**Seminar Report**

**ON**

**Sentiment Analysis: IPHONE Review**

**Name – Tarun**

**Roll-no - 2K15/CO/134**

**Aim**

To develop a program which can be used to find out whether given IPHONE review is positive or negative using 'Naive Bayes model' and other models

**INTRODUCTION**

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken. NLP is a component of artificial intelligence (AI).The development of NLP applications is challenging because computers traditionally require humans to "speak" to them in a programming language that is precise, unambiguous and highly structured, or through a limited number of clearly enunciated voice commands. Human speech, however, is not always precise -- it is often ambiguous and the linguistic structure can depend on many complex variables, including slang, regional dialects and social context.

**Uses of natural language processing**

Most of the research being done on natural language processing revolves around search, especially enterprise search. This involves allowing users to query data sets in the form of a question that they might pose to another person. The machine interprets the important elements of the human language sentence, such as those that might correspond to specific features in a data set, and returns an answer.

NLP can be used to interpret free text and make it analyzable. There is a tremendous amount of information stored in free text files, like patients' medical records, for example. Prior to deep learning-based NLP models, this information was inaccessible to computer-assisted analysis and could not be analyzed in any kind of systematic way. But NLP allows analysts to sift through massive troves of free text to find relevant information in the files.

[Sentiment analysis](https://searchbusinessanalytics.techtarget.com/definition/opinion-mining-sentiment-mining) is another primary use case for NLP. Using sentiment analysis, data scientists can assess comments on social media to see how their business's brand is performing, for example, or review notes from customer service teams to identify areas where people want the business to perform better.

Google and other search engines base their machine translation technology on NLP deep learning models. This allows algorithms to read text on a webpage, interpret its meaning and translate it to another language.

Sentiment Analyses

**Sentiment analysis – otherwise known as opinion mining – is a much bandied about but often misunderstood term.**

In essence, it is the process of determining the emotional tone behind a series of words, used to gain an understanding of the the attitudes, opinions and emotions expressed within an online mention.

## Sentiment analysis uses

Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. Social media monitoring tools like [Brandwatch Analytics](https://www.brandwatch.com/brandwatch-analytics/" \t "_blank) make that process quicker and easier than ever before, thanks to real-time monitoring capabilities.

The applications of sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organisations across the world.

Shifts in sentiment on social media have been shown to correlate with shifts in the stock market.

The Obama administration used sentiment analysis to gauge public opinion to policy announcements and campaign messages ahead of 2012 presidential election. Being able to quickly see the sentiment behind everything from forum posts to news articles means being better able to strategise and plan for the future.

It can also be an essential part of your market research and customer service approach. Not only can you see what people think of your own products or services, you can see what they think about your competitors too.  The overall customer experience of your users can be revealed quickly with sentiment analysis, but it can get far more granular too.

The ability to quickly understand consumer attitudes and react accordingly is something that Expedia Canada took advantage of when they noticed that there was a steady increase in negative feedback to the music used in one of their television adverts.

Sentiment analysis conducted by the brand revealed that the music played on the commercial had become incredibly irritating after multiple airings, and consumers were flocking to social media to vent their frustrations.

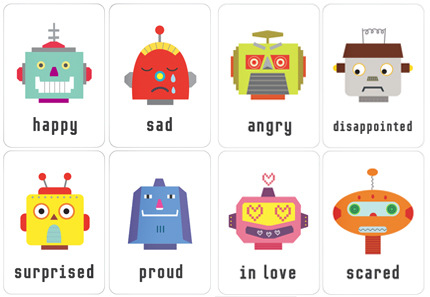
A couple of weeks after the advert first aired, over half of online conversation about the campaign was negative.

Rather than chalking up the advert as a failure, Expedia was able to address the negative sentiment in a playful and self-knowing way by airing a new version of the advert which featured the offending violin being smashed.

## Contextual understanding and tone

But that is not to say that sentiment analysis is a perfect science at all.

The human language is complex. Teaching a machine to analyse the various grammatical nuances, cultural variations, slang and misspellings that occur in online mentions is a difficult process. Teaching a machine to understand how context can affect tone is even more difficult.



Humans are fairly intuitive when it comes to interpreting the tone of a piece of writing.

Consider the following sentence: “My flight’s been delayed. Brilliant!”

Most humans would be able to quickly interpret that the person was being sarcastic. We know that for most people having a delayed flight is not a good experience (unless there’s a free bar as recompense involved). By applying this contextual understanding to the sentence, we can easily identify the sentiment as negative.

Without contextual understanding, a machine looking at the sentence above might see the word “brilliant” and categorise it as positive.

### How natural language processing works

### Current approaches to NLP are based on deep learning, a type of AI that examines and uses patterns in data to improve a program's understanding. Deep learning models require massive amounts of labelled data to train on and identify relevant correlations, and assembling this kind of [big data](https://searchdatamanagement.techtarget.com/definition/big-data) set is one of the main hurdles to NLP currently.

Earlier approaches to NLP involved a more rules-based approach, where simpler machine learning [algorithms](https://whatis.techtarget.com/definition/algorithm) were told what words and phrases to look for in text and given specific responses when those phrases appeared. But deep learning is a more flexible, intuitive approach in which algorithms learn to identify speakers' intent from many examples, almost like how a child would learn human language.

### Importance of NLP

The advantage of natural language processing can be seen when considering the following two statements: "Cloud computing insurance should be part of every service level agreement" and "A good SLA ensures an easier night's sleep -- even in the cloud." If you use national language processing for search, the program will recognize that *cloud computing* is an entity, that *cloud* is an abbreviated form of cloud computing and that *SLA* is an industry acronym for service level agreement.

These are the types of vague elements that appear frequently in human language and that machine learning algorithms have historically been bad at interpreting. Now, with improvements in deep learning and artificial intelligence, algorithms can effectively interpret them.

This has implications for the types of data that can be analyzed. More and more information is being created online every day, and a lot of it is natural human language. Until recently, businesses have been unable to analyze this data. But advances in NLP make it possible to analyze and learn from a greater range of data sources.

### ****What is scikit-learn?****

Scikit-learn is probably the most useful library for machine learning in Python. It is on NumPy, SciPy and matplotlib, this library contains a lot of effiecient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

Please note that scikit-learn is used to build models. It should not be used for reading the data, manipulating and summarizing it. There are better libraries for that (e.g. NumPy, Pandas etc.)

**Components of scikit-learn:**

Scikit-learn comes loaded with a lot of features. Here are a few of them to help you understand the spread:

**Preprocessing The Data**

#### Standardization

>>> from sklearn.preprocessing import StandardScaler

>>> scaler = StandardScaler().fit(X\_train)

>>> standardized\_X = scaler.transform(X\_train)

>>> standardized\_X\_test = scaler.transform(X\_test)

#### Normalization

>>> from sklearn.preprocessing import Normalizer

>>> scaler = Normalizer().fit(X\_train)

>>> normalized\_X = scaler.transform(X\_train)

>>> normalized\_X\_test = scaler.transform(X\_test)

#### Binarization

>>> from sklearn.preprocessing import Binarizer

>>> binarizer = Binarizer(threshold=0.0).fit(X)

>>> binary\_X = binarizer.transform(X)

#### Encoding Categorical Features

>>> from sklearn.preprocessing import LabelEncoder

>>> enc = LabelEncoder()

>>> y = enc.fit\_transform(y)

#### Imputing Missing Values

>>>from sklearn.preprocessing import Imputer

>>>imp = Imputer(missing\_values=0, strategy='mean', axis=0)

>>>imp.fit\_transform(X\_train)

#### Generating Polynomial Features

>>> from sklearn.preprocessing import PolynomialFeatures

>>> poly = PolynomialFeatures(5)

>>> oly.fit\_transform(X)

### Training And Test Data

>>> from sklearn.model\_selection import train\_test\_split

>>> X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y,random\_state=0)

### Create Your Model

### Supervised Learning Estimators

**Linear Regression**

>>> from sklearn.linear\_model import LinearRegression

>>> lr = LinearRegression(normalize=True)

**Support Vector Machines (SVM)**

>>> from sklearn.svm import SVC

>>> svc = SVC(kernel='linear')

**Naive Bayes**

>>> from sklearn.naive\_bayes import GaussianNB

>>> gnb = GaussianNB()

**KNN**

>>> from sklearn import neighbors

>>> knn = neighbors.KNeighborsClassifier(n\_neighbors=5)

### Unsupervised Learning Estimators

**Principal Component Analysis (PCA)**

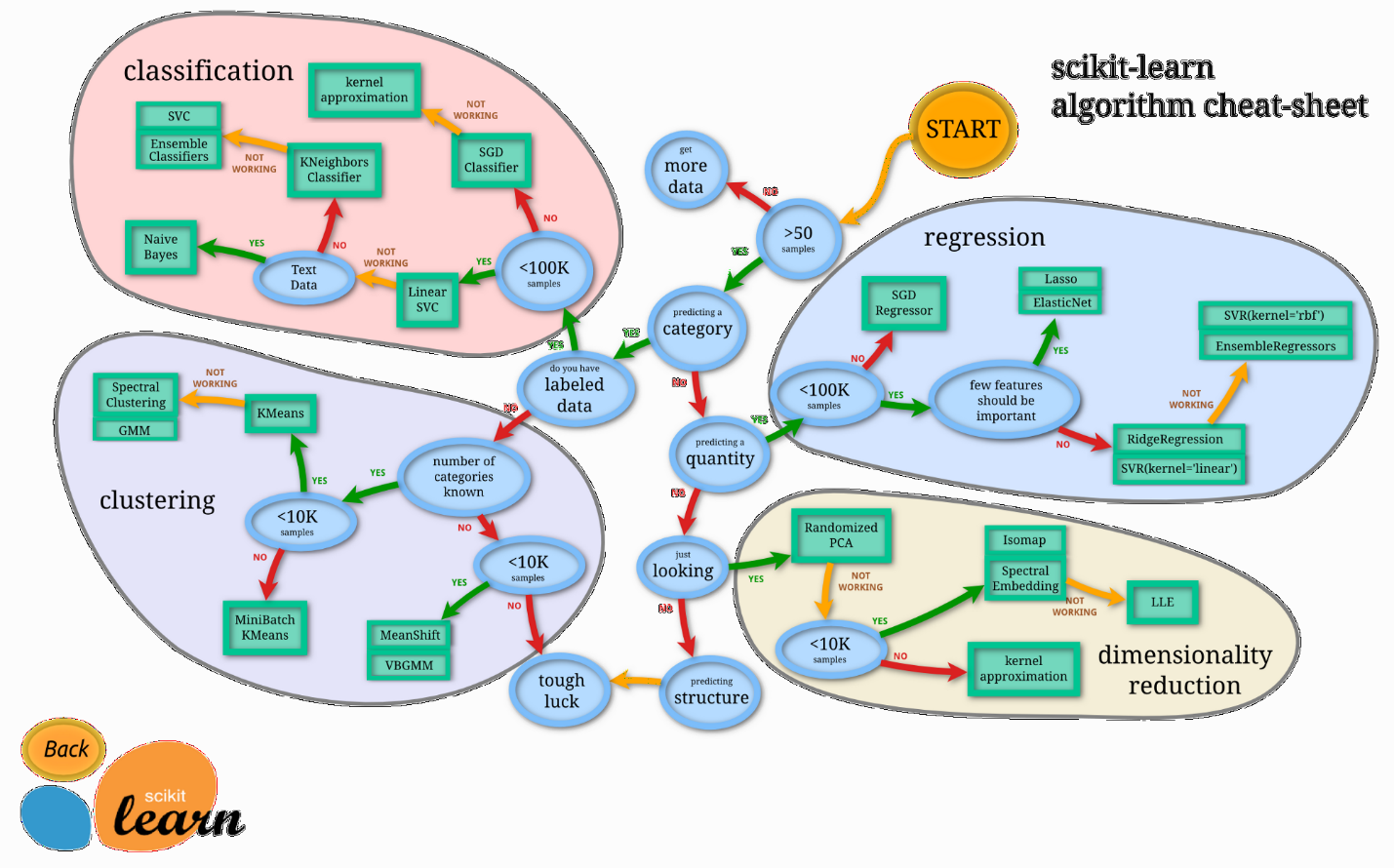
>>> from sklearn.decomposition import PCA

>>> pca = PCA(n\_components=0.95)

**K Means**

>>> from sklearn.cluster import KMeans

>>> k\_means = KMeans(n\_clusters=3, random\_state=0)



**DATA**

Dataset consist of 2 columns one of text and other tells whether the text is positive or negative.Text consists of features of iPhone like battery, camera, price etc.

**High Level Steps**

The high-level sequence involved in processing is as follows:

1) Raw data collection

2) Sentiment labeling

3) Transform into train/test sets for classifier

4) Bag of Words

5) Transform train/test sets for final classification by classifier

**Preprocessing**

*Tokenization*

Tokenization is the process of chopping up sentences into smaller pieces (words or tokens). The segmentation into tokens can be done with decision trees, which contains information to correctly solve the issues you might encounter.

***Stop Word Removal***

In order to check accuracy of prediction with stop words Removal (common words in English, but with no sentiment information) from the text reviews Porter corpus of stop words from Natural Language Toolkit (NLTK) was used.

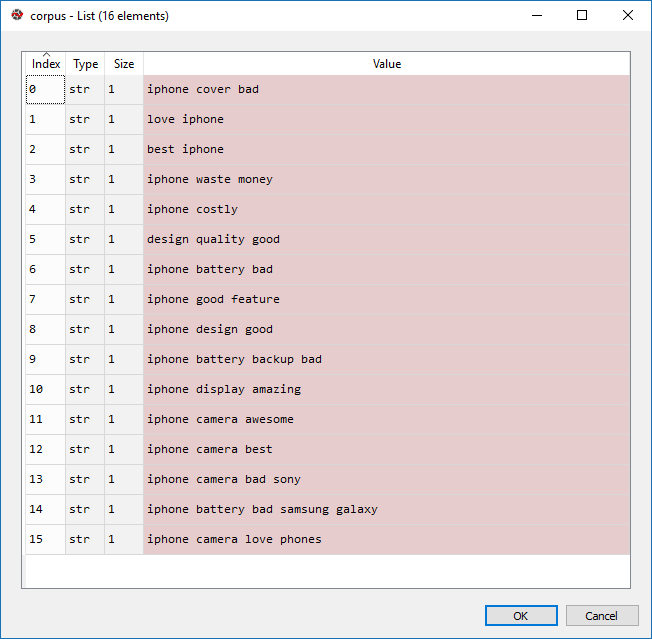
**Bag of Words**

This project uses Bag of words approach for feature selection. The bag-of-words model is one of the simplest language models used in NLP. It makes a unigram model of the text by keeping track of the number of occurrences of each word. This can later be used as a features for Text Classifiers. In this bag-of-words model you only take individual words into account and give each word a specific subjectivity score. If the total score is negative the text will be classified as negative and if its positive the text will be classified as positive. It is simple to make, but is less accurate because it does not take the word order or grammar into account. Steps Followed -

* To keep track of the number of occurrences of each word, tokenized the text and added each word to a single list. Then by using a Counter element kept track of the number of occurrences.
* Made a DataFrame containing the class probabilities of each word by adding each word to the DataFrame as we encounter it and dividing it by the total number of occurrences afterwards.
* These words in this constructed Sentiment Lexicon were used to give a value to the subjectivity of the reviews in the test set.

**Classifier**

Naive Bayes classifier was used using sklearn, having 80% train data and 20% test data.



Result

Sentences Predicted Sentiment

Iphone battery is bad than Samsung galaxy Negative

iPhone Camera Is best Positive

with increase in dataset performance of classifier can be improved.and using other algorithms like KNN, logistic regression, random forest classifier performance of classifier can be enhanced.

REFERENCES

[1] [Online]. Available: <https://datajobsboard.com/wp-content/uploads/2017/01/Naive-Bayes-Kevin-Murphy.pdf>

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol. 10, pp. 79-86, 2002

[3] X. Yun, X. Wu and Q. Wang, "Sentiment Analysis of Yelp’s Ratings Based on Text Reviews", pp. 1-5, 2014. [Online]. Available: <http://cs229.stanford.edu/projects2014.html>

[4] J. Jong, "Predicting Rating with Sentiment Analysis," pp. 1-5, 2011. [Online]. Available: <http://cs229.stanford.edu/proj2011/JongPredictingRatingwithSentimentAnalysis.pdf>