National College of Ireland

# Study of Topic Modelling and Sentiment Analysis with Word Vectorization for a Hotel Review dataset

MSc Research Project

**Siddhant Bakshi**
Student ID: 21143846

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

*National College of Ireland*
*Project Submission Sheet School of Computing*

National College of Ireland

| | |
|---|---|
| **Student Name:** | Siddhant Bakshi |
| **Student ID:** | 21143846 |
| **Programme:** | Research Project (MSCDAD_JAN22B_I) |
| **Year:** | 2022 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Vladimir Milosavljevic |
| **Submission Due Date:** | 14/12/2012 |
| **Project Title:** | Study of Topic Modelling and Sentiment Analysis with Word Vectorization for a Hotel Review data |
| **Word Count:** | 5506 excluding the first two pages |
| **Page Count:** | 20 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | siddhant bakshi |
| **Date:** | 14th December , 2022 |

*PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:*

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | Q |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | Q |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | Q |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Study of Topic Modelling and Sentiment Analysis with Word Vectorization for a Hotel Review dataset

Student Name: Siddhant Bakshi
Student Id: 21143846
MSCDADJAN22B

14 December 2022

**Abstract**

Text classification and Analyzing Sentiments are essential in order to interpret and understand the general opinion within a text. Topic Modelling and Sentiment Analysis are used for the purpose of classifying and categorize unstructured text and for determining the polarity in it. For this research Sentiment Analysis and Topic Modelling is performed on a dataset of Hotel reviews. It is both useful for the customer and the business owner to understand the polarity od reviews, over the years review feedback have become important and have re-shaped the hotel industry. Machine learning and Deep learning models have been used for performing Topic modelling and Sentiment Analysis, Random Forest ,SVM and LSTM are used for Sentiment Analysis. LDA and LSA are used for Topic Modelling. Vectorization of words is also performed to improve computation and word2vec and TFIDF are used for the same. Also hyper-parameterized tuning is done for the machine learning models.

***Keywords- LDA, LSA, Sentiment Analysis, Topic Modelling, SVM, NLP.***

## 1 Introduction

With advancement in technology over the years, more number of people have connected with the internet and the number of online users is ever increasing. Billions of people across the world use the internet and use digital platforms such as

Twitter, Facebook, Instagram, Reddit etc. Such platforms gives users a means to express their opinions, considering the same many other service providers have also added a review platforms along these services. They collect this information to get an understanding of the customer opinions towards the services respectively. These platforms act as a feedback associated along the respective services which is both useful for the customers and the service providers. For the customers it gives them an idea about what they are paying for, whether it is worth paying for or not. Ratings gives customers an idea about the quality of services. For the service providers it can be extremely useful in understanding the customer behavior towards their services whether it is up to the mark or not, what changes they may do in order to improve their services, in general it helps the service providers to understand their customer's expectancies and helps them improve their business. This feedback can be highly useful and must be properly analyzed as it can provide the service providers with critical information that can help them improve their business. Data gets produced in astronomical amounts everyday, it can be highly useful but such large quantity of data cannot be manually analyzed and as a result we mine this data using machine learning and deep learning algorithms. Natural language processing is an important field of data science, machine learning and deep learning algorithms are used for understanding the text and analyzing it. Machine learning and deep learning have a wide range of application and cover a wide range of topics related to NLP such as, sentiment analysis, topic modelling ,text summarization, text categorization, text classification etc.

Since the arrival of OTA the way customers choose their accommodation has changed. Earlier the customers used to focus mainly on ratings of the hotels, they focus of features such as size, capacity and hotel grade, on the main or broad aspects (Molinillo et al. 2016), however OTA has given a new shape to preferences while booking of hotels. It has been seen that customer booking the hotels are increasingly influenced by customer feedback (Yusliani et al. 2019), they no longer just look at the ratings to decide but also they give high consideration to the reviews by the past customers that share their reviews on these platforms. Reviews that get posted by the customers on OTA platforms are been stored and become a textual data source. We study the past reviews of the customers and try to interpret the main highlights the concerns, the goods and the bads from the past review that could be useful for both the new customers as well as the hotel owners. It can help understand the level of quality of services that one can expect during their stay duration, hotel owners can benefit by understanding the key features that customer find up to the mark and the areas that they need to improve, it can also be useful to develop marketing strategies, can lead to financial gains as it gives an insight

and understanding about customer level of satisfactions and expectations. Text analytics tools are being used to aid the interpretation and understanding since the data is in quite a large quantity. Using such tools it simplify the procedure as doing it manually is highly unpractical. There have been advancements over the years in text analysis since there are misspellings, abbreviations, local dialect and emojis, the process of text analysis has been ever evolving. It began in the 1980's and more broadly by the late 90's, NLP natural language processing was transformed (Rafea & GabAllah 2018) and different models were being sought after by the researchers , they have ever since try to come up with ways of improving the accuracy of prediction and fast computational results by building models and improving over time. Text analytics has found various applications across the fields. Topic Modelling is one such renowned technique of text analytics (Rosyidi 2019) , LDA being a topic model used for its study.

Since the online hotel reviews are in such a large quantity been produced daily they tend to be beyond the human capabilities to be analyzed manually or visually. There exists a need for innovation when it comes to analyzing the reviews, there exists a need to find ways of interpreting and understanding the reviews via machines and computers. As a result of this existing demand sentiment analysis or opinion mining comes to the role of identifying and interpreting the reviews (Sutherland & Kiatkawsin 2020). Sentiment analysis has the desired capability of automatically interpreting the reviews. Mining the opinions online or interpreting the polarity of reviews online is a tedious or complex task as there are many challenges surrounding it. One challenge is that the data need to be crawled from the websites where search engine and web spiders play a major role. It is also essential to separate the data of non-reviews and reviews. After this step the process of sentiment classification can be proceeded with. (Tian et al. 2016) found text mining algorithms do not perform as good as they do on topic-based categorization, on sentiment classification (Putri & Kusumaningrum 2017). It is easier to identify topics through keywords but sentiments are represented in a subtle manner. Topic based classification are easier to perform in comparison to sentiment analysis that require a good understanding to begin with (Prasanna & Rao 2019). The aim of sentiment analysis is to interpret the written text in form of a review or an opinion towards a product or service offered in terms of their polarity that is negative, positive or neutral . Sentiment analysis finds its use across various domains such

as product reviews, movie reviews, legal block, customer feedback reviews, hotel reviews etc (Kim & Cho 2020). There exists many other applications such as trying to identify the opinion of public towards a topic of discussion on blog pages or online social platforms and also can be useful to integrate a hot topic with search engine after analyzing the sentiments in them as it enhances the statistical use or incline towards the topic of discussion. Areas of tourism and hotel industry is another such area where we can apply and make use of sentiment analysis. For performing opinion mining the literature indicates the use of two different types of techniques , that include semantic orientation and machine learning (Nguyen & Shirai 2015). The latter follows the rules of simple statistic. Turney, Nasukawa and Yi tried to make out the whole tendency of text using simple statistics (Pietsch & Lessmann 2018). Generally the method is applied for the case of document-level sentiment analysis, that can be seen when Tsou makes the use of statistics on news article for its sentiment orientation and also try depict the celebrities from the public and measures their opinions and uses sentimental orientation of words for the same and also take to his consideration the density, spread and even considers the semantic intensity of existing polarity elements(Pietsch & Lessmann 2018). Sentiment analysis occupies a good weigh and is considered in case of orientation study even tough it belongs to the method of coarse-grained orientation classification due to its good accuracy and simplicity.

Sentiment analysis is widely used in Natural language Processing, it can be simply understood as processing the text to understand the sentiments in it, it gets interpreted in binary form that is 0 and 1 which are for negative and positive respectively. Sentiment analysis is helpful for understanding the overall or general opinion of the public one such area where it finds usage is analyzing reviews. By Analyzing reviews one gets to understand whether the opinions are in favor or against the topic of discussion, where the topic is either the service or a product that the users give a review about. Hotel industry is a Service based industry which is directly based on the quality of services that they provide, they are always trying to understand the nature and behavior of the customers and the opinions they carry towards the service that the hotel industry provides them. Over the years as the digital platform has taken over reviews and ratings have come into consideration and are given high importance. Ratings play an important role for any hotel industry, they determine the quality of services offered by the respective hotel. To understand the rating one must try to understand the reviews as the ratings are a direct reflection of the reviews. Reviews shape the rating, thus analyzing the reviews is very important, one can determine the overall polarity of review

through sentiment analysis but as the reviews can be highly diverse as numerous services are offered and reviewed by the customers. Topic modelling comes to our aid when it comes to understanding the topic of discussion, as the data is highly unstructured and the reviews being highly random it can be difficult to predict what might be the main discussed elements in a given review, thus by the use of topic modelling we can understand the key elements that have been focused in the discussion. By simultaneously using sentiment analysis and topic modelling one can understand the polarity of opinions as well as the main topic of discussion in the given unstructured and random texts. One can make out the key elements of discussion.

There exists many different machine learning and deep learning methods for performing Sentiment analysis and Topic Modelling. The research aims at interpreting the polarity in text and classifying the text into topics. The research question is How can Sentiment Analysis and Topic Modelling be useful in understanding the polarity of opinions and topics of discussion within a text and which model gives us good output? The methods that have been chosen are Random Forest, SVM ,LSTM for performing Sentiment Analysis and LSA, LDA as methods for performing Topic Modelling. Along with applying the methods word vectorization is done using model word2vec. Also tfidf vectorization is performed as well as hyper-parameterized tuning is done for the machine learning models SVM, Random Forest.

# 2 Literature Review

For determining the polarity of context two main approaches were proposed by the authors in paper (Lima et al. 2015). One is a knowledge based approach that is on lexical dictionary and the other one is based on machine learning algorithms. When lexical dictionary approach is used it is found to be vulnerable towards informal language where abbreviations, slangs or swear words are used, in the case of short messages, when special characters exists, in case of usage of mixed language, absence of explicit sentiment etc. Therefore the authors decide to proceed by combining both the approaches under a single framework. This framework has been designed keeping under consideration short tweets messages, where it was difficult to interpret due to slangs and abbreviations etc. The machine learning approach used consists of two different stages that is machine learning algorithm and also an automatic classified ie knowledge based. It is seen that the framework tends to be more effective and it provides a modular approach and is also agile that can be upscale to different modules by making some changes.

The authors chose different parameter settings for the respective models that are Na¨ıve Bayes, Decision tree and Support Vector Machine in the paper (Sidorov et al. 2012).]. Selection of features was done with the selected features being normalized words ngrams and a set of combinations were carried out on neutral,

negative and positive and also on informative set of classes. The above procedures were been implemented on topics that relate to data from tweets regarding the Mexican presidential election and cell phones in the Spanish language.

The authors used the proposed methodologies to analyze opinions for sentiments on web forum in the paper (Abbasi et al. 2008).Stylistic and syntactic features were used for the purpose of classification. They used information gain and genetic algorithm for the purpose of feature selection in addition to the two feature sets. There exists two steps involved in system design, the first one includes the selection of the feature set and the other step is to perform feature selection. The system design is applicable to Arabic language as well other than the English. To achieve a better selection of features out of the given independent set pf variables, entropy weighted genetic algorithm has been used with the aim of achieving a high accuracy as was met in this case with 95.5

SVM classifier was trained via a twitter dataset in this paper (Zgheib & Barbar 2017). Tweets were retrieved by using hashtags that were used for the dataset. The dataset comprises of a total number of 9000 tweets of which 3000 tweets each for the three polarities that is positive, neutral and negative respectively, so the dataset was symmetrically distributed and well balanced. The polarity of tweets were interpreted using twitter hashtags. A random list was used by the author to determine the accuracy of proposed technique ,a random list of 100 tweets were taken of which the classifier could accurately classify upto 80 tweets thus an accuracy of 85

As observed from the comparative studies conducted on RNN and CNN when applied for natural language processing (Yin et al. 2017), RNN gave a much better result that CNN. But there also seems limitation of RNN which can be seen as the time sequence tends to grow, the weights grow beyond control or vanish in the case of RNN. To deal with the vanishing gradient problem (Hochreiter & Schmidhuber 1997) ] To deal with the vanishing gradient problem (Hochreiter et al. 2001) LSTM that is Long Short-Term Memory is proposed. For analyzing sentiments from short texts LSTM are widely used. Word embedding techniques are often used alongside LSTM and they are highly useful as they by vector representation of words tend to reduce the computational time .Word2vec(Mikolov et al. 2013) and Glove(Pennington et al. 2014) are amongst the most popular word embedding techniques The authors have used LSTM along with word2vec for the purpose of research.

For classification of text documents LDA is been used. Latent Dirichlet Allocation was used by the authores in paper (Tasci & Gungor 2009) to perform text categorization. The challenge of Dimensionality problem was dealt by fea-

ture selection method. Topic analysis was dealt with initially and after that the word correspondence was done and the model also incorporate labels in (Ramage et al. 2009).From the results it can be seen that LDA can give us with better results in comparison to SVM. To classify twitter messages and the corresponding users into their respective categories two different methods were proposed that are Latent Dirichlet model and author topic model in the paper in paper(Hong & Davison 2010). Text classification was performed and analyzed using different methods that are LDA¡ Latent semantic indexing and vector space model In paper (Liu et al. 2011).

# 3 Methodology

Knowledge Discovery Database: For the research, the methodology used is KDD Knowledge Discovery Database, where all the steps have been followed in series, the process of KDD starts with collection of Data and involves selection of features, preprocessing of data, its transformation, followed by understanding of classification, clustering or regression and then we apply the corresponding algorithm to it and then interpret the results obtained from it. Below all the steps that have been followed and the methodologies that have been applied and techniques that have been involved are explained.

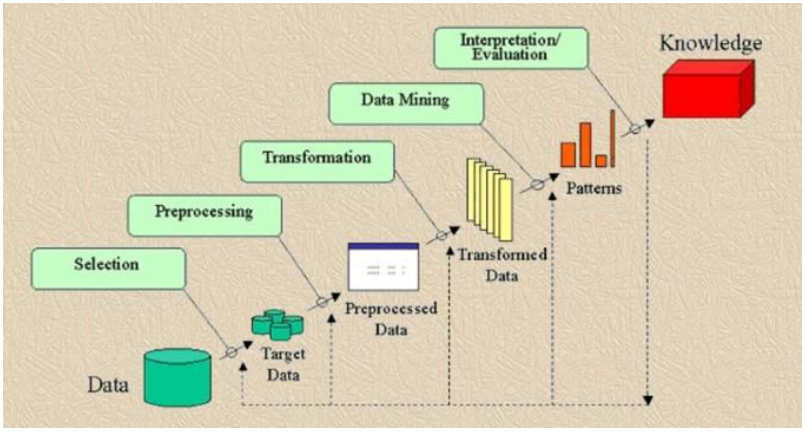



Figure 1: KDD Methodology

**Data Collection and Data Source**:

Data collection is the first step that involves understanding of area of research and collecting the data accordingly, for this research the topic of sentiment analy-

sis and topic modeling has been chosen as the research are and the data is collected accordingly. The dataset selected for the topic of research is Hotel Reviews and the source from which the data is collected is Kaggle which is an open source and the dataset is a public open dataset.

**Data Cleaning and EDA**:

It is essential to carry out data cleaning since raw data has a lot of anomalies, it consists of a lot of null values and even duplicates that need to be rectified before proceeding to any further step, raw data cannot be applied directly to any algorithm as no algorithm can process null values. For this research null values have been thoroughly checked, they have been treated by dropping the columns as only two columns had null value and that are not influential or significant towards the target. Exploratory data analytics is performed to understand the significance and influence of features that are present in the dataset, it gives an idea about selection of features that are of significance and what subset of features that are needed to be analyzed to carry or proceed with the research. Features that are found to not have much significance should be removed and must not be considered as it can lead to time complexity as well as computational errors as well. For this research since sentiment analysis and topic modelling is being performed the main features are the review text and their corresponding ratings as ratings are used to identify the validity of reviews. Review text is the main feature where the text is analyzed for sentiment analysis and topic modelling.

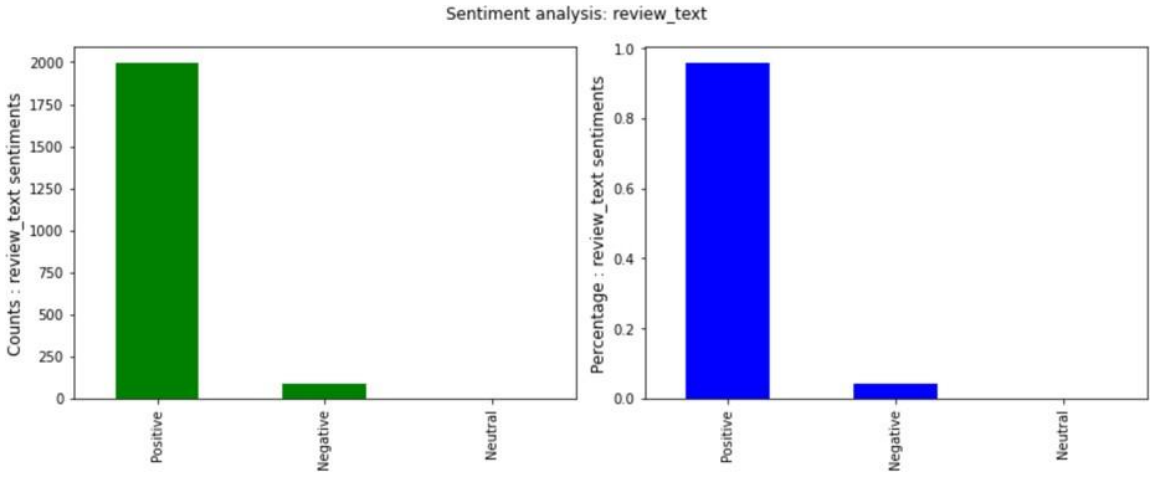



Figure 2: Review Count and percentage

**Data Pre-Processing**:

Data pre-processing is an initial step before feeding the data to the algorithm, this process also involves the removal of noise or reducing the dimensionality of subset of data been fed to algorithm. Noise present in dataset could be of the form

of emojis, punctuations, special characters etc. Lemmitization, stop word removal and stemming are the methods that are been used for removal of this noise, in this research a series of functions have been used for data pre processing and removal of noise. A total of nine functions have been used that are tokenize, remove stop word, remove$_{p}$*unct*
, lemmatizewords, stemwords, regexstringidentifier, removepattern, removee

# 4 Design

**4.1 Vector representation of data**:

Vector representation of words can be termed as the process of vectorising words in a way that words that have similar meaning are represented at closed distance on the word vector matrix , this is beneficial as it helps reduce the time complexity in a program as the words with similar meaning are close to one another so it makes the system more predictable and reduces the computational time for the system. Word vectorization can be done by various methods for this research work the methods chosen are word2vec and tfidf respectively.

**4.2 Word2Vec**:

Word2vec is a way to represent words spacially in a manner that words with similar meaning are closely represented on the word vector matrix. They are useful in case of NLP tasks and are widely used. They are part of the python library genism, they are mainly used to increase efficiency of a program and another advantage is that they are publically available. For this research genism library is imported and word2vec is used via function .models.word2vec

**4.3 TFIDF Vectorization**:

TFIDF is termed as term frequency – inverse document frequency that is used in machine learning and information retrieval that is used to quantify the relevance or the significance of representation of strings present in a document amongst a collection of documents also termed as a corpus. TFIDF vectorization can be useful when in case of information retrieval in tasks of NLP in cases of documents being present. For this research TFIDF vectorization is performed after splitting the dataset for hyper-parameterized tuning to input and response data, the function tfid vectorizer has been used.

**4.4 Hyper Parameterized Tuning**:

While we create any model in machine learning, we need to make design choices on which the model architecture is based. Sometimes we don't know what the choices need to be and we need to explore the possibilities. We are likely to let the machine perform this task and automatically find the optimal choices for the model architecture. Hyperparameters define the model architecture and hypermarameter tuning can be helpful in searching the ideal model architecture. For the machine learning models SVM and Random Forest hyperparameter tuning

is performed that help in building the model architecture.

**4.5 Random Forest**:

Random forest can be termed as an algorithm that is combination or a collection of many individual trees that together give a combined single output, this combination of trees to produce a single output results in a better accuracy compared to the individual outputs. Random Forest finds use in Sentiment Analysis for the following reasons: • The ability to give good results even in case of imbalanced data. • The ability to handle a high dimensional data that usually exists in case of text analysis. • Its simplicity also makes it a good choice.

**4.6 SVM**:

We can use Support Vector Machine SVM for both the problems that is classification and regression, it is a very common algorithm that is based on classifying distinctively the data points in N-dimensional space with respect to hyper plane. The following are the reasons for chosing SVM for Sentiment analysis: • High Dimensionality: SVM tends to be overfitting protected, they can handle large number of inputs. • Few irrelevant features: Since in text analysis the large number of features are important and significant an algorithm that can handle large number of features is preffered. • Document vector are sparse: SVM are well suited for sparse instances. • SVM can be useful to find the linear separators and most of the text categorization problems are linearly separable.

**4.7 LSTM**:

LSTM are unlike any other neural network, they tend to have feed forward connections. They can process data sequence rather that just processing the single points of data. They are widely used and ar quite common. They find good use in sentiment analysis due to the following reasons: • They are efficient and fast. • They can be applied to a large number of data, a dataset of significant size. • It can be useful in case of unstructured data. • IT can be applied in the case of long term dependencies.

**4.8 LDA**:

Latent Drichlet Allocation classify the documents by assigning them with topics such that assigned topics can capture the words in their respective documents. LDA is used in topic modelling and has the following advantages: • The biggest advantage is that there exists no need to know or predict initially what the topics may look like. • One can know document clustering and topic formation by tuning LDA to fit to different data shapes. • It measures in terms of probabilistic distribution for topics for occurrence of words.

**4.9 LSA**:

LSA is based on identifying and analyzing patterns in unstructured text and trying to retrieve information as well as relationships between them. • LSA is easy to understand and implement • LSA is much faster in comparison to other models as it uses only the term frequency decomposition matrix and nothing other than that.

# 5 Implementation

**5.1 Word Vectorization**:

• For implementing word vectorization word2vec has been used. • Genism library has been used for word2vec. • Genism.models.word2vec is the function been used for vectorization • For analyzing the data different word2vec functions such as .index to key and .most similar are used.

**5.2 TFIDF Vectorization**:

• Tfidfvectorizer function has been imported using the sklearn library.

• Vectorize function has been used in which the words have been transformed using tfidf fit • The data has been hyper-parameterized tuning after tfidf vectorization.

**5.3 Hyper-parameterized Tuning**:

• Hyper-parameterized tuning is performed after vectorization and splitting the dataset. • To define the parameters and use GridSearchCV() to perform tuning.

**5.4 Random Forest**:

• The function RandomForestClassifier has been imported from sklearn.ensemble library for performing classification using random forest.

• Random Forest has been applied after word2vec and tfidf vectorization.

• Cross value score is also analyzed and performing hyper parameterized tuning.

**5.5 SVM**:

• The function SVC is imported from the sklearn.svm library. • Support Vector Classification is applied after tfidf vectorization is performed,. • GridSearchCV and function SVC() are used for hyper-parameterized tuning of SVM.

**5.6 LSTM**:

• The library Keras has been used to import LSTM model in the program. • The maximum length of comments is kept in mind while defining the parameters of LSTM. • Label Encoder is used for the purpose of training the model on classifying he results. • Keras have been used, the function tokenizer to vectorize the words. • Finally the function sequential is used for building the Neural Network.

**5.7 LDA**:

• LDA has been implemented using the genism library. • For the purpose of Topic Modelling LDA is defined, where the text is converted to proper sentence and then to respective document before applying LDA on the corpus. • The perplexity of LDA and its coherence are used as a measure of identifying its performance.

5.8 LSA:

• TruncatedSVD is used to defint the Lsa model.

• LSA topic matrix is defined using the lsa model.fit transform for performing Topic Modelling.

• Get Top N words is used to interpret the top words for n=4 number of topics.

# 6 Result

Word2vec is the model used for word vectorization, the top 10 words present are shown above, it is clear from that and also can be seen by sentiment distribution that the positive reviews outweigh the negative also that can be verified by the rating as most of the rating is 5.

```
w2v.wv.index_to_key[:10]
```

```
['hotel',
 'room',
 'food',
 'pool',
 'good',
 'staff',
 'would',
 'time',
 'great',
 'day']
```

```
# Find the most similar words
w2v.wv.most_similar('amazing')
```

```
[('fantastic', 0.9992648363113403),
 ('brilliant', 0.9990436434745789),
 ('enjoyed', 0.9989259839057922),
 ('fabulous', 0.9988408088684082),
 ('everything', 0.9988296627998352),
 ('perfect', 0.998756468296051),
 ('especially', 0.9987457394599915),
 ('fault', 0.9986956715583801),
 ('best', 0.9986544251441956),
 ('job', 0.9986509680747986)]
```

Figure 3: Wor2vec

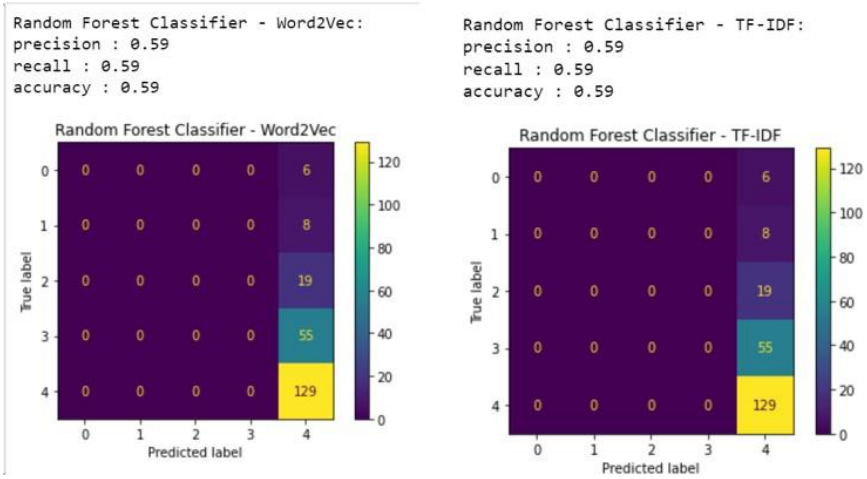



Figure 4: Random forest with word2vec and tfidf

Word2vec functions are used to observe the most similar words associated

with the positive word amazing, here w2v is the vector of size 100 applied on review text. Below are the outputs when tfidf and word2vec are applied to random forest classifier, it gives an accuracy of 0.59 and also the score is 0.59 in case of random forest.

```
# Instantiate and fit a basic Random Forest model on top of the vectors
rf = RandomForestClassifier(n_estimators = 5, criterion="log_loss", max_depth = 2,random_state = 32)
rf_model = rf.fit(X_train_vect_avg, y_sentiment_train.values.ravel())
rf_model.score(X_test_vect_avg, y_sentiment_test)
```

```
0.5944700460829493
```

```
              precision    recall  f1-score   support

           1       0.00      0.00      0.00         6
           2       0.00      0.00      0.00         8
           3       0.00      0.00      0.00        19
           4       0.00      0.00      0.00        55
           5       0.59      1.00      0.75       129

    accuracy                           0.59       217
   macro avg       0.12      0.20      0.15       217
weighted avg       0.35      0.59      0.44       217

[[  0   0   0   0   6]
 [  0   0   0   0   8]
 [  0   0   0   0  19]
 [  0   0   0   0  55]
 [  0   0   0   0 129]]
```

Figure 5: Random forest output

```
scores = cross_val_score(randomForest_model, X_sentiment_train_vec, y_sentiment_train.values.ravel(), cv = 5)
print(scores)
scores.mean()
```

```
[0.59195402 0.58959538 0.58959538 0.58959538 0.58959538]

0.5900671051757358
```

Figure 6: LSTM

Figure shows is the hyper parameterised output of random forest with the average of five cross value being as 0.59006. Only in the case of hyper-parameterized tuning there seems a difference in values that exists but is not that significant and its average being almost the same as tfidf and word2vec.

It can be seen that the output of Random Forest in all the cases is almost the same thus their seems to exist saturation point in terms of accuracy. For the future work one can work upon determining what could be the possible cause of saturation of accuracy.

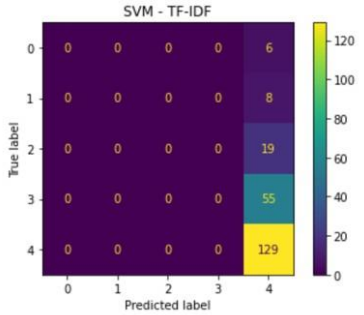
Figure shows the output of SVM produced after tfidf vectorization, it also gives the same accuracy of 0.59 which indicates that the model architecture needs further enhancement as their seems to exist a constant accuracy even after vectorization and hyper-parameterized tuning.

The model score is higher in the case of SVM, it gives a score of 0.63

Figure depicts building a neural network keeping in mind the maximum length of the review text, the use of keras is done to vectorize the words

Figure shows the output of last 10 of the 50 epochs chosen, loss and accuracy are measured in correspondence to the epoch. Also given below are diagrams that show variation of accuracy and loss with respect to the epochs that indicate that accuracy has saturated and become constant and remains unaffected with the increasing number of epochs, the loss fades after a few number of epochs and is also almost constant.

```
SVM - TF-IDF:
precision : 0.59
recall : 0.59
accuracy : 0.59
```




The model score is higher in the case of SVM, it gives a score of 0.63

```
score = model.score(X_sentiment_test_vec, y_sentiment_test)
score
```

```
0.631336405529954
```

Figure 7: SVM Output

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 693, 100)          571200

 lstm (LSTM)                 (None, 50)                30200

 dense (Dense)               (None, 5)                 255

=================================================================
Total params: 601,655
Trainable params: 601,655
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 8: LSTM built

```
Epoch 41/50
28/28 - 21s - loss: 1.1168 - accuracy: 0.5901 - val_loss: 1.1145 - val_accuracy: 0.5901 - 21s/epoch - 757ms/step
Epoch 42/50
28/28 - 21s - loss: 1.1150 - accuracy: 0.5901 - val_loss: 1.1140 - val_accuracy: 0.5901 - 21s/epoch - 755ms/step
Epoch 43/50
28/28 - 21s - loss: 1.1170 - accuracy: 0.5901 - val_loss: 1.1144 - val_accuracy: 0.5901 - 21s/epoch - 746ms/step
Epoch 44/50
28/28 - 21s - loss: 1.1162 - accuracy: 0.5901 - val_loss: 1.1140 - val_accuracy: 0.5901 - 21s/epoch - 743ms/step
Epoch 45/50
28/28 - 22s - loss: 1.1182 - accuracy: 0.5901 - val_loss: 1.1150 - val_accuracy: 0.5901 - 22s/epoch - 785ms/step
Epoch 46/50
28/28 - 21s - loss: 1.1167 - accuracy: 0.5901 - val_loss: 1.1140 - val_accuracy: 0.5901 - 21s/epoch - 751ms/step
Epoch 47/50
28/28 - 21s - loss: 1.1199 - accuracy: 0.5901 - val_loss: 1.1139 - val_accuracy: 0.5901 - 21s/epoch - 749ms/step
Epoch 48/50
28/28 - 21s - loss: 1.1174 - accuracy: 0.5901 - val_loss: 1.1141 - val_accuracy: 0.5901 - 21s/epoch - 750ms/step
Epoch 49/50
28/28 - 21s - loss: 1.1169 - accuracy: 0.5901 - val_loss: 1.1150 - val_accuracy: 0.5901 - 21s/epoch - 755ms/step
Epoch 50/50
28/28 - 21s - loss: 1.1168 - accuracy: 0.5901 - val_loss: 1.1146 - val_accuracy: 0.5901 - 21s/epoch - 760ms/step
```

Figure 9: LSTM epochs and output

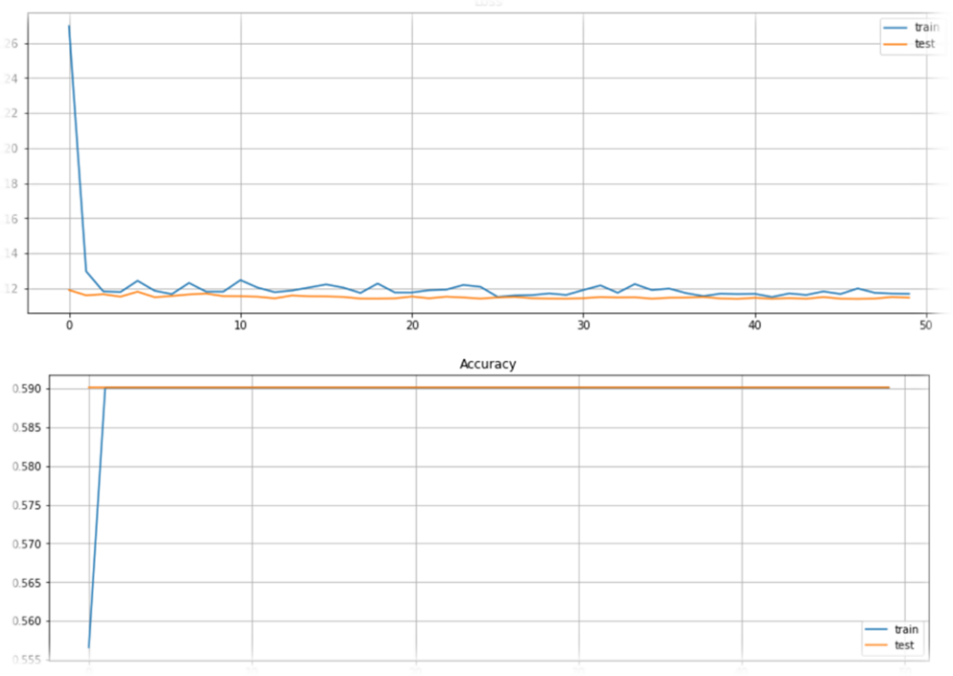



Figure 10: LSTM accuracy and loss v/s epoch

Figure 11: Word cloud

Given above is a word cloud indicating the most common positive words, given below is the output of LSA that shows a list of top 10 words and assigns topic to them, the topic count is 4.

```
Topic 0:  hotel room good great pool food staff would time restaurant
Topic 1:  hotel friendly back amazing everybody came staff holiday perfect friend
Topic 2:  pool unfortunately swimming enjoy teenage plenty also inclusive clean good
Topic 3:  hotel stayed star internet nice struggled located excellent carte fact
```

```
[['everything', 'marvelous', 'hotel', 'room', 'beautiful', 'garden', 'swimming', 'pool', 'beach', 'animation', 'team', 'great',
'hotel', 'cleanliness', 'everywhere', 'room', 'spacious', 'tidy', 'time', 'staff', 'friendly', 'enjoyed']]
```

Figure 12: LSA Topics

A corpus of documents has been created and the LDA model is applied on it,

```
[(0,
  '0.015*"get" + 0.013*"bar" + 0.012*"restaurant" + 0.010*"one" + 0.009*"day" '
  '+ 0.008*"pool" + 0.007*"main" + 0.006*"room" + 0.006*"area" + '
  '0.006*"buffet"'),
 (1,
  '0.032*"hotel" + 0.016*"room" + 0.014*"food" + 0.013*"good" + 0.012*"pool" + '
  '0.011*"staff" + 0.011*"would" + 0.011*"time" + 0.010*"great" + 0.009*"day"'),
 (2,
  '0.012*"said" + 0.011*"star" + 0.009*"rude" + 0.006*"told" + 0.006*"asked" + '
  '0.005*"system" + 0.005*"broken" + 0.005*"sleep" + 0.005*"lesson" + '
  '0.005*"terrible"'),
 (3,
  '0.011*"boat" + 0.006*"waterfall" + 0.006*"coast" + 0.006*"diver" + '
  '0.005*"sail" + 0.005*"dive" + 0.005*"ala" + 0.004*"regarding" + '
  '0.004*"titanic" + 0.004*"co"')]
```

below shows the output of the model in form of a list with its associated documents in this list of corpus.

Figure 13: LDA Corpus

The perplexity and coherence are measured, they indicate the fact of models performance that shows the model has performed well in terms of picking up the word since the perplexity is low.

```
Perplexity:  -7.061446913423162

Coherence Score:  0.36377641526810955
```

Figure14:LDAOutput

# 7 Conclusion

The research aims at Analyzing Sentiment and Modelling Topics using Deep Learning and Machine Learning models and compare their results. LSTM, Random Forest and SVM were applied for Analyzing Sentiments after vectorization of words was done, word2vec and TFIDF were used for the machine learning models that is SVM and Random Forest and Keras was used for LSTM. Hyper-parameterized was done for Random Forest and SVM. LDA and LSA were used for Topic Modelling. The accuracy of all the models is the same for analysis of sentiments that is 0.59, Random Forest on hyper parameterized tuning gave the highest accuracy of 0.591 for cv=5 but the average of the 5 cross value is 0.590 which is the same as other models gave on vectorization. SVM gave the highest model score of 0.63. Also for the case of LSTM 50 epochs were used and it was observed that the accuracy becomes almost constant after a few epochs and loss Is also almost constant after it decreases rapidly after a few epochs. SVM has edged the other models in terms of model score. Both LDA and LSA were able to classify the text into respective topics. LDA successfully classify the topics in the given corpus and it also gave a low perplexity which is a measure of good performance of the model and also coherence score is good for the model. For the future work one can try to work on the accuracy of models in case on Sentiment Analysis, the accuracy becomes constant after a few number of epochs and is the same for all the three models that is LSTM, Random Forest and SVM. One can try to analyze the cause behind the same, for the research vectorization and hyper parameterized tuning were used and still the accuracy became constant after a while so maybe one can work on trying to find the reason for it, one can look for data pre-processing or on tuning parameters or weather it is a reflection of unbalanced data and try to find the cause of why the accuracy becomes constant even after vectorization and hyper-parameterized tuning and try to find a way to increase the accuracy and fix the problem.

# References

Abbasi, A., Chen, H. & Salem, A. (2008), 'Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums', *ACM transactions on information systems (TOIS)* **26**(3), 1–34.

Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J. et al. (2001), 'Gradient

flow in recurrent nets: the difficulty of learning long-term dependencies'.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural computation* **9**(8), 1735–1780.

Hong, L. & Davison, B. D. (2010), Empirical study of topic modeling in twitter, *in* 'Proceedings of the first workshop on social media analytics', pp. 80–88.

Kim, S.-H. & Cho, H.-G. (2020), 'User–topic modeling for online community analysis', *Applied Sciences* **10**(10), 3388.

Lima, A. C. E., de Castro, L. N. & Corchado, J. M. (2015), 'A polarity analysis framework for twitter messages', *Applied Mathematics and Computation* **270**, 756–767.

Liu, Z., Li, M., Liu, Y. & Ponraj, M. (2011), Performance evaluation of latent dirichlet allocation in text mining, *in* '2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)', Vol. 4, IEEE, pp. 2695–2698.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013), 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781* .

Molinillo, S., Fernández-Morales, A., Ximénez-de Sandoval, J. L. & Coca-Stefaniak, A. (2016), 'Hotel assessment through social media–tripadvisor as a case study', *Tourism & Management Studies* **12**(1), 15–24.

Nguyen, T. H. & Shirai, K. (2015), Topic modeling based sentiment analysis on social media for stock market prediction, *in* 'Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)', pp. 1354–1364.

Pennington, J., Socher, R. & Manning, C. D. (2014), Glove: Global vectors for word representation, *in* 'Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)', pp. 1532–1543.

Pietsch, A.-S. & Lessmann, S. (2018), 'Topic modeling for analyzing open-ended survey responses', *Journal of Business Analytics* **1**(2), 93–116.

Prasanna, P. L. & Rao, D. R. (2019), 'A text mining research based on topic modeling using latent dritchlent allocation', *Int. J. Recent Technol. Eng* **7**(5), 308–317.

Putri, I. R. & Kusumaningrum, R. (2017), Latent dirichlet allocation (lda) for sentiment analysis toward tourism review in indonesia, *in* 'Journal of Physics: Conference Series', Vol. 801, IOP Publishing, p. 012073.

Rafea, A. & GabAllah, N. A. (2018), 'Topic detection approaches in identifying topics and events from arabic corpora', *Procedia computer science* **142**, 270–277.

Ramage, D., Hall, D., Nallapati, R. & Manning, C. D. (2009), Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora, *in* 'Proceedings of the 2009 conference on empirical methods in natural language processing', pp. 248–256.

Rosyidi, M. I. (2019), Indonesian online travel agencies: Profiling the services, employment, and users, *in* '3rd International Seminar on Tourism (ISOT 2018)', Atlantis Press, pp. 211–216.

Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Trevino, A. & Gordon, J. (2012), Empirical study of machine learning based approach for opinion mining in tweets, *in* 'Mexican international conference on Artificial intelligence', Springer, pp. 1–14.

Sutherland, I. & Kiatkawsin, K. (2020), 'Determinants of guest experience in airbnb: a topic modeling approach using lda', *Sustainability* **12**(8), 3402.

Tasci, S. & Gungor, T. (2009), Lda-based keyword selection in text categorization, *in* '2009 24th International Symposium on Computer and Information Sciences', IEEE, pp. 230–235.

Tian, X., He, W., Tao, R. & Akula, V. (2016), 'Mining online hotel reviews: a case study from hotels in china'.

Yin, W., Kann, K., Yu, M. & Schütze, H. (2017), 'Comparative study of cnn and rnn for natural language processing', *arXiv preprint arXiv:1702.01923* .

Yusliani, N., Primartha, R. & Marieska, M. D. (2019), 'Multiprocessing stemming: A case study of indonesian stemmi', *International Journal Computer and Applications (IJCA)* **182**(40), 15–19.

Zgheib, W. A. & Barbar, A. M. (2017), 'A study using support vector machines to classify the sentiments of tweets', *International Journal of Computer Applications* **975**, 8887.