## Contents

[1 Introduction 3](#_Toc9660)

[2 Related Work 4](#_Toc9661)

[3 Research Methodology 5](#_Toc9662)

[4 Expected outcomes 10](#_Toc9663)

[5 Time and Activity Plan 10](#_Toc9664)

[6 Risk Management 11](#_Toc9665)

# Introduction

Sentiment Analysis, also called as Opinion Mining is a field under Natural Language Processing, which tries to identify the opinion/expression of a person who mentions his/her intentions in a textual form. This method shows the polarity of the subject through the obtained text by transforming unstructured data into structured data. Since this field tries to connect human and computer intelligence, to detect and classify the opinions of the users, it is an active field of progressive research for some time now.

This is generally done by capturing the opinions of the users which may be in the form of reviews, suggestions, ratings, assorted as various expressions. The stakeholders, generally the management can use this data to analyze the user’s preference and give recommendations based on them. They also provide rankings and relevant data to the production houses. This in turn benefits them to attract more users.

However, the complex cases have not been focused enough on nor were they addressed properly in this field. The opinions that have been analyzed until now are either classified into positive, negative or neutral classes or classified based on a likert scale. These classifications do not capture the emotions of the users as originally intended nor give a sense of ambiguous sentences. This will restrict the analysis to generalize the opinions within the defined limits [4].

This problem is urgent and relevant to address, as this gives a better categorization within the opinions of the users. This helps the stakeholders to inspect the users opinion about a certain genre of movies. The analysis is reflected in the recommendations given to them, to keep the user engaged.

Hence, introduction of classes that describe the user emotions by choosing the most general ones from a wide array of emotions available, needs to be done to identify the original meaning of the user. The impact on accuracy when such modification is introduced to the cBLSTM classifier implemented by (Mousa et al., 2017) [6] is observed, by making a comparison between this method and the modified method, which is the context of the problem.

cBLSTM(Contextual bidirectional long short term memory) classifier is a variation of RNN (recurral neural network) that checks the context in both the directions, extracting features from the given input, which can be a part of the dataset given as sequential batches, or the entire dataset itself. This classifier is chosen for this study, to determine its accuracy when emotions are used to categorize the user opinions. Comparisons are drawn with the original cBLSTM classifier that works under the conditions set by (Mousa et al., 2017) [6] to find the same.

CURRENNT Toolkit introduced by Weninger et al. (2014) is used for the implementation of uni/ bidirectional RNNs like cBLSTM classifiers and its variants. It categorizes a given dataset into respective labels and also provides flexibility to introduce own guidelines for this categorization. The toolkit was used for implementation of the study (Mousa et al., 2017) [7].

The purpose of this study is to find how the cBLSM classifier performs when the dataset chosen is labelled as emotions, rather than generalized sense of opinions and how accurate will it be in identifying those emotions.

The objective of the study is to replicate the entire process in (Mousa et al., 2017), with a change introduced in the categorization of the dataset, and to determine how this change affects the process in terms of accuracy.

The actions performed to achieve the said objective are two-fold, first being the task of separating and processing the training data into the classes labelled and the next task is to train the classifier to do the same and test it to determine how accurate it is, for the dataset chosen. Likely implication of this study is to increase accuracy of the model, by introducing the proposed changes.

# Related Work

Sentiment analysis is an emerging field in the context of natural language processing in textual form, raising several questions relevant to its implementation. Wide range of implementations were successful due to classifiers like Support vector machines,Naive Bayes, Max Entropy (Amit gupte et al.,2014) [2]even in terms of feature space. Models like Unigrams, Bigrams, ngrams have been used on their own in some cases while in others, as a conjuncture with Parts of Speech(POS) model which was discussed by (Zhou et al.,2015) [8].

Other standardized classifiers like the ones mentioned in (Amit gupte et al.,2014) [2] which proved to be efficient in many aspects were also used. Turns out that even though Random Forest classifier has high accuracy and performance, it requires high training time and processing power. Due to the assumptions over the data distributed and the constraints on power and memory, Naive Bayes classifier is not a preferred model. Max Entropy (MaxEnt) can accommodate less training time, but as it requires powerful processing system and memory, this was not chosen either [8].

(Zhou et al.,2015; Amit gupte et al.,2014) [2] [8] studies highlight that neural network models are used to process the input either as a word sequence or as a syntactic parse tree and Convolutional Neural Network (CNN) as well as Recurrent Neural Network (RNN) are mostly used for this implementation . Another study by (mousa et al.,2017) [6] reveals that cBLSTM, a hybrid implementation of RNN, outperforms the combination of Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BLSTM) binary classifiers which used linear interpolation of probabilities.

Hence, certain elements of this study are chosen to be replicated like subjects, objects, instrumentation and dataset to test the classifier’s accuracy when the categories of the data on which training is done are changed.

However, all these studies make their point by using labelled/annotated data comprising of two or three generalized classes [1] and as a consequence, their classification of analysis is either positive or negative and in some cases neutral is also included. Very few approaches classified their analysis into more than three classes. We find this way of classification to narrow down the expression of the reviewer, as relative degree is not considered.

(Kang et al.,2016) [3] discusses about predictive algorithms that detect the emotions by using a Hierarchical Bayesian Model approach, where the labelled classes are termed after an extensive semantic exploration to determine the emotions in a blog. In the same lines, (Meo et al.,2017) [5] identified the emotions by drawing comparisons between different methods using models such as SVM, Naive Bayes, Random Forest. These emotions recognized are labelled into 7 different classes. This method conveyed the opinion of the user in a better way [5].

As these papers support the use of classes labelled as emotions, to classify the user opinion, instead of using generalized categorization on likert scales, this method is introduced as a modification in the study being replicated, to determine its effect on accuracy of the classifier.

Implementation of conscious refinement of the dataset to a detail where the original meaning is not changed or obscured while cleaning the data and usage of an emoticon dictionary to get an equivalent meaning of the emoticons specified has been done, which was proposed by (Meo et al.,2017) [5].

# Research Methodology

RQ-1: How is the dataset pre - processed before the cBLSTM classifier categorizes it into the labelled classes ?

RQ-2: How are the results of the classifier analyzed to evaluate the accuracy?

The method of research conducted for this study is an Experiment. Experiment provides means to make meaningful comparisons between existing methods and new methods, by providing control over the factors involved along with flexibility in the entire process . Hence comparison is drawn between the original method with modifications in the categorization of dataset to be classified and the original classifier itself. This method also grants flexibility over the choice of subjects, objects and instrumentation.

Furthermore, experiments provide an opportunity for replication which is crucial in such implementations, to evaluate and to improve a process under varied settings.

Cause – Effect relationship can be studied in this experiment, where the cause is the change in the choice of the number and categorization of the labelled classes, and the effect is the change in accuracy of the performance of the classifier.

We aim to analyze and classify the reviews from IMDB data-set into seven different classes/labels by processing them using cBLSTM classifier. For this, supervised learning is followed, by labelling the data-set into those classes and using these labelled data for training and testing purposes. Stratified sampling strategy is followed, considering these labelled categories as strata for random sampling.

All of these measures can be followed under a controlled setting, which is not feasible in a case study. Comparison cannot be drawn in a case study setting, which is the key treatment proposed, to draw a conclusion about the accuracy when modifications are induced in the cBLSTM classifier.

## Study Design

In this experiment, since the research questions are related, addressing consecutive parts of the solution, they share the same subjects, objects, variables, instrumentation, population and sampling strategy. However, separate treatments are proposed to solve each research question, as mentioned below.

Objects of our experiment are reviews,to capture the sense of feelings through the opinions given by the user.

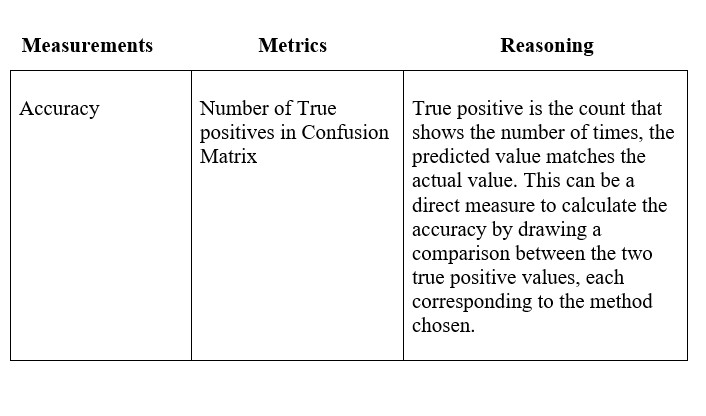
Reviewers are the subjects in this experiment. We use IMDB data-set that consists of movie reviews generated by these reviews for this study and we aim to classify user experience based on the user ratings they provide.

Variables:

The independent variables considered in this experiment are the classes, size of the data set and the dependent variable is accuracy.

Instrumentation: cBLSTM classifier.

Experiment Measures:



Population: IMDB data-set.

Sampling strategy:

Stratified sampling strategy is performed on the population after splitting the pre - processed data for training, testing and validation. The ratio chosen for training and testing data is 6:4, i.e., 60% of the data is used to train the classifier, while the remaining 40% of the data is used for testing and validation purposes.

Treatment:

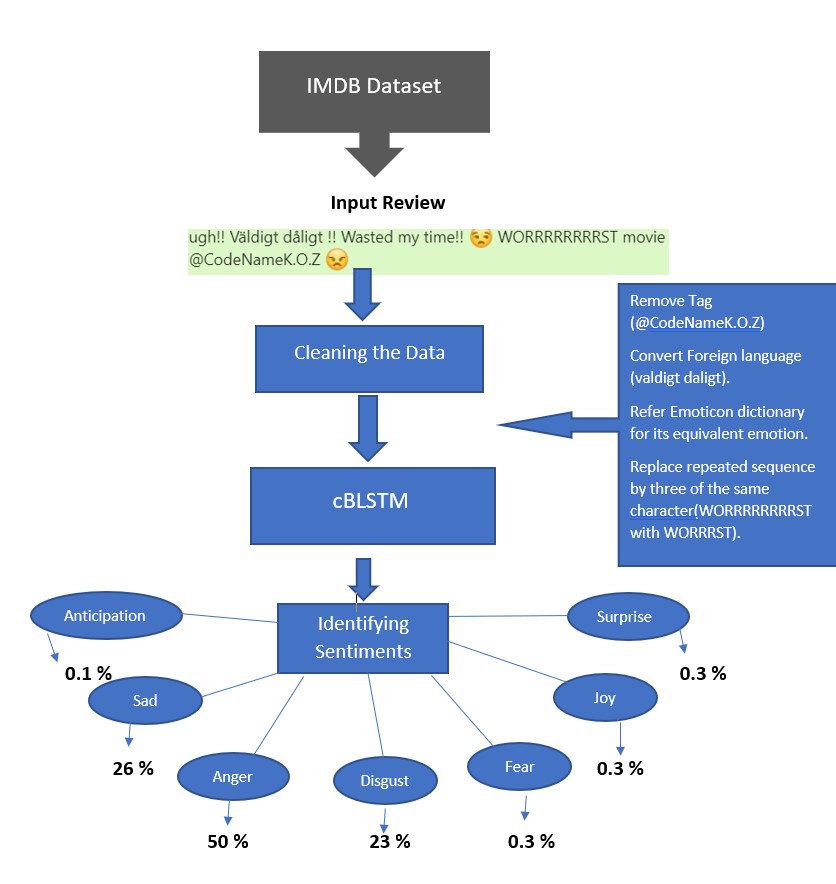
This experiment uses IMDB data set which is annotated by using CURRENNT toolkit and labelled under 7 different classes namely: joy, surprise, anticipation, fear, sad, disgust, anger.

RQ1:

The reviews are pre - processed and cleaned to preserve the original meaning. This can be done as follows:

* Using annotated dictionaries for emoticons, acronyms and stop words to map emoticons to their polarities [1].
* English translations over frequently used acronyms by using WordNet [1].
* Converting reviews in foreign languages to English by using Google translate. This eliminates language or location restrictions when mining the data [1].
* URLs and targets("@") replaced with tags [1].
* A sequence of repeated characters replaced by three of the same characters to consider the difference between regular usage and emphasized usage of the words [1].
* Tokenizing each text sentence to record their occurrences [1].
* Combining Word popularity and WordNet dictionary to determine pleasantness of each word. The scores are normalized to give a relative sense to each word [1].

For example, a review from IMDB dataset is pre - processed as:



These manipulations are implemented to clean the data for removing ambiguity and to give a direct sense of intended emotions. 60% of this refined data is used for training the cBLSTM classifier as intended and for this, the data is categorized into labels using the CURRENNT toolkit.

RQ2:

The cBLSTM classifier is trained using supervised learning method, where the classifier is taught to determine which label should be assigned to the new data, by training it to identify the similarities, if any. CURRENNT toolkit is used to categorize data under labelled classes for this purpose, and this data is fed as an input to this classifier.

A One factor, two treatment method of design is chosen to calculate the accuracy of this classifier. The factor being cBLSTM classifier, and the two treatments chosen for comparison are the cBLSTM classifier, from the paper written by (Mousa et, al., 2017) and the replication of instrumentation, design, variables and subjects of that study while incorporating a modification in the way the data is processed and classified.

Relative measure is calculated by cBLSTM for each review as the reviews may have mixed feelings and by attempting to capture the degree of relativity, opinions of the users are identified more clearly. This modification is implemented in the newer method to increase its accuracy. The comparison is made for checking if the incorporated change i.e., categorization of the labelled classes as emotions gives better accuracy and over the scenario of generalizing the opinion on a comparative scale.

Hypotheses formulated to compare the old and new methods are:

Null hypothesis:

*Ho* : *µANew <*= *µAOld*

Alternate hypothesis:

*H*1 : *µANew >µAOld*

where *µ* denotes the average, A denotes accuracy.

Null Hypotheses state that the accuracy of the new method to identify the user opinion is the same or lesser than that of the old method and the alternative hypotheses disprove it by stating that, the accuracy is more when new method is used.

The outcome of the experiment can be analyzed by drawing a comparison between the confusion matrix of each method. A confusion matrix is a matrix displaying false positives, false negatives, true positives and true negatives, generated for the given dataset. It is shown as:

Prediction outcome

p n total

|  |  |  |
| --- | --- | --- |
| True  Positive |  | False  Negative |

p0P0

|  |  |  |
| --- | --- | --- |
| False  Positive |  | True  Negative |

actual value

n0N0

total P N

The values from the two confusion matrices that correspond to each model can be compared consecutively, to draw a conclusion about the accuracy after the modification is introduced with respect to categorization of the dataset.

Validity threats

Internal Validity:

Over - fitting is observed when a model/classifier does not reflect its knowledge gained on the testing data. This is observed through reduced accuracy when testing data is run on the classifier. To avoid this, regularization can be done to the classifier, by using a dropout in the network.

Construct Validity:

level of construct - the size of the dataset could be a factor to effect the level of the construct i.e, if a larger dataset is used for training the classifier, better accuracy can be achieved over the current size of the dataset.

Conclusion Validity:

since the toolkit is subjected to periodic updates, this might effect the categorization of the dataset because of which there could be a difference in accuracy and by extension the outcome of the experiment.

External validity :

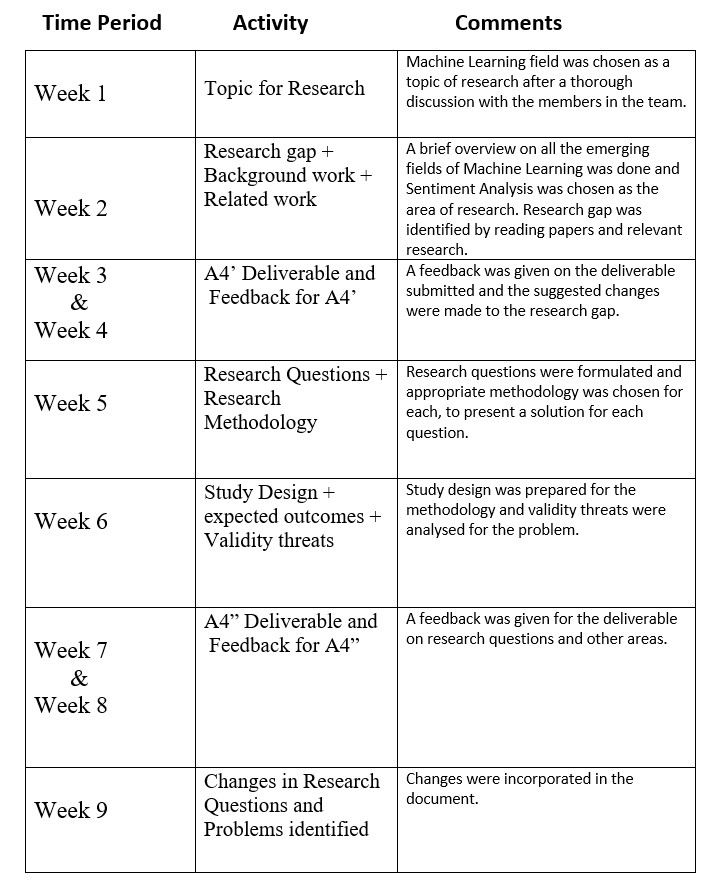
As the IMDB Dataset chosen falls under recreation and entertainment genre, the classes were labelled to capture the emotions of the user relevant to this category. However these results cannot be generalized to all the data sets, but are still relevant to analyze the datasets like BookMyShow that fall under similar categories.

# Expected outcomes

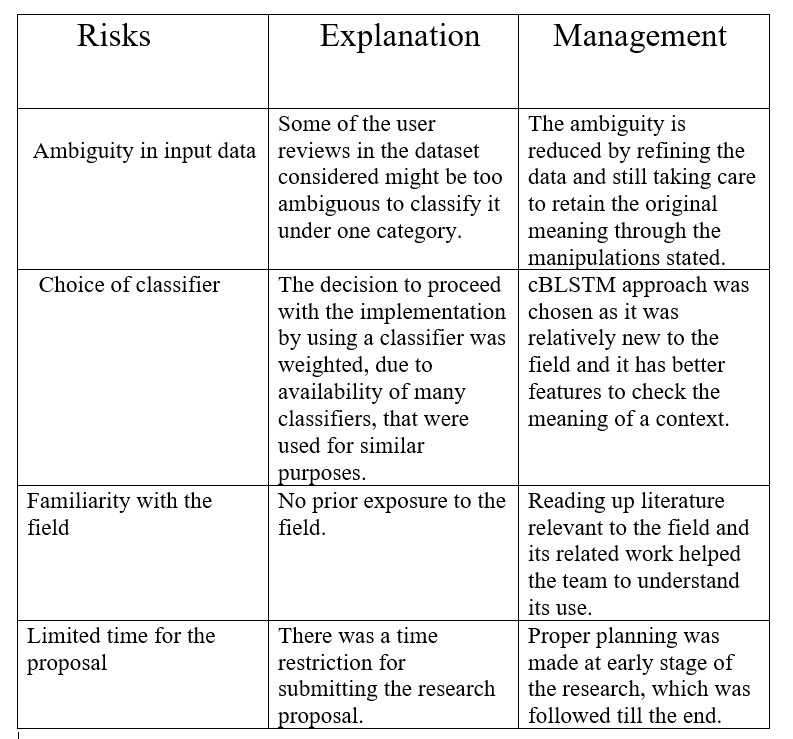
RQ1: The refined data, that is categorized into the labelled classes for training the classifier is obtained.

RQ2: The method which resulted in more accuracy is determined at the end of the comparison.

# Time and Activity Plan



# Risk Management



## References

1. Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, 2011.
2. Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam, and A Gupte. Comparative study of classification algorithms used in sentiment analysis. *International Journal of Computer Science and Information Technologies*, 5(5):6261–6264, 2014.
3. Xin Kang and Fuji Ren. Understanding blog author’s emotions with hierarchical bayesian models. In *2016 IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC)*, pages 1–6. IEEE, 2016.
4. PK Kumar and S Nandagopalan. Insights to problems, research trend and progress in techniques of sentiment analysis. *International Journal of Electrical and Computer Engineering*, 7(5):2818, 2017.
5. Rosa Meo and Emilio Sulis. Processing affect in social media: a comparison of methods to distinguish emotions in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(1):7, 2017.
6. Amr Mousa and Björn Schuller. Contextual bidirectional long short-term memory recurrent neural network language models: A generative approach to sentiment analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 1023–1032, 2017.
7. Felix Weninger, Johannes Bergmann, and Björn Schuller. Introducing currennt: The munich open-source cuda recurrent neural network toolkit. *The Journal of Machine Learning Research*, 16(1):547–551, 2015.
8. Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*, 2015.