Komal D. Anadkat, Dr. Hiteisha M. Diwanji , Dr. Shahid Modasiya
EFFECT OF PREPROCESSING IN HUMAN EMOTION ANALYSIS
USING SOCIAL MEDIA STATUS DATASET

RT&A, No 1 (67)
Volume 17, March 2022

# EFFECT OF PREPROCESSING IN HUMAN EMOTION ANALYSIS USING SOCIAL MEDIA STATUS DATASET

Komal Anadkat[1], Dr. Hiteishi Diwanji[2], Dr. Shahid Modasiya[3]

•

[1]Ph.D. Scholar, Information Technology department, Gujarat Technological University, Ahmedabad, India. komalanadkat@gecg28.ac.in

[2] Associate Professor, Information Technology department, L.D. College of Engineering, Ahmedabad, India. hiteishi.diwanji@gmail.com

[3]Assistant Professor, Electronics & Communication department, G.E.C., Gandhinagar, Gandhinagar, India. shahid@gecg28.ac.in

## Abstract

*Emotion analysis using social media text is the emerging research area now a day. It helps the researcher to recognize the emotional state of the users and identify mental health-relevant problems like depression or anxiety, which may lead to suicide if not cured. The social media platforms like WhatsApp, Facebook, Instagram, etc. are widely used as these applications provide an affordable and reliable medium for transferring data, sharing thoughts, and even for routine informal communication. Social media status is normally analyzed to recognize the mood, emotion, thought process, or mental state of the individual as people generally share status for what they feel. On the other hand, pre-processing is the crucial step for any kind of text data analysis. In this paper, the social media status dataset is first pre-processed using various methods, given for feature extraction and classification purpose. For the machine learning approach, we have used count vectors and TF-IDF techniques for extracting the different features of the data. Using count vector feature extraction accuracy achieved by pre-processed data is 68.90%, 69.33%, 70.59%, 64.95%, 69.33% for naïve Bayes, LDA, Random forest, SGD and MLP respectively. Similarly, using TF-IDF feature extraction accuracy achieved by pre-processed data is 65.76%, 69.96%, 68.49%, 65.96%, 70.80% for naïve Bayes, LDA, Random forest, SGD and MLP respectively. The experimental results show that pre-processing helps to improve the accuracy of the classifier and CNN outperforms the traditional approach and achieves 79% accuracy.*

**Keywords:** Pre-processing, Emotion analysis, Feature extraction, classifiers, Count Vectors, TF-IFD

## I. Introduction

Emotions are defined as psychological conditions caused by neurophysiological changes, associated with thoughts and feelings of happiness or dislike. The advent of internet technology and communication makes them a great platform for people to express their feelings and ideas. If a person is having positive emotions like happiness, it can help to improve the health and work efficiency of a person [1]. On the other side, constant negative emotions like sadness and anger can cause major physical or mental disorders Persistent negative emotions can cause psychological issues like depression, which in some cases can lead to suicide. Nowadays people are using social media platforms to share their moods and feelings. Many popular applications like Twitter, Whats App, Facebook etc. are not only used to share static forwarded messages or multimedia contents but

Komal D. Anadkat, Dr. Hiteisha M. Diwanji , Dr. Shahid Modasiya
EFFECT OF PREPROCESSING IN HUMAN EMOTION ANALYSIS
USING SOCIAL MEDIA STATUS DATASET

RT&A, No 1 (67)
Volume 17, March 2022

also used to express feelings and emotions. To provide such insight, machine learning strategies may provide some unique features that can help explore unique patterns hidden in online communication and process them to reveal emotions like happy, sad, anger, surprise, or disgust. In addition, there is a growing body of literature on the role of social media concerning the structure of social networks such as broken relationships, mental illness smoking and drinking, sexual abuse, and suicidal thoughts among users of social media.

Many social media applications like WhatsApp, Facebook, Instagram etc. has a feature called status or story. The WhatsApp messenger started in 2009 with the purpose of instant communication and transfer of information. People use to share an image, video, text, GIFs in the status of WhatsApp which disappears automatically after 24 hours of uploading. To predict the mood or emotion of the person, WhatsApp status or Facebook story data can be very useful as people use it to post on the status what they usually feel. WhatsApp users are very careful about what they post as their status because their contacts will make meanings from their posts. People are sensitive to WhatsApp profile status; this implies there is a lot of significance attached to it. This phenomenon shows that users are sensitive, alert, and mindful of WhatsApp profile statuses.

In this paper, we have observed that social media status can be used to classify emotions effectively. For extracting meaningful information from the textual status data, the essential step is data preprocessing [5][2]. We used count vectors and TF-IDF feature extraction methods and report experimental results with different machine learning classifiers with the effect of different preprocessing methods. At last, we demonstrated the results of CNN on the same dataset. Our main contributions are, first we have developed a system that could find the emotion label like happy, sad, and anger for any piece of text, especially status and stories on social media with their probability for each emotion using its textual features, Preprocessing techniques, and various machine learning and deep learning algorithms. Second, we were able to achieve an accuracy of around 79% using a CNN classifier with a dataset available on Kaggle and last we Show the effect of pre-processing techniques on the final accuracy of the model.

## II. Existing Research

Since the last decade, many researchers are working in the area of emotion analysis and classification using social media datasets. The main goal of emotion classification is to identify the specific text and decide the relevant emotion in a different category.

In [3], the authors have proposed a novel hybrid approach to determine the sentiment hidden in the individual tweet. They have also shown the effectiveness of the pre-processing using different methods like slang/abbreviations and lemmatization and correct and stop words removal. In [4], the authors conducted an extended comparison of sentiment polarity classification methods for Twitter text. Furthermore, they proceeded to the inclusion of a combination of classifiers in the compared set and the aggregation and use of several manually annotated tweets for the evaluation of the methods. In [9], the authors have worked on the Spanish language for the mobile phone and Mexican presidential elections relevant tweets .they proposed parameter settings for SVM, naïve bays, and decision trees. They used n-gram as a feature extraction method and presented results for positive, negative, and neutral classes. Wiebe et al. [5] has manually annotated the emotions, opinions, and sentiments of a sentence, which are taken from news articles. Sana Shahid [7], the author has analyzed the popular social media platform WhatsApp and concluded that it helps in building aimed to explore that WhatsApp helps in building the interpersonal relationship and used by Professionals for meeting business goals .he also analyzed the usage pattern and kind of conversation users made by observing the frequency and composition of users. Church, K. et al [6], have compared SMS to MIM and give the reason for the success of WhatsApp. They conclude that

Komal D. Anadkat, Dr. Hiteisha M. Diwanji , Dr. Shahid Modasiya
EFFECT OF PREPROCESSING IN HUMAN EMOTION ANALYSIS
USING SOCIAL MEDIA STATUS DATASET

RT&A, No 1 (67)
Volume 17, March 2022

SMS is more private, formal, and more reliable and WhatsApp messages are more social, informal, and conversational. Yeboah and Ewur [8], authors have analyzed the performance of students in Ghana and conclude that WhatsApp can help in making communication easy and fast. In [11], The Lexicon-based methods have experimented for the instant messaging system as it uses sentiment expression keywords from the corpus and uses this to classify the emotions in different categories. The proposed method first identifies the emotion-relevant content in the text using keywords. After this to detect the emotional meaning of the text, they extracted syntactic features. Other functions like gestures, postures can be used to help the users to communicate with individual users. Aman et al. [12] have proposed a novel classification method that is based on emotional intensity knowledge. The author has achieved around 66% of accuracy rate for the blog corpus emotion classification task. Alm et al. [13], proposed an SVM-based text sentiment prediction method combined with SNoW (sparse network of winnows) architecture. In this approach, 22 fairy tales of Grimm's are taken as the input dataset. These fairy tales are divided into six emotion categories named happy, sad, fearful, angry, disgusted, surprised by selecting 30 features such as the first sentence of the story and specific connectives. Then the author has compared the result of the proposed method SVM + SNoW with existing well-known approaches and achieved 69.37% accuracy.

However, many researchers have done lots of work in the field of emotion classification and sentiment analysis using various social media datasets. But after a thorough investigation of the related literature, we came up with the conclusion that still there is a scope in the field of emotion classification using social media status. Hence, we present the effect of pre-processing in emotion analysis, and a report on experiment results demonstrating that feature selection and representation can affect the classification performance positively.

## II. Proposed Methodology

The ultimate goal of this paper is to evaluate the effect of pre-processing methods for emotion classification for social media status datasets. Several pre-processing techniques are available but we have applied some of them like cleaning Html tags, converting to lower, removing special characters and stopping words, and performing stemming and evaluating the effectiveness of these methods for different classifiers. Figure. 1 shows the flow for analyzing the emotions using social media status. In this paper, Emotion analysis has been performed by two different approaches. One is using the traditional machine learning methods, which extract the hand-crafted features from the data using TF-IDF and Count Vector methods. The other is using the deep learning methods, which don't require manual extraction of features.

### I. Dataset Description

The input dataset is available on Kaggle (https://www.kaggle.com/sankha1998/emotion). It has social media status collected by scraping from different sources. The dataset contains a textual status for three basic emotions happy, sad, and angry. The .csv file for each emotion category has two columns for status and relevant sentiment respectively.

### II. Preprocessing Techniques

Preprocessing is the basic mandatory step for any kind of classification problem. It tries to clean the data by removing noise and preparing suitable data for the machine-learning model. We have applied the following methods to convert raw data into a suitable form.
- Cleaning Html tag: - Unstructured text contains a lot of noise, especially here, we use

Komal D. Anadkat, Dr. Hiteisha M. Diwanji , Dr. Shahid Modasiya
EFFECT OF PREPROCESSING IN HUMAN EMOTION ANALYSIS
USING SOCIAL MEDIA STATUS DATASET

RT&A, No 1 (67)
Volume 17, March 2022

techniques like web or screen scraping.

- Converting to lower: - Converting all data to lowercase helps in the process of pre-processing.
- Remove special characters:-Special characters and symbols are usually non-alphanumeric characters, which add to the extra noise in unstructured text.
- Remove stop words: - Words that have little or no significance, especially when constructing meaningful features from the text, are known as stop words. Typically, these can be articles, conjunctions, prepositions, and so on. Some examples of stop words are a, an, the, and the like.
- Perform stemming: - Word stems are also known as the base form of a word, and we can Create new words by attaching affixes to them in a process known as inflection. The reverse process of obtaining the base form of a word from its inflected form is known as stemming. Stemming helps us in standardizing words to their base or root stem, irrespective of their inflections.
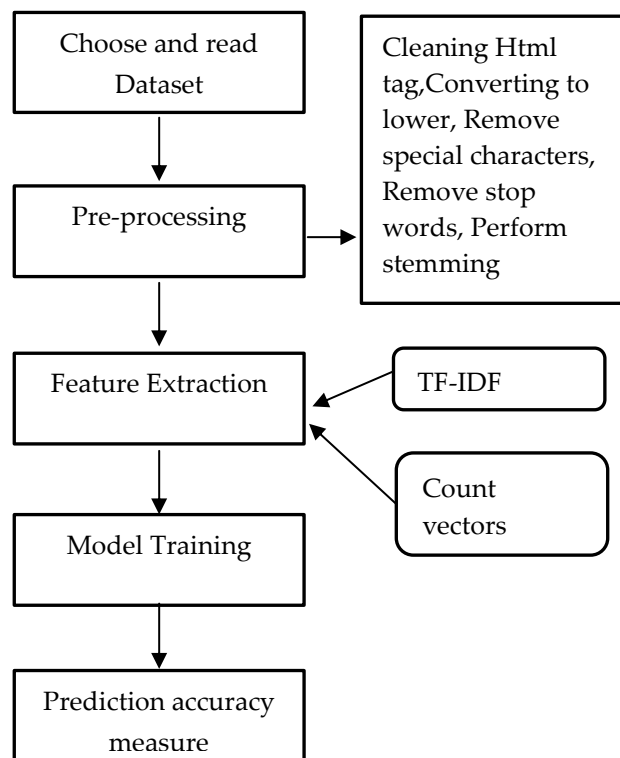


**Figure 1:** *Work Flow basic of emotion analysis*

## III. Feature Extraction

The preprocessed data is high-dimensional and unstructured. So different data is structured in different ways, which makes it difficult for the different classifiers to work on these data. The process to reduce the dimensionality of these kinds of data and extract the useful features from the text is known as feature extraction.

- TF-IDF: - Term frequency and inverse document frequency method is normally used for information retrieval from text purposes. It shows the relevance of particular tern is in the

Komal D. Anadkat, Dr. Hiteisha M. Diwanji , Dr. Shahid Modasiya
EFFECT OF PREPROCESSING IN HUMAN EMOTION ANALYSIS
USING SOCIAL MEDIA STATUS DATASET

RT&A, No 1 (67)
Volume 17, March 2022

document.[10] It converts simple words into vectors and assigns weights to not relevant words. We have used it as a feature extraction method to determine which terms in a document are most distinguishing for that document.

- Count Vector:-This method counts the number of occurrences of each word that appears in the particular document and makes a vector to store the result of it.

## IV. Classifiers

The ultimate goal for emotion recognition is to classify the input into relevant emotion classes like happy, sad, or anger. Many classification models are available for this purpose. We have chosen some well-known classifiers, named Naïve Bayes, Linear Discriminant Analysis (LDA), Random Forest (RF), Multi-Layer Perceptron, SGD, and CNN to evaluate their performance depending on the preprocessing method applied to the input dataset.

Each classifier has its own merits and demerits, so as per the input data and require the appropriate classifier can be chosen. The naïve Bayes algorithms are widely used for text classification purpose because it is efficient even though the dataset size is less and resources are scarce. LDA works well for the linear classification of data as it has good generalization ability. Random forest and MLP work well if data is high-dimensional. CNN is the updated version of a simple neural network, as it tries to make the network less complex by sharing weights and making a local connection. As shown in figure:-2, we have sequentially implemented the deep model, layer by layer. To add the layers we have used add () function. In the out model, we have a total of three dense layers and two Dropout layers as shown in the figure:-2 Activation function has a dominant effect on the performance of the network. We have used ReLu in our network as it works normally effectively with the neural network. Finally, we compiled our model by using 'adam' as out optimizer, 'accuracy' as a metric, and 'categorical_crossentropy' for our loss function.
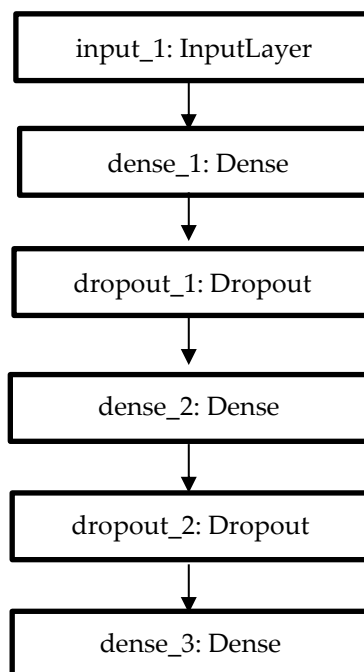
```
input_1: InputLayer
        ↓
dense_1: Dense
        ↓
dropout_1: Dropout
        ↓
dense_2: Dense
        ↓
dropout_2: Dropout
        ↓
dense_3: Dense
```

**Figure 2:** *Deep architecture for emotion classification*

Komal D. Anadkat, Dr. Hiteisha M. Diwanji , Dr. Shahid Modasiya
EFFECT OF PREPROCESSING IN HUMAN EMOTION ANALYSIS
USING SOCIAL MEDIA STATUS DATASET

RT&A, No 1 (67)
Volume 17, March 2022

## IV. Experimental Results and Discussion

We have performed various Pre-processing techniques on the dataset as mentioned in the previous section and will show the results in this section. We have tried Count vector and TF-IDF feature extraction methods and will show the results. Then, will show the results obtained from deep neural networks.

## I. Count Vector Feature Experiments

First, we have used Count Vector Features with different classifiers named Naïve Bayes, LDA, Random Forest, Multilayer Perceptron, and SGD. As shown in the table:-1 the result shows that maximum accuracy can be achieved by using a random forest classifier which is 70.59% by applying pre-processing which is 4% higher than without applying pre-processing techniques.

**Table 1:** *Performance Comparison of Count vector features*

| Classifier | Accuracy(%) without Pre-processing | Accuracy(%) with Pre-processed data |
|---|---|---|
| Naïve Bayes | 66.38 | 68.90 |
| LDA | 67.23 | 69.33 |
| RF | 66.59 | 70.59 |
| MLP | 63.27 | 64.95 |
| SGD | 65.75 | 69.33 |
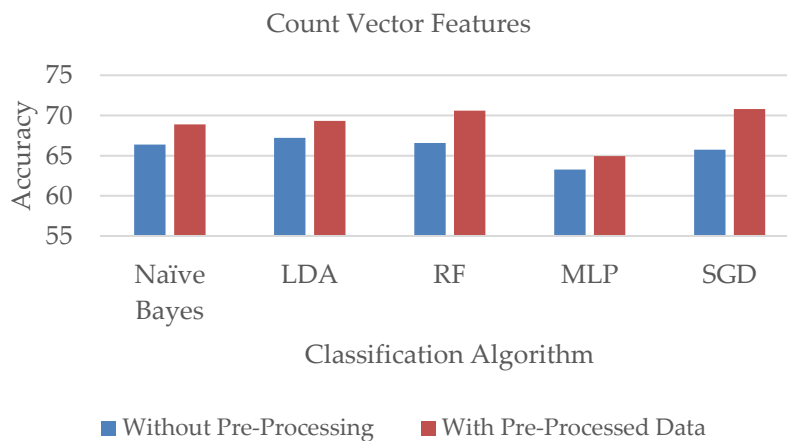


## II. TF-IDF Feature Experiments

Now, we tried TF-IDF Features with different classifiers named Naïve Bayes, LDA, Random Forest, Multilayer Perceptron, and SGD. As shown in the table:-2 The result shows that maximum accuracy can be achieved by using SGD which is 70.80% by applying pre-processing which is 5% higher than without applying pre-processing techniques. Here, Random forest gives good results without pre-processing which is 67.43%.

Komal D. Anadkat, Dr. Hiteisha M. Diwanji , Dr. Shahid Modasiya
EFFECT OF PREPROCESSING IN HUMAN EMOTION ANALYSIS
USING SOCIAL MEDIA STATUS DATASET

RT&A, No 1 (67)
Volume 17, March 2022

**Table 2:** *Performance Comparison of TF-IDF features*

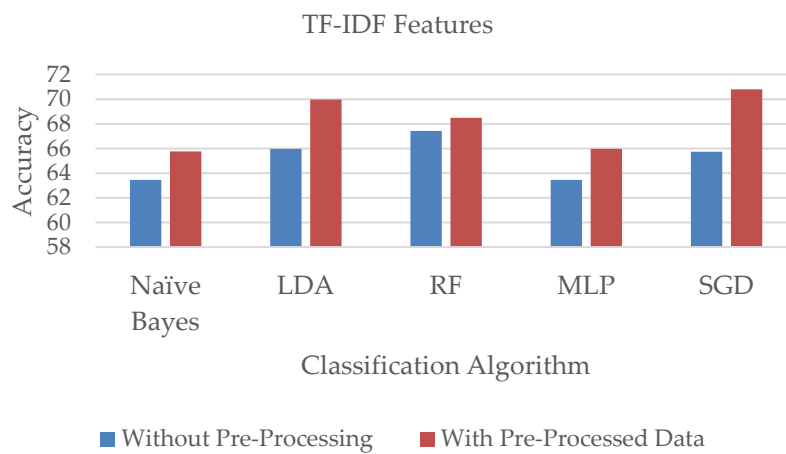| Classifier | Accuracy(%) without Pre-processing | Accuracy(%) with Pre-processed data |
|---|---|---|
| Naïve Bayes | 63.44 | 65.76 |
| LDA | 65.96 | 69.96 |
| RF | 67.43 | 68.49 |
| MLP | 63.44 | 65.96 |
| SGD | 65.75 | 70.80 |

**Figure 4:** *Effect of pre-processing on different classifiers for TF-IDF features*

## III. CNN Based Experiments

CNN architecture we have used is already explained in section III. As given in table 3, we have achieved 79% validation accuracy by implementing this model for all three emotion categories. Figure: - 5 shows the confusion metrics for happy, sad, and angry emotions. It clearly shows that the deep model achieves almost 9% higher accuracy than manual feature extractions.

**Table 3:** *Class-wise Precision, F1-score, and support*

| | Precision | Recall | F1-score | support |
|---|---|---|---|---|
| Angry | 0.79 | 0.87 | 0.83 | 237 |
| Happy | 0.79 | 0.75 | 0.77 | 228 |
| Sad | 0.79 | 0.75 | 0.77 | 208 |
| Accuracy | NA | NA | 0.79 | 673 |
| Macro avg | 0.79 | 0.79 | 0.79 | 673 |
| Weighted avg | 0.79 | 0.79 | 0.79 | 673 |

Komal D. Anadkat, Dr. Hiteisha M. Diwanji , Dr. Shahid Modasiya
EFFECT OF PREPROCESSING IN HUMAN EMOTION ANALYSIS
USING SOCIAL MEDIA STATUS DATASET
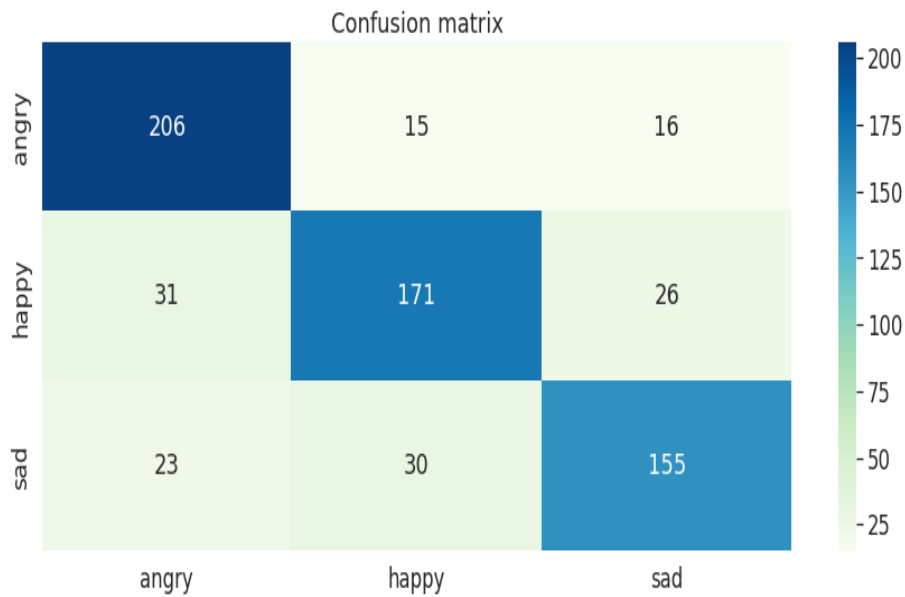
RT&A, No 1 (67)
Volume 17, March 2022

**Figure 5:** *Confusion Matrix for different emotions*

## V. Conclusion and Future Work

Even though detail research has been done in the area of analyzing social media text for emotion detection, still it faces many problems like poor accuracy when applied to real-life data. For making classification tasks efficient, pre-processing and feature extraction should be performed properly on the raw dataset. We applied different preprocessing techniques to the dataset of social media status and evaluated the performance of five different machine-learning classifiers and CNN for deep learning classification. Our experimental results concluded that the appropriate pre-processing methods could effectively improve the efficiency of the classifiers by around 4-5%. By the traditional approach, the highest accuracy has been achieved by SGD classifiers which is 70.59% for count vector features whereas the highest accuracy has been achieved by random forest classifiers which is 70.80% for TF-IDF features. For the deep learning approach, CNN achieved 79% accuracy for classifying the emotions in three different categories which is almost 9% higher than the traditional approach. However, this work can be further extended by incorporating the different pre-processing methods to find the optimal settings. In future, these results can be cross verified on other wide and real-world datasets of social media status for emotion classification.

## References

[1] Akrivi Krouska, Christos Troussas and Maria Virvou (2016).The effect of preprocessing techniques on Twitter Sentiment Analysis.*7th International Conference on Information, Intelligence, Systems & Applications (IISA).*

[2] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang(2018). A Review of Emotion Recognition Using Physiological Signals., *sensors.*

[3] F. H. Khan, S. Bashir, and U. Qamar(2014). TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems.* 57(1), 245-257, 2014.

[4] E. Psomakelis, K. Tserpes, D. Anagnostopoulos, T. Varvarigou(2014). Comparing methods for twitter sentiment analysis.*KDIR 2014 - Proceedings of the Int. Conf. on Knowledge Discovery and Information Retrieval*, 225-232.

[5] Janyce Wiebe, Theresa Wilson, and Claire Cardie (2005).Annotating expressions of

Komal D. Anadkat, Dr. Hiteisha M. Diwanji , Dr. Shahid Modasiya
EFFECT OF PREPROCESSING IN HUMAN EMOTION ANALYSIS
USING SOCIAL MEDIA STATUS DATASET

RT&A, No 1 (67)
Volume 17, March 2022

opinions and emotions in language. *Language resources and evaluation*. 39(2-3):165–210, 2005.

[6] Church, K., de Oliveira(2013). What's Up with Whatsapp: Comparing Mobile Instant Messaging Behaviors with Traditional SMS .*In: Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services*, 352–361.

[7] Sana Shahid (2018). Content Analysis of Whatsapp Conversations: An Analytical Study to Evaluate the Effectiveness of Whatsapp Application in Karachi. *International Journal of Media, Journalism and Mass Communications (IJMJMC).* Volume 4, Issue 1, 2018, 14-26 ISSN 2454-9479.

[8] Rice, D. R. and Zorn, C. (2013).Corpus-Based Dictionaries for Sentiment Analysis of Specialized Vocabularies. *Proceedings of NDATAD*, 1–17.

[9] G. Sidorov, S. Miranda-Jiménez, F. Viveros-Jiménez, and J. Gordon (2012). Empirical study of machine learning based approach for opinion mining in tweets. *11th Mexican International Conference on Artificial Intelligence, MICAI.*

[10] Zach CHASE, Nicolas Genain, and Orren Karniol-Tambour (2013). Learning Multi-Label Topic Classification of News Articles. *Semantic scholar.*

[11] Ma Chunling, Osherenko A, Prendinger H, et al (2005). A chat system based on emotion estimation from text and embodied conversational messengers. *Proc of the 4th Int Conf on Active Media Technology. Piscataway, NJ: IEEE*, 546-548.

[12] Aman S, Szpakowicz S. Identifying expressions of emotion in text (2007). Proc of the 10t Int Conf on Text, Speech and Dialogue. *Berlin: Springer*, 196-205.

[13] Alm C, Roth D, Sproat R.(2005). Emotions from text: Machine learning for text-based emotion Prediction. *Proc of the 2nd Conf on Human Language Technology and on Empirical    Methods in Natural Language Processing. New York: ACM*, 579-586.