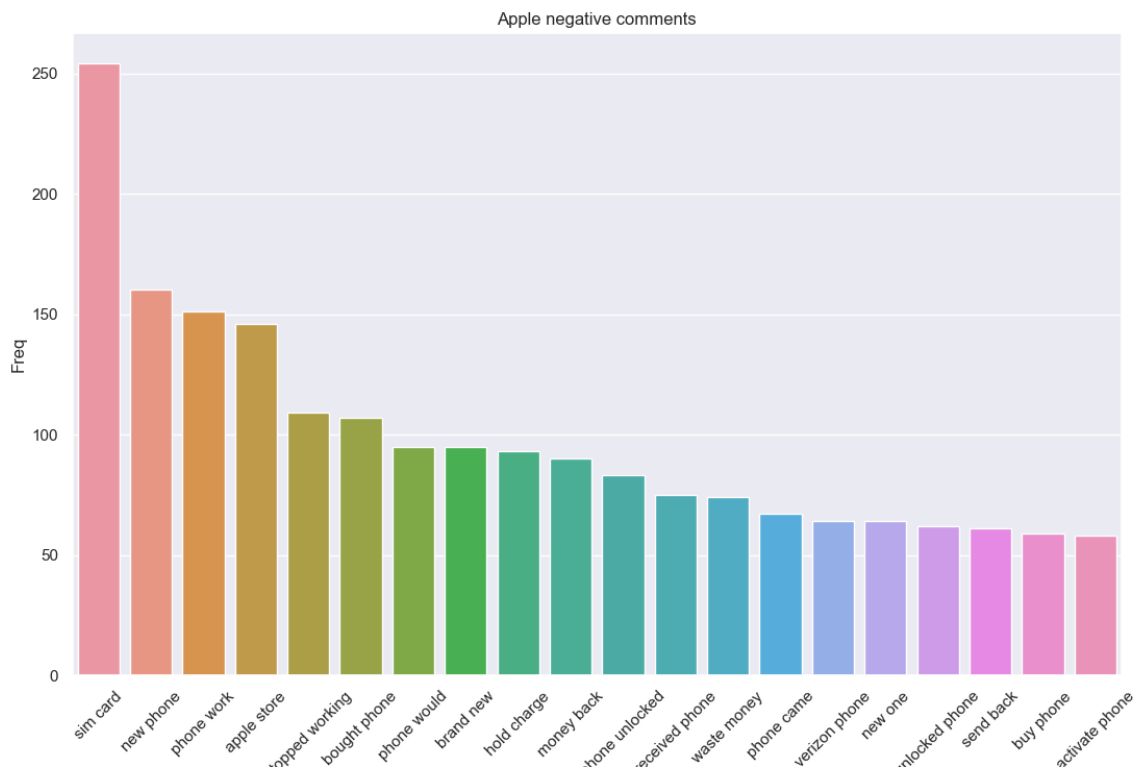
The brand “BLU” is the second most commented brand in our dataset. We can see that there are many similarities with Samsung, “sim card” is still an issue often mentioned. However, we can observe that “customer service” and “battery life” are two of this brand’s weaknesses. Blu customer service is often criticized to have slow response times and being not at all helpful at resolving problems. Battery life dies too quickly when using particular applications and doesn’t last much in general after it was charged a whole night.



This brand's strengths are the product prices ("phone price", "great price") and "easy use". In this case "battery life" seems to contradict previous weaknesses' observations. Since the frequency is a lot higher than in the previous graph, we can assume that all considered the battery life is a positive trait for this brand. Maybe a more in-depth study about the relation between battery life and in-use applications could reveal more.

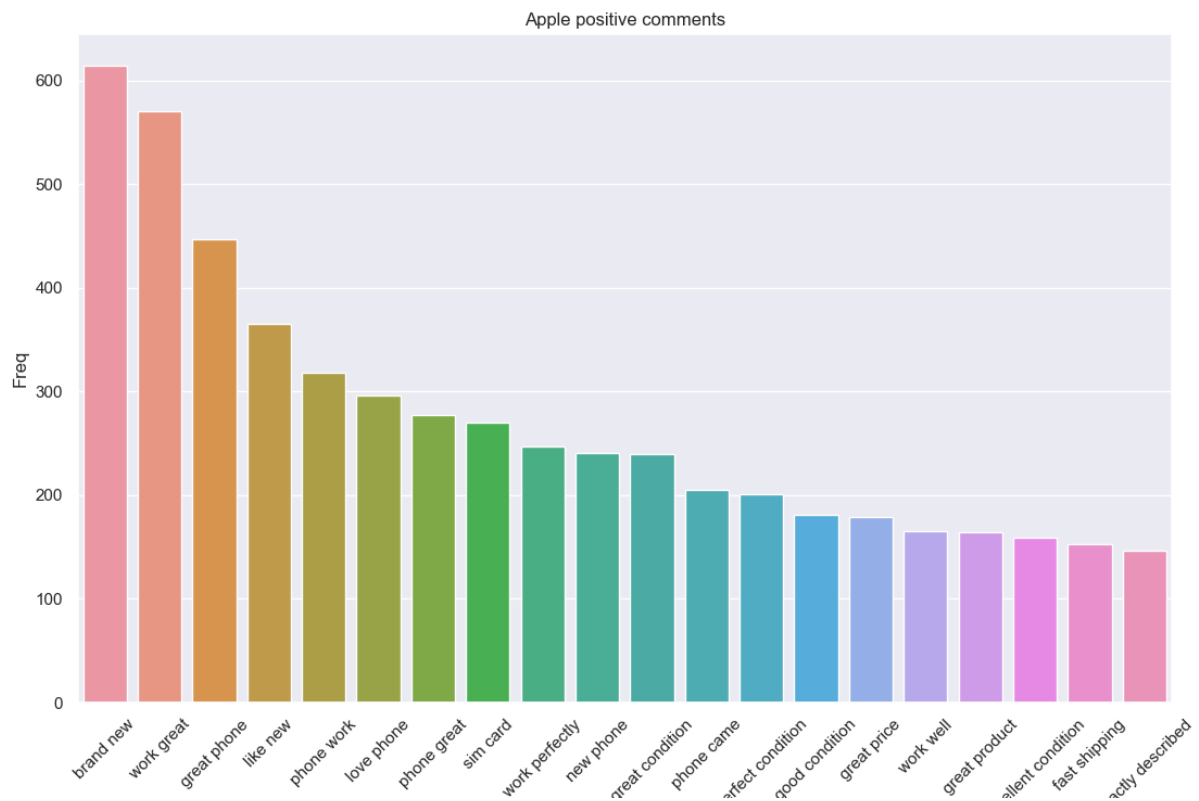
Blu brand is characterized by very low-cost mobile phones, buyers appreciate its good balance of performance for the price.

Apple



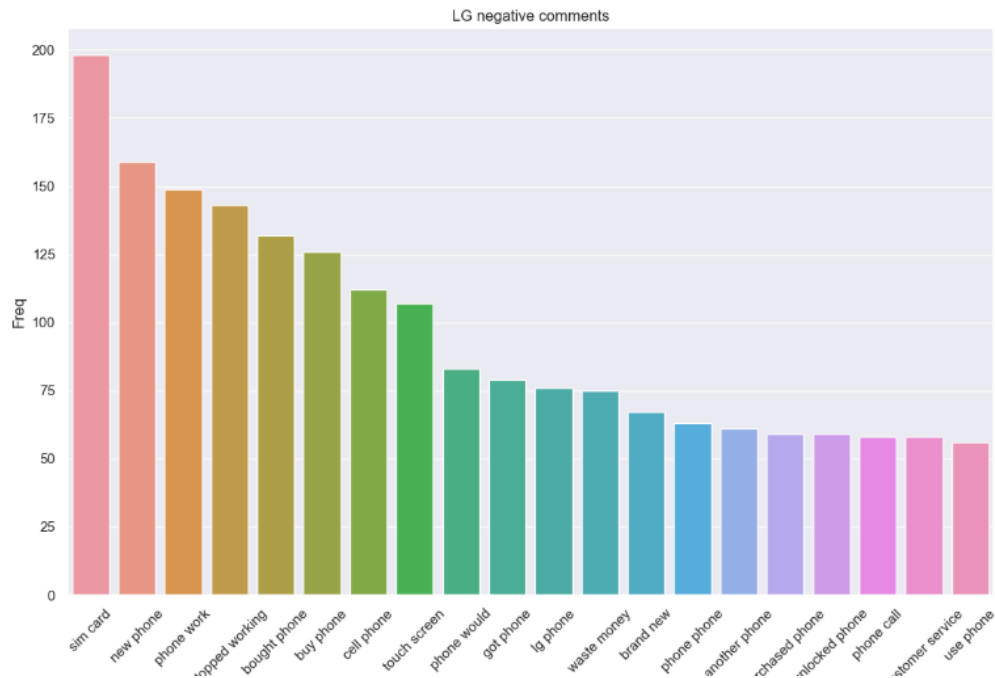
Let's now analyze the most frequent pair words in negative comments about Apple products. Many customers found out that their new phone had a battery problem and didn't "hold charge". This seems to be batteries malfunctioning and not only a short-duration issue.

Looking at comments containing "stopped working" we didn't discover any particular aspect in Apple phones that brought the cell to stop working after some time (sometimes it was a camera problem, some other time the home button, the touch screen and/or the on-off button). To repair them they were brought to an apple store, which is why "apple store" compare in the above plot (it's not a direct critique of apple stores).

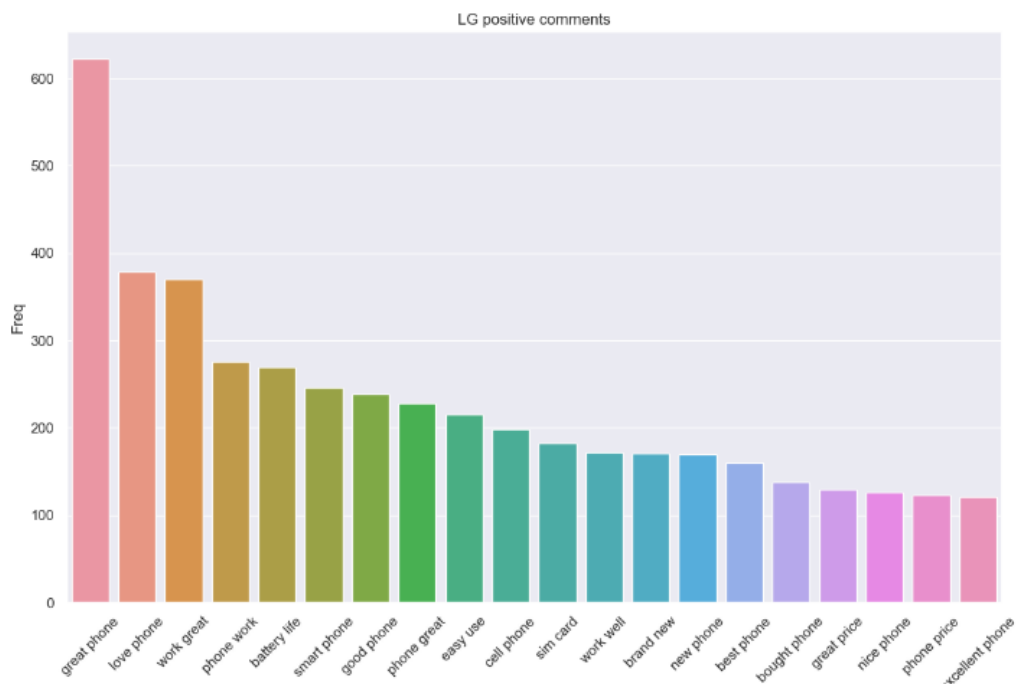


About positive comments we can notice “like new” as one of the most frequent pair words. It’s used by customers who bought a used product and found it in “perfect condition” and “great condition”. In this plot the word “battery” is not present. We can conclude, unlike previous brands, that the battery of their products is Apple’s weakness.

LG

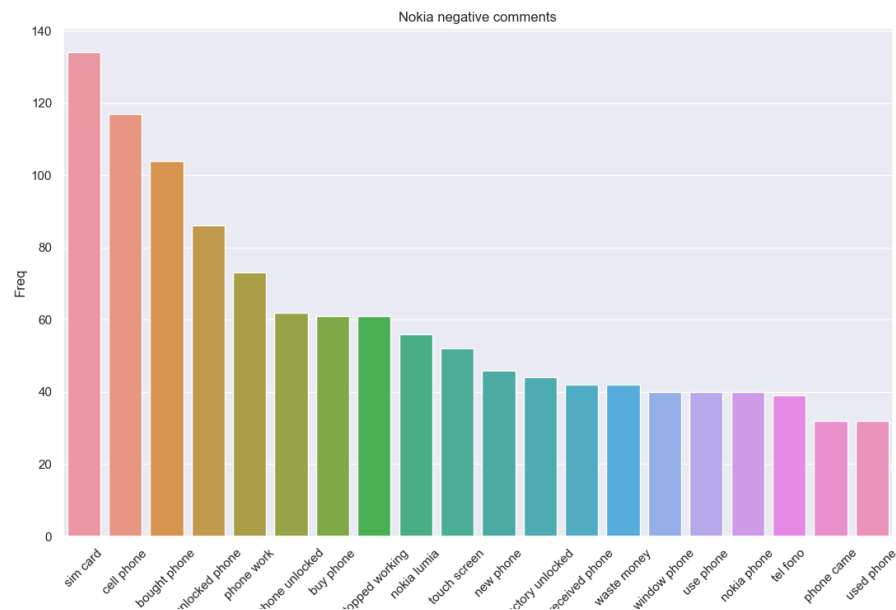


Looking at this bar plot we notice that many customers were disappointed with the “touch screen” of LG phones and the “customer service” (like the brand BLU). Checking the comments containing “stopped working” we confirmed that the most frequent malfunction was the touch screen, just a very few comments showed camera and microphone problems. We should consider in fact that many of these brand phones still have a keyboard and is possible that it was a choice to have touch screens of low quality compared to other brands like Samsung or Apple which already didn’t have a keyboard.

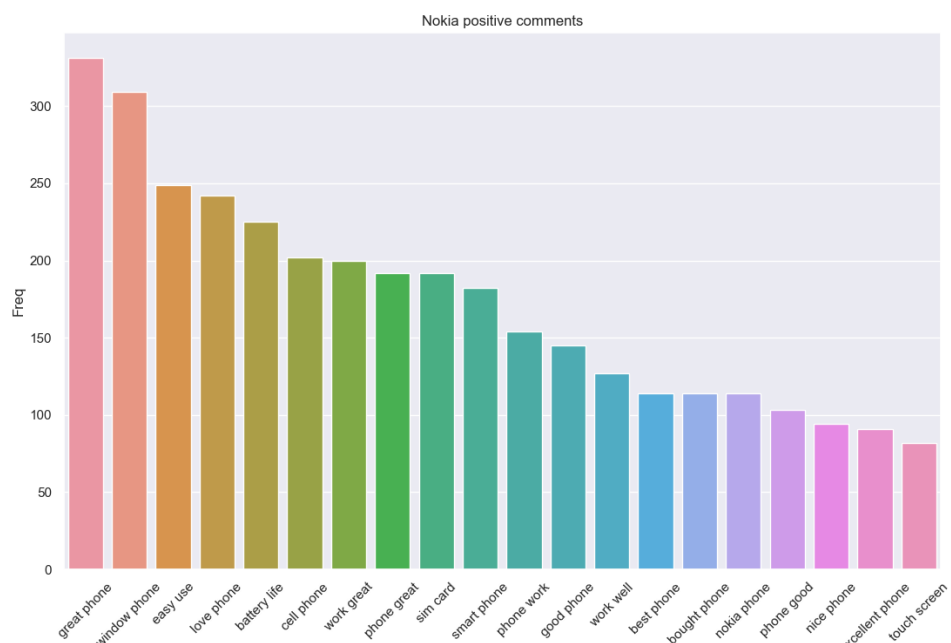


LG's positive characteristics are "battery life", "great price" and "easy use". For example seniors found the keyboard with big numbers very helpful and easy to use. The colors on the screen are brighter than on other phones and this is an appreciated characteristic.

Nokia



The last brand considered in our study is Nokia. Even in this case “touch screen” is a weak point of the brand, in fact like for LG some of these brand phones still have a keyboard.



Nokia’s greatest positive points are “great price”, “easy use” and “battery life”. Nokia has collaborated with Microsoft and produced phones that use Windows operating system and not Android. It seems that this operating system left the customers pleased.

Conclusion

We have seen how it's possible to identify differences between products of different brands using keywords extraction, answering the second question of our study. According to recent research by HubSpot, 49% of consumers selected positive consumer reviews in their top 3 purchase influences.

Analyzing reviews is a critical point for brands to identify customers' needs and frustrations and subsequently improve their market strategies for higher customer satisfaction and firm growth, increasing profits. Companies have an opportunity to earn trust through the recommendations and opinions of others, showing that they are experts in the industry.

In conclusion this study can be used by: customers to make more informed decisions, marketing to make more attractive advertising campaigns, and companies in defining business strategies.