Poznan University of Technology
Faculty of Computing
Institute of Computing Science

Master's thesis

RESTAURANT RECOMMENDER SYSTEM (RSS) BY SENTIMENT ANALYSIS ON YELP DISH REVIEWS

Omer Gokdere

Supervisor
dr. Adam Wojciechowski

Poznan , 2018

Acknowledgement

Firstly, I would like to express my sincere to my graduation committee for their valuable advice and feedback on my research

Secondly, I would like to thank my supervisor dr. Adam Wojciechowski, for all the assistance and supervision during the project.

Also, special thanks to prof. dr hab. Jerzy Nawrocki for guidance and providing me with feedback in the project.

At last, I am extremely grateful to my brother and my parents for having faith in me and for extending their support at all times. In addition, I would also like to thank my friends for making everything bearable.

Summary

The present thesis provides an exploratory research on text classification in English language. This project is based on customers' reviews on Yelp restaurants.
Sentiment analysis on customers' reviews to identify positive and negative reviews on on specific dishes of the restaurants.

The main objective of this study sentiment analysis on customers' reviews on specific dishes of the restaurants as "positive" or "negative" to give reliable restaurant recommendation for the user to make an easy and quick decision.

In this thesis, we proposed a solution for people to make an easy and quick decision with good quality related to customers' reviews of the dishes of restaurants.

List of Figure

List of Tables

Contents

Chapter 1

Introduction

1.1 Introduction

It is human nature to ask acquaintances when it comes to take a decision. Most of their decisions in the real world are affected by the thinking that, how other people would perceive/see their decision. Almost fifteen years ago, when the internet was not so accessible, people used to make decisions about buying some service or product based on their friends' or experts' recommendation. Nevertheless, the available volume of information for decision making was limited. However, with the usage of the internet, the big data explosion and the capacity of the people to discover and exploit the web has made the massive amount of information available, which can be used to take decisions objectively.

Nowadays opinion mining is a quite growing topic as proportional as usage of the internet by people and sharing information globally. Due to the increase in social networking people started to share their opinions, feelings, reviews, and criticism with each other.

The development of e-grocery allows people to order online food. As a number of customers' rates and reviews on the same product or restaurant, allows a user to decide whether to buy a product or not. However, customer reviews are more important than ratings. In customer reviews, we can find good and bad reviews about foods. However, ratings are not showing which product quality is not good or average. This information is necessary for the organization that is selling or manufacturing products in order to make changes in design and another configuration of the product.

Yelp is a local-search service powered by crowd-sourced review forum, as well as an American-multinational organization was founded in 2004 to help people to find great local businesses such as dentists, hairdressers, and repair shops. The primary purpose of Yelp is to provide a platform for customers to write a review along with providing a star-rating along with an open-ended comment. Yelp data is reliable, up-to-date and has full coverage of all kinds of businesses. Millions of people use yelp, and empirical data demonstrated that Yelp restaurant reviews affected consumers' food choice decision-making.[1]

The proposed solution gives an enhanced summarized result of the recommended restaurants based on the machine learning algorithms for users to take a fast decision. It extends the level of feature-based opinion classification. A soft computing technique is used to classify opinion into positive and negative reviews. By using the proposed system, it is easy for the user to order good quality food as well as to take a fast decision.

1.2 Aims and Scope

This thesis aims to examine if it is possible to use Machine Learning methods to identify customers' reviews. We limit the type of reviews to restaurants. Also, we focus on reviews expressed in English.

Thesis organized as follows:

- Chapter 2 presents Yelp dataset, the technologies that used during machine learning process and other studies concerning customer reviews and introduces Natural Language Processing (NLP), sentiment analysis, and supervised machine learning.

- Chapter 3 presents how we collected, filtered and applied pre-processing methods to our dataset also supervised classification methodologies that implemented to find the best accuracy.

- Chapter 4 is an experimental validation of the proposed method for identifying positive and negative reviews. We present the research methodology of the validation study. Then, we prepared the interface of the RRS.

- Chapter 5 summarizes the main findings of this thesis and discusses some future directions of research.

Chapter 2

Background and Related Works

Since we have taken yelp dataset for the project, there are many advantages. The yelp services supply these data freely for the user and encourage programmers to participate in yelp dataset challenge in which the participant can come up with an algorithm which can predict the business rating efficiently based on the presented dataset and generate a reasonable comparison with the expected release of the dataset. There are many similar projects available based on yelp dataset.

2.1 Yelp Dataset

For the present research, the yelp academic dataset used which provided by Yelp corporation. Yelp connects people with local businesses, and the dataset provides valuable information about customer's experiences at every business via reviews, tips, check-in and business attributes since 2004. The range of local businesses in the chosen dataset is mostly in USA, Canada and some parts in Germany and UK. Yelp provides a way for users to explore, rate and review the businesses they visit. Businesses can highlight their products and services that may attract users to them and finally rate the business.[2]

Yelp dataset contains a wide variety of businesses, like restaurants, bars, cafes, doctors, pharmacies, hotels and so on. Users can give a star rating from 1 to 5 for a business, and can also write a text review which clarifies the rating. These ratings are very helpful for users who are searching local business and help them in deciding which one would be the best for them. These features of Yelp make it a highly recommended system.

2.2 Python

Python is an object-oriented programming language with high-level built in data structures, combined with effective typing and effective binding, make it very attractive for Rapid Application Development. [8] Python's simple, easy to learn syntax to maintain readability. Mostly, programmers started to prefer Python programming language because of the enhanced productivity it provides. On the other hand, there is no compilation step, the fastest way to debug a program is adding some statements to the source: the fast edit-test-debug cycle makes this simple approach very useful.

Some core libraries that is used by python during the project is given below [9].

NumPy	library allows vectorization of mathematical operations on the NumPy array type, which improves performance and hence speeds up the execution
SciPy	provides effective numerical methods as numerical integration, optimization, and many others via its specific submodules.
Pandas	a perfect tool for data operations designed for fast and simple data collection, manipulation and visualization.
Seaborn	is focused on the visualization of statistical models; such as heat maps that summarize the data but still represent the overall relationships.
NLTK	provides lots of operations such as text tagging, classification, tokenizing, stemming and semantic reasoning
SciKit-Learn	exposes a brief and consistent interface to the standard machine learning algorithms, producing it as easy as it brings ML into production systems

Table 1: Some core python libraries

2.2.1 Pandas in Python

For data scientists, managing with big data is typically separated into two steps: data pre-processing and modeling it, then providing the results of the analysis into plotting or tabular display.

Pandas is one of the most useful Python package that provides quick, adaptable, and significant data structures designed to work with "labeled" data both intuitive and straightforward [9]. It proposes to be the fundamental high-level building block for doing practical, real-time data analysis in Python. Additionally, pandas has the wider goal of becoming the most robust and flexible open source data analysis and manipulation tool available in any language.

Pandas support many different sets of data:

- Tabular data as in an SQL table or Excel spreadsheet
- Ordered and unordered series data.
- Random matrix data with row and column names

2.3 Natural Language Processing

In the late 1980s and mid-1990s, there was a leap in technological developments. More advanced computers created an innovative era for machine learning.[3] Since many types of Natural Language Processing researches have depended on statistical data (as called machine learning), implementing NLP systems has become manageable and affordable.

Natural Language Processing (NLP) is a field of Artificial Intelligence and Linguistics, loyal to making computers understand the words and declarations spoken or written in human languages. Natural language processing tries to close the gap between human and machine, by extracting useful information from natural language messages.

Some of the significant fields of NLP: Question Answering Systems (QAS), Summarization, Machine Translation, Speech Recognition, Document classification. Researchers have hard worked on NLP to bring NLP today as its level. Sentiment Analysis, Parts of Speech (POS) Tagging, Chunking, Named Entity Recognition (NER) mostly used methods of NLP.

The purpose of NLP is to provide one or more specialties of a system. The basis of NLP evaluates an algorithmic system which allows for the integration of language generation and linguistic understanding.[4] purposed a unique modular system for \cross lingual case extraction for English, Dutch and Italian texts by using various pipelines for several languages.

2.4 Sentiment Analysis

Sentiment Analysis (aka Opinion mining) is the computerized process of understanding an opinion about a presented subject from written or spoken languages. It is a natural language processing in which the opinion of the person is classified to be of positive, negative or neutral polarity.

The data source for sentiment analysis is varied, from the web [blogs, social media, online customer reviews] to handwritten documents. It uses lexicons-based methods or different machine learning algorithms to extract words/phrases from the text and classify them as positive or negative or neutral opinion.[5]

Sentiment analysis and methods:

The following figure shows the sentiment analysis process for customers' reviews as the input data:

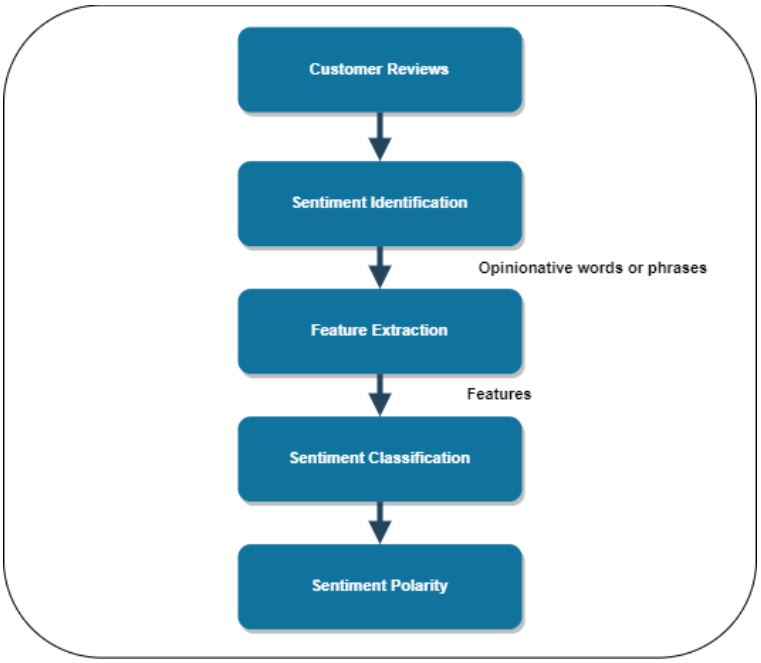

Figure 1: Sentiment analysis process (Source: Sentiment analysis algorithms and applications: A survey. (n.d.))

Customers' product reviews, the terms/expressions are extracted from the input in Sentiment Identification. Depend on the features [Part-Of-Speech tags N-grams, frequency, negations...] of the words; they are selected in Feature Selection creating the feature vector. These features are then classified in Sentiment Analysis to detect the Sentiment Polarity

The sentiment classification methods are given in the figure below:

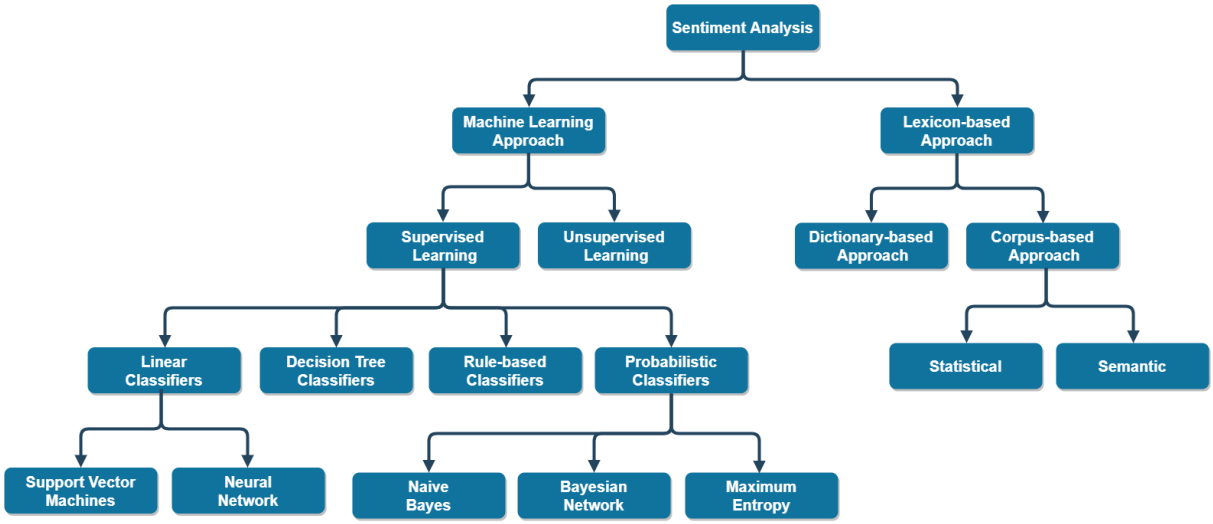

Figure 2: Sentiment Analysis methods (Source: Sentiment analysis algorithms and applications: A survey. (n.d.))

Levels of Sentiment Analysis:

Sentiment analysis can be conducted at different levels which are word-level, sentence-level, document-level and aspect-level.

The figure below shows the different levels.

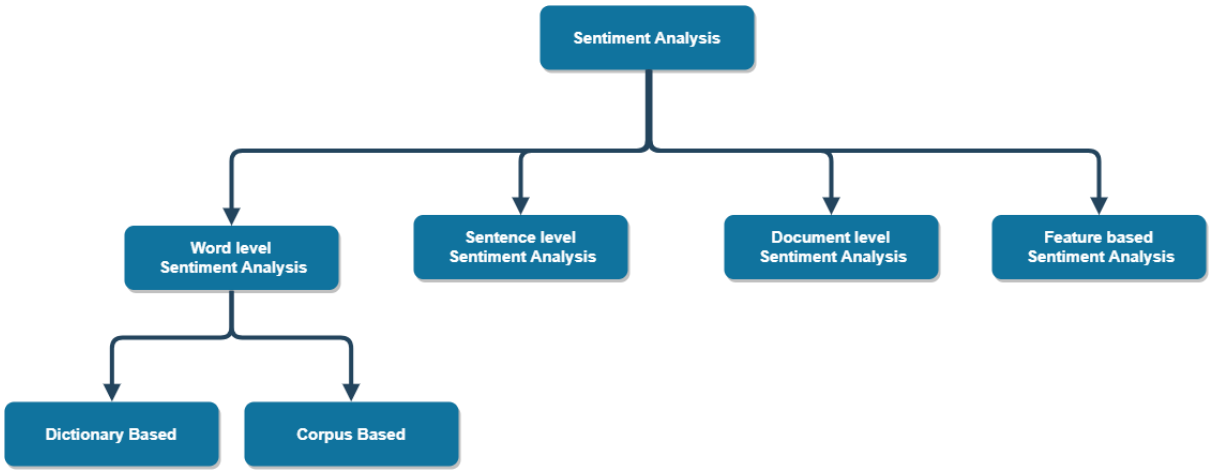

Figure 3: Levels of Opinion mining or Sentiment Analysis (Source: Sentiment analysis algorithms and applications: A survey. (n.d.))

The word-level sentiment analysis determines the basis of document-level, and sentence-level sentiment analysis as the phrases in sentences analyzed for positivity or negativity of the opinion.

There are two approaches to word-level sentiment analysis:

- Dictionary-based approach, in which the inadequate list of words that created with known polarity and the list in prolonged with synonyms and antonyms [assigning each to proper polarity] from online words source, an online dictionary
- Corpus-based approach in which the word polarity is determined from its co-occurrence with another known polarity word, relying on syntactic and statistic methods.

Document-level sentiment analysis calculates the opinion presented by the subjective and objective sentences in the document as a whole. At sentence-level sentiment analysis, the subjective sentences are analyses, and opinion polarity is detected.

In aspect-level or feature based sentiment analysis the person's opinion about the features of the article is analyzed, which cannot be extracted from sentiment or document level analysis, as a different person can have a different opinion about the features of the entity

2.5 Supervised Machine Learning

In the classification process, the class label attached to the input. The class labels are pre-defined. The supervised classifier is created based upon the training set of input data, corpora, holding a specific label for each input. Input data are considered to be independent of each other. [6]
The figure below shows the Supervised Classification:

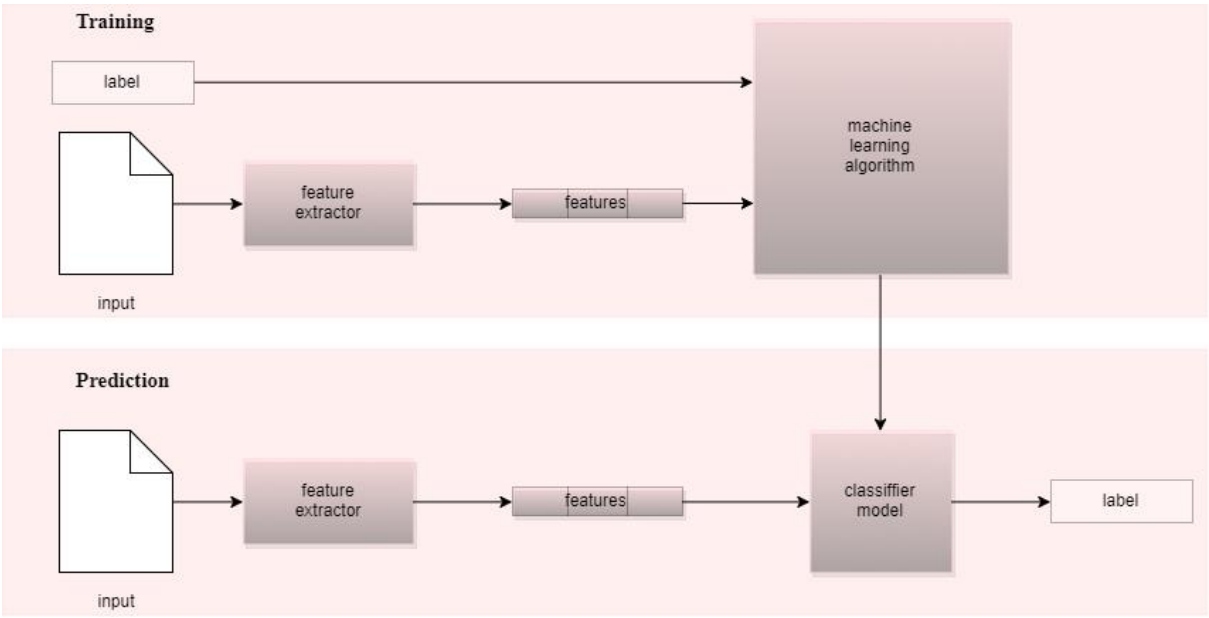

Figure 4: Supervised classifier (Source: Bird, S., Klein, E., & Loper, E., 2009)

Supervised text classification works on training and testing principle. We support labeled data to the machine learning algorithm to process. The algorithm is trained on the labeled dataset and returns the desired output (the pre-defined categories). During the testing phase, the algorithm is filled with unobserved data and classifies them based on the training phase.

Chapter 3

Restaurant Recommender System

In this chapter exploration and creation of the dataset and experiments that are done to achieve a better classifier.” Dataset” section consists of information about how the dataset collected, data selection methods and general statistics about the dataset. Then the” Preprocessing” section will be mentioning the steps that were taken to reduce the number of ambiguous attributes and prepare the best version of the data possible for the classification phase. At last,” Classification” challenges, strategies, experiments in various points throughout the work will be written in this section.

3.1 Dataset

We use the Yelp Challenge Dataset stored in JSON format which consists of 6.1 million reviews from 1.5 million users from 189 000 local businesses such as restaurants, bars, cafes, local events, doctors, pharmacies, hotels. (to around 7 GB of data). The chosen dataset is mostly in USA, Canada, and some parts in Germany and the UK. The dataset contains data about business id, name, location, category, users, reviews, dates, stars, longitude and latitude values, zip codes, attributes.

For our research, since we are only interested in the restaurant data, we have considered out only those businesses that categorized as restaurants or foods in the range of local businesses in the USA. This filtering reduced the number of reviews to 3 million, business to 35 000. (to around 3.5 GB of data in CSV format).

Ratings	Count
1	353705
2	275112
3	373696
4	747312
5	1210334

Table 2: Statistics of the dataset after collection of data

We consider that the star rating is a critical measure for the sentiment opinion of the review. The star rating of a business review is an integer from one to five. We decided to extract all star ratings and their corresponding review which are equal to three and to keep all ratings below three which

considered as “negative” sentiments, and also to keep all ratings above four which considered as “positive” sentiments. We collect a dataset containing 2.6 million reviews (to around 2.9 GB of data).

The business dataset merged with the reviews’ dataset by the “business id” attribute. RRS is filtering dataset according to entered input by user which are “zip code”, “distance”, and “desired food or restaurant type”.

Later, words in each review separated and the punctuations were removed so that a “bag of words” was generated for each review. Finally, we stemmed, lemmatized and filtered out the stop words in each bag of words using both the in-built list in Python’s NLTK package. The merged review-business data were randomly separated into training, validation and testing set according to ratio 3:2:5.

3.2 Preprocessing

In nearly all research about Yelp sentiment analysis pre-processing techniques are incorporated. Pre-processing is one of the most data mining tasks which includes the preparation and transformation of data into a suitable form to mining procedure. Pre-processing techniques are essential to obtain a more clean dataset. Cleaning the dataset increases the performance of the following classification system significantly. We performed pre-processing on our dataset to enhance the accuracy of our models. Possible clarifications methods applied, like filtering letters, punctuations, words, tokenization and improving simple errors.[7]

Tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation. Improving the simple type errors, such as a misspelling and repeated letters based on dictionaries. The dictionaries used to fix the errors. Thus, abbreviations and acronyms replaced with words from a predefined dictionary.

We have applied stopword removal technique to remove useless content. Stopword removal is a pre-processing method that tries to extract tokens from documents that are considered that very little significance in the classification of sentiment analysis. For instance, conjunctions, like “and” and “but” may be of no relevance to the ultimate sentiment score and can be extracted from documents.

Stemming algorithms are also the method that we used in our pre-processing step. Group of different tokens that might be deriving from the same root of the word. Stemming is a method to remove the conjugations of verbs to their stem, the original root form. For example, “make” and “making” may be replaced by the same token ‘make’ before training the dataset, as the opinion classification can be considered to be remarkably similar.

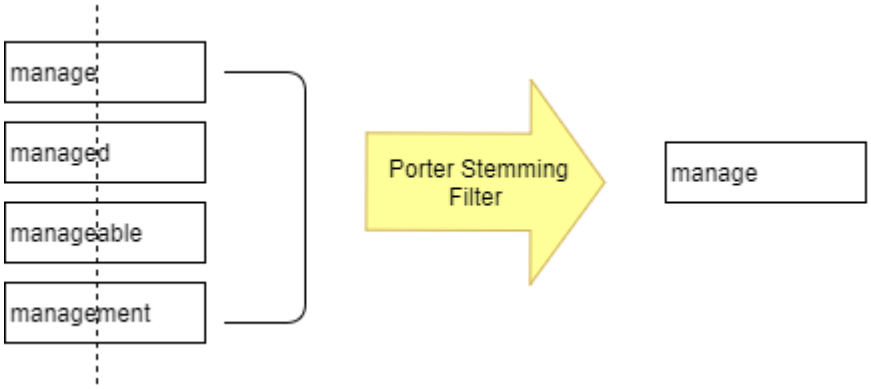

Figure 5: Porter Stemming Filtering Process

Another useful technique called a part-of-speech tagger (or shortly a POS tagger) is the method of specifying a 'label/category' to every token in a given sentence. For example, "phone/**NN** is/**VB** great/**JJ**" ... In the English language, common POS categories are:

Description	Tag	Description	Tag
Nouns singular	NN	Prepositions	IN
Nouns plural	NNS	Conjunctions	IN
Verb, base form	VB	Adverbs	RB
Verb, past tense	VBD	Adverbs, comp	RBC
Verb, present	VBG	Adverbs, sup	RBS
Adjective	JJ	Personal pronoun	PRP
Adjective, comp	JJR	Wh-pronoun	WP
Adjective, sup	JJS	Possessive pronoun	PRP$
Cardinal number	CD	Interjections	UH

Table 3: Common POS Tagging categories.

3.3 Supervised Classification

3.3.1 Feature Extraction

3.3.1.1 N-gram model

The n-gram technique is a relatively simple algorithm. The items are generally letters or words. A n-gram is an adjoining sequence of n items from a written or spoken data. In the case of unigrams ($n = 1$), each text represents a document and split up into words. Calculating the frequency of all the words in all documents results in a word frequency table. [7]

There are two possible steps of estimating the frequency among whole documents: the summed term frequency and document frequency. The term frequency *(tf(w, d))* is the number of times that a word w occurs in document d, given in the equation *[1.1]*. In computing the term frequency, all events of a word in a document are counted. Therefore, the term frequency can assume a value in the interval *[0-n]*, where n is the absolute number of words in the document.

The term presence *(tp (w; d))* only checks if a word w is present within a document d, which results in binary value. This calculation is given in equation *[1.2]*.

$$tf(w,d) = |\{w \in d\}| \qquad [1.1]$$

$$tf(w,d) = \begin{cases} 1, & if\ w \in d \\ 0, & if\ w \notin d \end{cases} \qquad [1.2]$$

The summed term frequency *(df(w, D))* counts all terms' frequencies of a word w across all documents D, the formula given in equation *[1.3]*. The frequency of the document refers to the number of documents in which a word occurs. The formula is given in equation *[1.4]*, where the document frequency is *df(w, D)* of word w across all documents D.

$$stf(w,D) = \sum_{d \in D} tf(w,d) \qquad [1.3]$$

$$df(w,D) = |\{d \in D : w \in d\}| \qquad [1.4]$$

According to higher order n-grams, the texts divided as n-length items. In the case of $n = 2$ (bigrams), the items consist of the following two words. The same holds for $n = 3$ (trigrams). The set includes all series of two words or three words that are following in the original text. Intuitively, those higher order n-grams seem to catch the relation of the consecutive words more meaningful, e.g., in negation. Finally, a selection produced of the most valuable words. Only the top k most relevant n-gram features, according to the weight measurements are selected to form the feature vector.

3.3.1.2 Term Frequency – Inverse Document Frequency

The TF-IDF (term frequency-inverse document frequency, *tf-idf*) measure is a statistic that shows the importance of a word across a set of documents. This measure is composed of two individual measures: the term frequency and the inverse of the document frequency. We have mentioned the term frequency and the normal document frequency at previous section *3.3.1.1* and their formulas sequentially given in equation *[1.1]* and *[1.4]*.

The inverse document frequency is used to measure the rareness of a word across all the documents. How higher the value of the inverse document frequency, how rarer the word across the set of documents is. The inverse document frequency, $idf(w; D)$ of a word w across all documents D given in equation *[2.1]*. Combination of term frequency to the tf-idf measure shown in equation *[2.2]*.

In this equation, the tf-idf is the $tf\text{-}idf(w; d; D)$ of a word w in document d across a set of documents D.

$$idf(w, D) = log\ \frac{|D|}{df(w,D)} \qquad [2.1]$$

$$tf - idf(w, d, D) = tf(w, d) \cdot idf(w, D) \qquad [2.2]$$

However, it is the question of whether the tf-idf is a good measure for feature selection in this research. The tf-idf value is high when a word often occurs in a document and does not often occur within all documents.

Derivatives:

A number of term-weighting schemes have derived from tf–idf. One of them is TF–PDF (Term Frequency * Proportional Document Frequency).[13] TF–PDF was introduced in 2001 in the context of identifying emerging topics in the media.

3.3.1.3 Naïve-Bayes Classifier

Naive Bayes (NB) algorithms based on using Bayes theorem with a naive assumption, that every feature is independent of each other, in order to predict the category of a given sample. [12]NB algorithms are probabilistic classifiers that calculate the probability of every category using Bayes theorem and outputs the category with the highest probability. NB classifiers have implemented too many domains, particularly NLP. Bayes' Theorem provides a calculation of the probability of a hypothesis given our prior knowledge.

Bayes' Theorem declared in the given formula *[3.1]*

$$P(h|d) = \frac{P(d|h) * P(h)}{P(d)} \qquad [3.1]$$

In the classification process, our hypothesis *(h)* may be the class to assign for a new data instance *(d)*. *P(h)* refers probability of hypothesis h being true (regardless of the data d) and where *P(h|d)* is the probability of hypothesis h given the data d. Also, *P(d)* refers to the probability of the data (regardless of the hypothesis). Lastly, P*(d|h)* is the probability of data d given that the hypothesis h was true.

We perform the multinomial NB event model with additive smoothing. This NB model provides a solid beginning for our semantic analysis. The multinomial NB model generates one term from the vocabulary in each position of the document. First, we divided texts into their respectively ranking, so the occurrence of particular words for a particular star rating fits into the same bin.

Later we have trained every single word in either a negative or positive rating, we have tested the resulting words of an unknown review for its possible star rating. Multinomial NB is comparatively basic to other NB implementations because it measures the probable existence of particular words within a particular classification. As an example, it ignores word sequence and grammar, and it does not return a proper probability ratio of two-word similarities.

3.3.1.4 Support Vector Machine

Support vector machine technique is highly efficient statistical classification algorithm that classifies data by dividing into two classes with the help of a functional hyperplane. A marginal area of functional hyperplane also called as "danger zone" is defined to be the area between two parallel hyperplanes that are determined by the average distances of the support vectors from the two classes

to functional hyperplane. The border size based on the furthest points among each group. We can say that the main idea behind the support vector machine is the give more attention to the locations where the classification is possible to fail. This technique reduces the statistical risks during the classification.

There are two types of SVM which are linear and non-linear SVM. Linear SVM divides the data points using a linear decision boundary.[11] Non-linear SVM divides the data points using a non-linear decision boundary. We perform a support vector machine (SVM) that uses a linear kernel.
For a linear SVM, the equation for the decision boundary is given in *[4.1]*

$$w \cdot x + b = 0 \qquad [4.1]$$

where *w* and x are vectors and the direction of *w* is perpendicular to the linear decision boundary. Vector *w* is determined using the training dataset. For any set of data points (x_i) that lie above the decision boundary the equation given in *[4.2]* also, for the data points (x_j) which lie below the decision boundary, the equation given in *[4.3]*

$$w \cdot x_i + b = k, \text{ where } k > 0, \qquad [4.2]$$

$$w \cdot x_j + b = k`, \text{ where } k` < 0. \qquad [4.3]$$

By rescaling the values of w and b the equations of the two supporting hyperplanes (h_{11} and h_{12}) can be defined as

$$h_{11}: w \cdot x + b = 1 \qquad [4.4]$$

$$h_{12}: w \cdot x + b = -1 \qquad [4.5]$$

The distance between the two hyperplanes (margin "*d*") is obtained by

$$w \cdot (x_1 - x_2) = 2 \qquad [4.6]$$

$$d = 2/||w|| \qquad [4.7]$$

Chapter 4

Validation

4.1 Research Methodology

The purpose of this section is to explain the evaluation metrics applied in this work to evaluate the predictions for the selected algorithm models. For validation schemes, accuracy is not a good measure for imbalanced sets because failing algorithms may always predict" good" and still get 80 % accuracy. In validation, Accuracy, F-measure, and Matthew correlation values were taken into account. Computational time is not estimated to be a criterion for the selection of the algorithm model

4.1.1 Research questions

How reliable recommender system model can be proposed which uses machine learning techniques to classify "positive" and "negative" customers' reviews on restaurant dishes?

4.1.2 Validation procedure

We use the Yelp Challenge Dataset 6.1 million reviews from 1.5 million users from 189 000 local businesses such as restaurants, bars, cafes, local events, doctors, pharmacies, hotels. After collecting the data, the dataset has been filtered to eliminate non-restaurant reviews since our study based on restaurant reviews.

The star rating is an essential measure for the sentiment opinion of the review. The star rating of a business review is an integer from one to five. We decided to extract all star ratings and their corresponding review which are equal to three and to keep all ratings below three which considered as "negative" sentiments, and also to keep all ratings above four which considered as "positive" sentiments.

Then, the preprocessing is examined on the dataset to prepare for the classification process. Naive Bayes, and SVM algorithms used for predictions. After collecting the results, we have classified 2000 review by ourselves as a collection of survey for the comparison of two algorithms with the predictions of the same reviews gave information about how realistic the predictions. The comparison between the collection and the predictions, answer the question of whether classifying poor customer reviews based on the statistics be as good as manual classification.

4.1.3 Validation Schemes

The training set has the" stars" attribute that is used to train the classifiers. All tests run on 10 times cross validation.

4.1.4 Prediction quality measures

The prediction quality measured F-measure, Matthews correlation and the mean absolute error. [12]Precision simply is the ratio of true positives (TP) to all positives retrieved $(\mathrm{TP}+\mathrm{FP})$ and formulation is given in *[5.1].*

$$\frac{TP}{TP+FP} \qquad [5.1]$$

Recall on the other hand is the ratio of true positives to the all real positives and formulation is given in *[5.2].*

.

$$\frac{TP}{TP+FN} \qquad [5.2]$$

\

F-Measure (aka f-Score or F1-score) uses precision and recall to calculate a score which can be interpreted as 'Weighed average of precision and recall.'

Formula of F1-score is the mean of precision and recall given in *[5.3].*

$$\frac{2 * Precision * Recall}{Precision+Recall} \qquad [5.3]$$

The Matthews correlation coefficient (also known as MCC) is a measure of the variety of binary classifications.[13] MCC works on true and false positives and negatives, and it returns a value between 1 and +1. The coefficient of +1 represents a perfect prediction, 0 no better than random prediction and 1 shows the total conflict between prediction and observation. Formulation of MCC is given in *[5.4]*

$$\frac{\mathrm{TP}*\mathrm{TN}-\mathrm{FP}*FN}{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)} \qquad [5.4]$$

Mean Absolute Error (MAE) is a measure of how close estimates or predictions to the real outcomes. Alternatively, it is the difference between two set of variables or following variables.

Formulation is given in *[5.5].*

$$\sum_{i=1}^{n} \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \qquad [5.5]$$

4.2 Results

Algorithm	Accuracy	The lowest F-measure	The lowest Mcc	Mean absolute error
J48	81.5 %	0.32	0.32	0.21
Naive Bayes Multinomial	95.5 %	0.81	0.82	0.02
SVM	82.7 %	0.44	0.77	0.16
Adaboost + SVM	81.9 %	0.88	0.56	0.15
RealAdaboost + SVM	89.9 %	0.61	0.65	0.1
Bagging + SVM	84.1 %	0.22	0.52	0.14

Table 4: Experiment on classification

4.3 Threats to validity

In this section, we address aspects starting with external threats. There are two external threats, both concerning the metrics that used. After that, we discuss the internal threats; one is a result of the way we pre-process the data and the other petioles from the way the models tested.

4.3.1 Internal Threats

The dataset has two parts, training, and test sets. Training and test sets collected from Yelp API. Since two sets consisted reviews from same data sources, we have included many reviews from many different cities which written for many different venues. This has increased the variety and reduced the heterogeneity of items in the training set. Review predictions have made on the whole dataset to obtain a significant quality of reviews.

RRS is using another API which is ZipCodeAPI to collect zip codes according to entered distance and zip code. Dataset is preparing for reliable results after its filtered by user input data which are "desired dish," "zip code" and "distance."

4.3.2 External Threats

Firstly, the way the participants have been selected. The participants were not selected among a uniform group randomly. The survey has given to mostly graduate students, acquaintances, friends. This has limited the survey group in the name of the location and age group. For example, the participants were mostly from the cities of Poznan, age population of the participants between 18 and 27. This threats the validity of the survey results. Second threat is the survey answers. The survey

participant is asked to classify a review, into ”positive" or " negative. “Comparing the survey results with 3-class predictions is not an issue, but the binary predictions needed a prior work of conversion.

Another question of validity is, if the dataset truly represents the general. The dataset has collected from big cities in the United States. Those cities located in different regions; thus it grasps the general. Also, the smaller cities do not contain as many reviews about local businesses as the big cities; this means the level of homogeneity of the reviews for those cities is low.

Another observation was that the negative reviews were mostly more longer than the others. Length of negative reviews can be due to the fact that in negative reviews, people tend to tell a short story or an experience. Oppositely, the positive reviews tend to stay short and tell the best part of the venue such as smooth coffee or delicious cakes. Another thing which is worth to mention is that the number of positive reviews is much higher than the negative ones. However, when we observe the content of the positive reviews, a part of them can be classified as ”average.”

Chapter 5

Conclusions

In this thesis, we proposed a solution for people to make an easy and quick decision with good quality related to customers' reviews of the dishes of restaurants.

The results are evaluated using the cross-validation method on restaurant review based on the classification accuracy. Our dataset is depending on user input data which "desired dish or restaurant type," "ZIP code," and "distance."

We have tested four different zip code with the same distance and desired dish type, and accuracy of evaluated models given in order 89%, 90%, 86%, 88% over 5000 reviews.

Figure 6: Interface of Restaurant Recommender System

Bibliography

[1] Michael Luca "Reviews, Reputation, and Revenue: The Case of Yelp.com". Harvard Business School (2016), Working Paper 12-016

[2] Kyle Carbon, Kacyn Fujii, and Prasanth Veerina. "Applications of Machine Learning to Predict Yelp Ratings". Stanford University, CA

[3] Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association : JAMIA*, *18*(5), 544-51.

[4] Isabelle Moulinier and Jean Ganascia. "Applying an existing machine learning algorithm to text categorization". In: *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing* (1996), pp. 343–354

[5] Li, Xinmiao et al. "A global optimization approach to multi-polarity sentiment analysis" *PloS one* vol. 10,4 e0124672. 24 Apr. 2015, doi:10.1371/journal.pone.0124672

[6] Kamal Nigam, Andrew McCallum, and Tom Mitchell. "Semi-supervised text classification using EM". In: *Semi-Supervised Learning* (2006), pp. 33–56

[7] Sakr, S., Elshawi, R., Ahmed, A. M., Qureshi, W. T., Brawner, C. A., Keteyian, S. J., Blaha, M. J., … Al-Mallah, M. H. (2017). Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercIse testing (FIT) project. *BMC medical informatics and decision making*, *17*(1), 174. doi:10.1186/s12911-017-0566-6

[8] Evaluating Rapid Application Development with Python for Heterogeneous Processor-based FPGAs Andrew G. Schmidt, Gabriel Weisz, and Matthew French Information Sciences Institute, University of Southern California

[9] William W Cohen and Yoram Singer. "Context-sensitive learning methods for text categorization". In: *ACM Transactions on Information Systems (TOIS)* 17.2 (1999), pp. 141–173.

[10] Data Structures for Statistical Computing in Python - Wes McKinney

[11] "Toward a Progress Indicator for Machine Learning Model Building and Data Mining Algorithm Execution: A Position Paper" *SIGKDD explorations : newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining* vol. 19,2 (2017): 13-24.

[12] Du, J., Xu, J., Song, H., Liu, X., & Tao, C. (2017). Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. *Journal of biomedical semantics*, *8*(1), 9. doi:10.1186/s13326-017-0120-6

[13] Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PloS one*, *12*(6), e0177678. doi:10.1371/journal.pone.0177678