2020/2021

A picture containing shape

Description automatically generated

**732A92 Text Mining Final Project**

**Chatbot for a service company**

Student

K.M. Mudith Chathuranga Silva (kogsi273)

Examiner

Prof. Marco Kuhlmann

Table of Contents

[Abstract 3](#_Toc58941557)

[1. Introduction 4](#_Toc58941558)

[1.1 Motivation 4](#_Toc58941559)

[1.2 Objective 4](#_Toc58941560)

[2. Theory 5](#_Toc58941561)

[3. Data 9](#_Toc58941562)

[Bibliography 14](#_Toc58941563)

# Abstract

Since the development of social media networks, more and more businesses have been using chat services to attract more customers by providing real-time information for their potential or existing customers. Medium or large scale businesses have a dedicated customer service person to cater to their online customers’ needs. But for small-scale businesses are unable to compete with their rival companies since they have a limited amount of employees and unable to pay for an extra person for online customer handling. And also there are some special service-related businesses in which the owner needs to directly interact with customers. So they can’t use an employee to handle their customer because of the business model or they can’t expose customer information with another party.

Over the last few years, we could see that Text Mining and Natural language processing has been spreading to varies kind of industries. So we could use the Text mining knowledge to address the above situations and create an automated chatbot. We could extract existing customer's text messages from a specific industry and create a simple chatbot that could communicate with customers and solve their problems in real-time.

# 1. Introduction

## Motivation

I have a friend in Sri-Lanka who is running a small-scale saloon. And also he has unique customers such as Actresses, Businessman’s … etc. Most of his customers make/cancel and remove appointments through WhatsApp and Facebook Messenger (Private profile and Business profile). By the way, he won’t use the phone when he is providing a service to a client. Therefore sometimes he can’t reach the phone to reply to his exiting customer or a new customer who is asking general questions or even ask appointments over the chat. Some of the important clients' required a reply instantly and new customers are also demanding a reply from the saloon as soon as possible. Since I’m aware of this problem, I thought that Text Mining may help to resolve this problem somehow.

## Objective

It is always recommended to have a look at the dataset before implementing the solution to the problem. So sample chat records were requested for initial analysis. Then several question groups were identified. Automated replies could be generated from the saloon side by checking the client question type. But the tricky part is to extract customer information for the appointment placement. Then generate an appropriate text according to the customer's reply. The chatbot would be an ideal solution for this kind of scenario. A chatbot could be trained by using previous customer questions and provide a general reply. And also python libraries such as Spacy could be used to identify customer replies and store required data.

# 2. Theory

There are some frequently used terminologies in Text mining and it is required to get some idea about those terms and the theory behind them. Generally ‘Spacy’ and ‘sci-kit’ libraries used for the project and the below terms are corresponding with the library functionalities.

Tokenization

During processing, spaCy first tokenizes the text, i.e. segments it into words, punctuation and so on. This is done by applying rules specific to each language. For example, punctuation at the end of a sentence should be split off – whereas “U.K.” should remain one token. (Spacy - Tokenization, n.d.)

Stop Words

A “stop list” is a classic trick from the early days of information retrieval when search was largely about keyword presence and absence. It is still sometimes useful today to filter out common words from a bag-of-words model. To improve readability, STOP\_WORDS are separated by spaces and newlines, and added as a multiline string. (Spacy - Stop Words, n.d.)

Count Vectorizer

Number of words appear in the given document and represent the count in a vector form. This approach may be ignore rare words when creating the model. To overcome this problem inverse document frequency method introduced. (scikit - CountVectorizer, n.d.)

Inverse Document Frequency

Overall weighted document of a word. This helps to detect most frequent words and penalize them. TfidfVectorizer weights the number of word count in the document by calculating the frequency of the word. In scikit-learn, the tf–idf weight is computed as:- (Wikipedia - tf-idf, n.d.)

* 𝑁 \_denotes the number of documents in the collection.

Text Classification pipeline

Document

Vectorizer

CountVectorizer

Matrix of Document Vector

Tfidtransformer

Tf-id Transform

Matrix of Document Vector

Eg:- SGD,MNB…

Predictor

Vector of Class Labels

*Figure 1.1*

Accuracy, Precision, Recall and F1-Measure

The accuracy of a classifier is the proportion of documents for which the classifier predicts the gold-standard class.

Precision is the proportion of correctly classified documents among all documents for which the system predicts specific class.

Recall is the proportion of correctly classified documents among all documents with gold-standard class.

The F1-measure is the harmonic mean of the two values. A good classifier should balance between precision and recall.

Naïve Bayes Classifier

𝐶 A set of possible classes

𝑉 A set of possible words; the model’s vocabulary

𝑃(𝑐) Probabilities that specify how likely it is for a document to belong to class 𝑐 (one probability for each class)

𝑃(𝑤|𝑐) Probabilities that specify how likely it is for a document to contain the word 𝑤, given that the document belongs to class 𝑐 (one probability for each class–word pair)

Predicted class for the document.

W Count of the word W in the document

Logistic Regression Classifier

The logistic model extends the linear model by adding a pointwise logistic function 𝑓.

The term logistic regression refers to the procedure of learning the parameters of the logistic model. For Binary classifications :-

Chart, line chart, histogram

Description automatically generated

# 3. Data

All the chat messages were extracted from the messenger application from the saloon owner's mobile phone. Chat messages were copied to an excel sheet manually. Some of the messages were contained sensitive information, such as names, phone numbers, general greetings….etc. By the way, it’s essential to remove above information since we need to train the model to identify question groups in general and model do not need to be aware of different kind of names, phone numbers, …etc. Otherwise, the model may be overfitted by considering several names or user related entities. Here are the preprocessed steps on the chat dataset.

Any person name by introducing him/her self :- **username**

Eg:- Hi I’m Jason. Nice to meet you => Hi I’m **username**. Nice to meet you

Any greetings (eg:- Good Morning, Morning, Evening..) **usergreet**

Eg:- Good Morning! I need to know about general hair cut prices =>

**usergreet**! I need to know about general hair cut prices

Any welcome interjections (eg:- Hi, Hello …) **interjection**

Eg:- Hi! I need to make an appointment => **interjection**! I need to make an appointment

Any saloon services (eg:- Hair cut, Coloring, ….) **gservice**

Eg:- I need to know about your saloon hair coloring prices =>

I need to know about your saloon **gservice** prices

Above words were used to train the model to remove redundant words and introduce the generic word to represent all the word groups. All the preprocessed chats were stored on ***UserChatData.xlsx.*** A brief summary of the chats and gold label classes could be generated from the data.

|  |  |
| --- | --- |
| 1 - General Greeting | Used Hi for all injection words (Hey, Hi,..) |
| 2 - Greeting + Checkback | Replaced usergreetings -> Good Morning |
| 3 - Ask Available Services | Eg:- All available services |
| 4 - Ask General Services | Eg :- Hair cut, Colouring … |
| 5 - Make Reservation |  |
| 6 - Change Reservation |  |
| 7 - Remove/Cancel Reservation |  |

*Figure 3.1 – Gold Label Classes*

Table

Description automatically generated

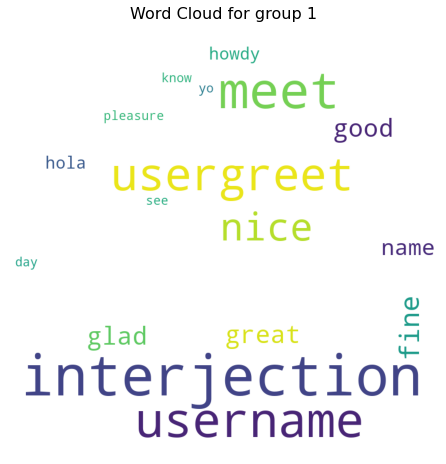
*Figure 3.2 – Data frame head(10)*

Chart, bar chart

Description automatically generated

*Figure 3.3 – Count Plot for Gold label classes*

Word Cloud for Gold label classes.

A picture containing pie chart

Description automatically generated

Text

Description automatically generatedText, application, chat or text message

Description automatically generatedText

Description automatically generatedText

Description automatically generatedText

Description automatically generated

# Bibliography

*Spacy - Tokenization*. (n.d.). Retrieved from https://spacy.io/usage/linguistic-features

*Spacy - Stop Words*. (n.d.). Retrieved from https://spacy.io/usage/adding-languages#stop-words

*scikit - CountVectorizer*. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.feature\_extraction.text.CountVectorizer.html

*Wikipedia - tf-idf*. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Tf%E2%80%93idf