Problem Statement

The proliferation of information is the current trend. Information can be found through text, audio, video, and numbers. It is essential to know how to extract the instructional facts we are interested in from a big amount of data. Customer reviews over online websites like Amazon, Twitter, etc plays a vital role in understanding customer behavior and providing better user experience. However, product reviews can take many forms, which presents difficulties for researchers and data miners. Sentiment Analysis has been a hot topic for research in recent years due to both practical needs and theoretical difficulties.

Example: The customer service was pretty bad.
HUMAN : Interprets in a negative sense.
COMPUTER: confused ? pretty -> positive, bad -> negative

In this project, we have taken the classification of Amazon product reviews particularly for the cell phones industry a step ahead. We classify the reviews into a set of not pre-defined topics using topic moedlling, as well as perform sentiment analysis of the review. Hence given a new review (text) , our model will classify it as good or bad , and identify which topic the review was targeted, thus generating appropriate tags.

Example :

INPUT : “Nokia A655 has a great battery life an efficient charging”
OUTPUT : sentiment: good, topic: battery

Literature Survey

1. A lot of prior research has been done in this field where words and phrases have been classified with prior positive or negative polarity. This prior classification is helpful in many cases but when contextual polarity comes into the picture, the meaning derived from positive or negative polarity can be entirely different. The contextual polarity of the phrases was taken into consideration and ambiguity was removed.[1]

2. Also, a refined method has been devised to establish contextual polarity of phrases by using subjective detection that compresses reviews while still maintaining the intended polarity.[2]

3. Delineated study has been conducted on tweets available on Twitter or on movie reviews to build the grounds on sentiment analysis and opinion mining. A sentiment classifier has been built to categorize positive, negative and neutral sentiments not only in English but also for other languages using corpus from Twitter. The polarity of smartphone product reviews has been found only on the basis of positive and negative orientation of the review [3]

4. A system has been built using a support vector machine where sentiment analysis is carried out by taking into consideration sarcasm, grammatical errors and spam detection. An enhanced Naïve Bayes model by combining methods like effective negation handling, word n-grams and feature selection has been utilized to conduct sentiment analysis. [4]

5. Sentiment analysis on travel reviews using three machine learning models namely, Naïve Bayes, SVM and character based N-gram model has been performed in which SVM and N-gram approaches have better performance than Naïve Bayes. It has been observed that in the maximum number of cases SVM showcases best performance in comparison to other classification models.[5]

6. Sentiment Analysis is not limited just to reviews or Twitter data but is also applicable on stock markets, news articles or political debates. Sentiment analysis can be used to flourish consumer products related business. It uses a rule-based approach for sentiment analysis to extract topic words of negative opinion from sentences and thus promote the competitors of the products receiving negative feedback.[6]

7. Xing Fang and Justin Zhan studied the subjective contents that tackled a fundamental problem of sentiment polarity categorization based on sentence-level and review-level. The experimental results for both levels of categorizations revealed promising outcomes.[7]

We have been able to reveal the hidden patterns in data thanks to data analytics. Many IT departments are unable to keep up with the daily data generation due to its sheer volume and unrelenting speed. E-commerce websites are flooded with a wide range of unique product reviews. Reviews like these can help us understand consumer behavior and make wise judgments. Both structured and unstructured reviews are possible. By removing irrelevant data, useful business insights can be obtained. Big Data has made it possible for firms to grow and improvise based on data rather than gut feeling. It helps with acquiring knowledge about more precisely focused social influencer marketing, customer base segmentation, sales and marketing prospects, fraud detection, risk quantification, and more.

Till now the models that were developed to cater to the needs of the industry, classified the reviews based on a broad level of their sentiments only. But in the fast growing data world, only broad level classification is not sufficient. Rather knowing what actually the customer is speaking about in his review, be it positive or negative is instrumental in policy making. So we have decided to come up with a new model that incorporates certain key performance indicators along with the sentiments to make the analysis more comprehensive. The data set used to train the model is ‘Cell phone and accessories review data” from Amazon, taken from the following source: https://nijianmo.github.io/amazon/index.html

Project Schedule

Project Component	**Status**	**Date**	**Notes**
Problem Statement	Completed	27.10.22 - 28.10.22	Brainstorm the topic for the project.
Literature Review	Completed	29.10.22 - 08.11.22	Go through previous research works done in this field.
Dataset & Text Preprocessing	Completed	09.11.22 - 14.11.22	Select appropriate dataset and apply various standard preprocessing techniques as well as some project specific techniques.
Word Embeddings	Completed	14.11.22 - 16.11.22	Apply different pre-trained word embedding for the input corpus in the training phase
Training Model	Completed	16.11.22 - 20.11.22	Develop and experiment with various sentiment and classification NLP models
Validation & Testing	Completed	21.11.22 - 24.11.22	Fine-tune various hyper parameters using methods such as k-fold cross validation to ensure unbiased training
Results	Completed	25.11.22 - 29.11.22	Compare the observations and findings of the testing phase
Report Writing	Not started	30.11.22 - 05.12.22	Document the report for final project submission

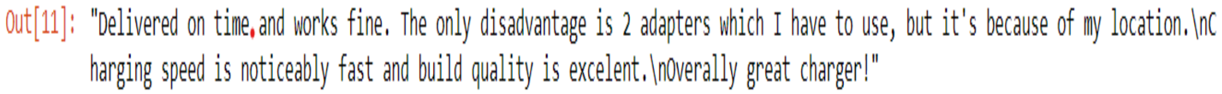
Need of Text Processing

Raw text data might contain unwanted or unimportant text due to which our results might not give efficient accuracy and might make it hard to understand and analyze. So, proper pre-processing must be done on raw data. Thus, the dataset is preprocessed in order to check missing values, noisy data, and other inconsistencies before training the model. Depending on how data gets pre-processed, the results also differ.

The most important steps involved in text pre-processing are:

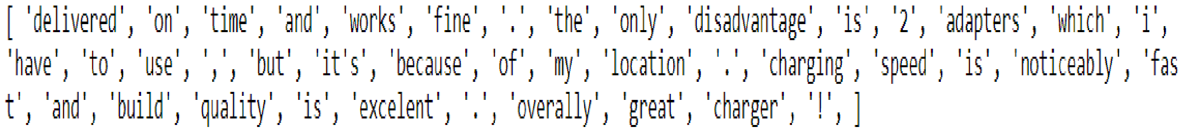
1) Tokenization
2) Removal stop words
3) Stemming
4) Lemmatization

These are explained further in the following section and applied to a test example from the dataset used in the project. The example is given below:



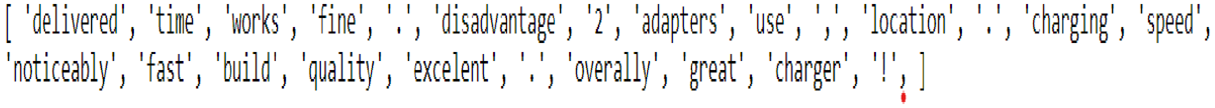
```
Out[11]: "Delivered on time,and works fine. The only disadvantage is 2 adapters which I have to use, but it's because of my location.\nC
         harging speed is noticeably fast and build quality is excelent.\nOverally great charger!"
```

Tokenization:
It is the process of breaking a whole sentence into individuals such as symbols, keywords, and phrases known as a token. In tokenization some characters are removed. Applying the technique of tokenization on the above sentence would yield the following output:



```
[ 'delivered', 'on', 'time', 'and', 'works', 'fine', '.', 'the', 'only', 'disadvantage', 'is', '2', 'adapters', 'which', 'i',
'have', 'to', 'use', ',', 'but', 'it's', 'because', 'of', 'my', 'location', '.', 'charging', 'speed', 'is', 'noticeably', 'fas
t', 'and', 'build', 'quality', 'is', 'excelent', '.', 'overally', 'great', 'charger', '!', ]
```

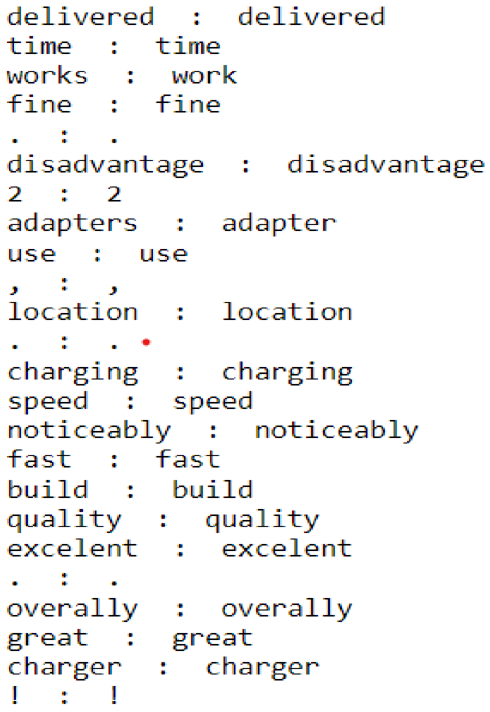
Stop words:
Stop words are those objects in a sentence which are not required in any segment of text mining, so usually, these sentences are removed to increase the efficiency of analysis. The idea is simply removing the words that occur commonly across all the documents in the corpus. Typically, articles and pronouns are generally classified as stop words. These words are removed since they are not discriminative. Stop word removal when applied to the example produces the following output:



```
[ 'delivered', 'time', 'works', 'fine', '.', 'disadvantage', '2', 'adapters', 'use', ',', 'location', '.', 'charging', 'speed',
'noticeably', 'fast', 'build', 'quality', 'excelent', '.', 'overally', 'great', 'charger', '!', ]
```

Stemming
It is the process of alloying the modified forms of a word into a common interpretation. It is a technique used for information retrieval (IR) in text processing based on the statement in documents. In NLP use cases such as sentiment analysis, spam classification, restaurant reviews etc., getting the base word is important to know whether the word is positive or negative. Stemming is used to get that base word.

Lemmatization
In linguistics, it is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form. It is different from stemming as it not only looks at the spelling of words but also the meaning of the word.
The techniques of stemming and lemmatization when applied on the example, output the following:



```
delivered  :  delivered
time  :  time
works  :  work
fine  :  fine
.  :  .
disadvantage  :  disadvantage
2  :  2
adapters  :  adapter
use  :  use
,  :  ,
location  :  location
.  :  .
charging  :  charging
speed  :  speed
noticeably  :  noticeably
fast  :  fast
build  :  build
quality  :  quality
excelent  :  excelent
.  :  .
overally  :  overally
great  :  great
charger  :  charger
!  :  !
```

References:

[1] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” in Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005, pp. 347–354.

[2] B. Pang and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts,” in Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004, p. 271.

[3] M. WAHYUDI and D. A. KRISTIYANTI, “Sentiment analysis of smartphone product review using support vector machine algorithm-based particle swarm optimization.” Journal of The-oretical & Applied Information Technology, vol. 91, no. 1,2016

[4] D. N. Devi, C. K. Kumar, and S. Prasad, “A feature based approach for sentiment analysis by using support vector ma-chine,” in Advanced Computing (IACC), 2016 IEEE 6th Interna-tional Conference on. IEEE, 2016, pp. 3–8.
V. Narayanan, I. Arora, and A. Bhatia, “Fast and accurate sen-timent classification using an enhanced naive bayes model,” in International Conference on Intelligent Data Engineering and Automated Learning. Springer, 2013, pp. 194–201.

[5] S. Tan and J. Zhang, “An empirical study of sentiment analysis for chinese documents,” Expert Systems with applications, vol. 34, no. 4, pp. 2622–2629, 2008.

[6] L.-C. Yu, J.-L. Wu, P.-C. Chang, and H.-S. Chu, “Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news,” Knowledge-Based Systems, vol. 41, pp. 89–97, 2013.

[7]M. Hagenau, M. Liebmann, and D. Neumann, “Automated news reading: Stock price prediction based on financial news using context-capturing features,” Decision Support Systems, vol. 55, no. 3, pp. 685–697, 2013.
T. Xu, Q. Peng, and Y. Cheng, “Identifying the semantic orien-tation of terms using s-hal for sentiment analysis,” Knowledge-Based Systems, vol. 35, pp. 279–289, 2012.
I. Maks and P. Vossen, “A lexicon model for deep sentiment analysis and opinion mining applications,” Decision Support Systems, vol. 53, no. 4, pp. 680–688, 2012.
https://link.springer.com/article/10.1186/s40537-015-0015-2#Sec2

Jagdale, R.S., Shirsat, V.S. and Deshmukh, S.N. (2018) “Sentiment analysis on product reviews using Machine Learning Techniques,” Cognitive Informatics and Soft Computing, pp. 639–647. Available at: https://doi.org/10.1007/978-981-13-0617-4_61.

Sentiment analysis of Amazon product reviews based on NLP (no date) IEEE Xplore. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&amp;arnumber=9513125&amp;tag=1 (Accessed: November 12, 2022).

Fang, X. and Zhan, J. (2015) “Sentiment analysis using product review data,” Journal of Big Data, 2(1). Available at: https://doi.org/10.1186/s40537-015-0015-2.

Real-time sentiment analysis on e-commerce application (no date) IEEE Xplore. Available at: https://ieeexplore.ieee.org/document/8743331 (Accessed: November 12, 2022).

A review of natural language processing techniques for sentiment analysis using pre-trained models (no date) *IEEE Xplore*. Available at: https://ieeexplore.ieee.org/abstract/document/9076502 (Accessed: November 12, 2022).

Sun, S., Luo, C. and Chen, J. (2016) A review of natural language processing techniques for opinion mining systems, Information Fusion. Elsevier. Available at: https://www.sciencedirect.com/science/article/pii/S1566253516301117 (Accessed: November 12, 2022).

Dataset Citation:
Justifying recommendations using distantly-labeled reviews and fined-grained aspects
Jianmo Ni, Jiacheng Li, Julian McAuley
Empirical Methods in Natural Language Processing (EMNLP), 2019