Emotion Detection From Tweets

Utsav Baghela
utsav21101@iiitd.ac.in
MTech CSE, IIIT Delhi
New Delhi, India

Saurabh Pandey
saurabh21077@iiitd.ac.in
MTech CSE, IIIT Delhi
New Delhi, India

Kirpali
kirpali21040@iiitd.ac.in
MTech CSE, IIIT Delhi
New Delhi, India

Ekta Gambhir
ekta21025@iiitd.ac.in
MTech CSE, IIIT Delhi
New Delhi, India

Karan Singh
karan21038@iiitd.ac.in
MTech CSE, IIIT Delhi
New Delhi, India

1 PROBLEM STATEMENT

Social Media Analysis provides an overview of people's opinions and sentiments towards certain entities. Users post their thoughts and insights on these networking sites.

The main aim of our project work is to comprehend and classify the tweets posted on Twitter, into five classes of emotions. Previously, a very broad classification scheme such as positive-negative, happy-sad, etc. has been developed by the NLP community. Since very little work is done on an in-depth classification of tweets into several classes, we decided it was a worthy problem to tackle.

With our work we seek to create a multi-class classification system that segregates tweets into one of several classes that are less general and offer a profound idea of the sentiment behind it.

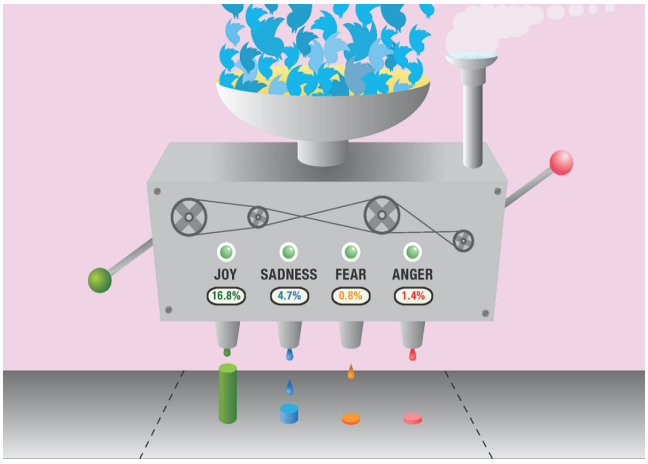

Figure 1: Tweets to Emotions

2 MOTIVATION

More than a decade of being launched, Twitter now has an active user base of around 300 million people. The platform has a strong impact and influence over the world. Being used for awareness about social issues, to raise an opinion or to spread political messages, today's youth is very much involved in all that's going on. There is a lot of interest in what people are tweeting about in a lot of domains like politics, business and climate change.

Every tweet made has an associated feeling and emotion which is termed as the Sentiment. Even with the character-limit of 280, a lot is expressed. Emoticons and hashtags are also the high impact forms of expression. Thus, a lot of work is being done to analyse the posts made on twitter and to determine the emotion behind them.

3 LITERATURE REVIEW

[1] The team built two SVM classifiers (obtaining the most optimal parameters using cross-validation), one to detect the sentiment of messages such as tweets and SMS (message-level task) and one to detect the sentiment of a term within a message (term-level task), obtaining an F-score of 69.02 in the message-level task and 88.93 in the term-level task. They also generated two large word–sentiment association lexicons, one from tweets with sentiment-word hashtags, and one from tweets with emoticons.

[2] The corpus contains 5,553 tweets and is developed using small-scale content analysis, They classified twitter tweets into 28 emotion categories. Out of 28 emotional categories 33% of the tweets containing emotion are positive, 13% are negative and only 3% are neutral.. They used Machine Learning models to predict the classifier and metric for result used was Micro-averaged F1 for multi-class-single (MCS) and multi-class-binary (MCB).Existing classifiers achieve only moderate performance in detecting emotions in tweets even those trained with a significant amount of data. BayesNet classifiers produce consistently good performance for fine-grained emotion classification.

[3], The paper mentions about the use of NLP in Social Media posts like twitter. Further it very randomly define as to how applying NLP models is very different for Traditional posts/articles/blog compared to Social Media posts/tweets etc and the challenges that it create. Moreover, it explains in about how to collect twitter dataset and modify it as per the challenges of the target model like handling emoji's and lexical lexical relationships.

[4] Research article focuses on classifying Amazon product reviews into three classes - positive, negative and neutral. It uses data collected from Amazon product reviews and adopts a machine tagging approach, implemented via bag of words model.Prior to machine tagging, several pre-processing steps are namely tokenisation, lemmatization and part of speech tagging (POS-tagging) were implemented. This bag of words model counts the occurrences of positive and negative tokens in a sentence and assigns it a ground truth label based on that. Feature vectors were developed based

on total tokens and several models such as SVM, Naive Bayesian Model and Random Forests were used out of which the Random Forest classifier which was essentially an ensemble method using bagging, outperformed the rest.

[5] The team extracted tweets with the help of the Twitter Streaming API around 520,000 tweets as raw data performing text preprocessing and feature augmentation to add additional attributes that are effective for emotion identification. The decision tree, decision forest and decision majority rule are used to classify the tweets into the eight classes. The proposed classifiers are implemented on both WEKA and Apache Spark system over Hadoop cluster for scalability purpose.

4 BASELINE RESULTS

4.1 Dataset

The data is basically a collection of tweets annotated with the emotions behind them. We have three columns: id, emotions, and text. In "text" we have the raw tweet. In "emotions" we have the emotion behind the tweet. To suit our 5 class classification problem we manually re-annotated the tweets, primarily the sentiment column to fit our classes.

4.2 Data Preprocessing

As per the above foreseen challenges of a social media text which are comparatively a lot more informal and thus complex to process, we need to perform suitable and specific preprocessing that can allow our model to get more accurate results. For this purpose we need to utilize the common informal features of tweets.

• **URLs:** URLs present in the tweets mostly provide no significant value to identify the emotions of the author. So we remove them to get a cleaner data.

• **Mentions:** The mentions are used to get the attention of some other user which proves as not valuable for detecting the emotions, and thus can be removed.

• **Hashtags:** Hashtags are considered very important in a tweet. But they can be hard to process as they can very commonly happen to be a combination of words or abbreviations. So we need to carefully process the hashtag texts to get the meaningful data. For this we can divide the hashtags as follows:

• **Emoji's:** The emoji's are of key importance when it comes to emotion detection. For this purpose we classify the existing common emoji's that can help in emotion detection as per the categories that we need. And then based on that category we can give a probability score to the hashtag of the expected category.

• Next we further clean the data by checking for spelling errors and repetitions in words in the remaining text of the tweets and fetch the valid words from it.

4.3 Word Vectorization

The Word2Vec makes a vocabulary of all the tokens present in the tweets. Each word is converted to a numerical vector. The number of vector values is equal to the chosen size. These are the dimensions on which each token is mapped in the multi-dimensional space. It takes into account the context in which a token appears in the space. Thus, words that are similar in meaning are closer to each other in the multi-dimensional space.

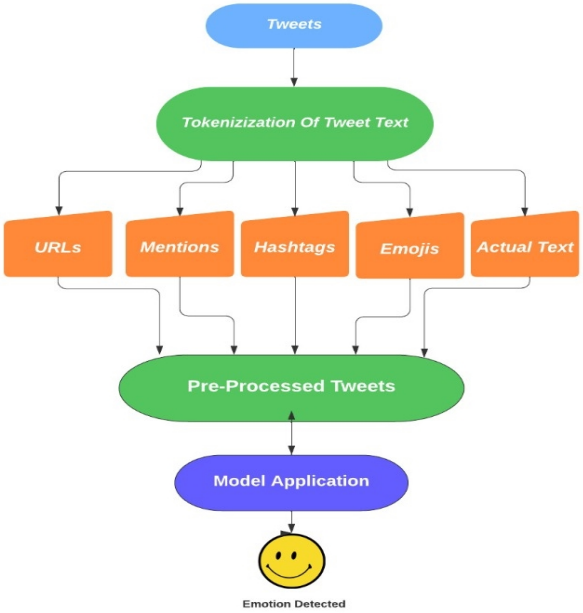

Figure 2: Workflow

4.4 Models

The models have been trained on three different classifiers - Decision Trees, Random forest and Logistic Regression.

•Logistic Regression outperforms the Random Forest classifier very minutely.

•The best performance on the test set comes from the Logistic Regression with features from Word2Vec.

Model	Accuracy
Logistic Regression	65.37178976707148
Random Forest Classifier	61.602876766872384
Decision Tree Classifier	46.430171212422856

Figure 3: Classification Models

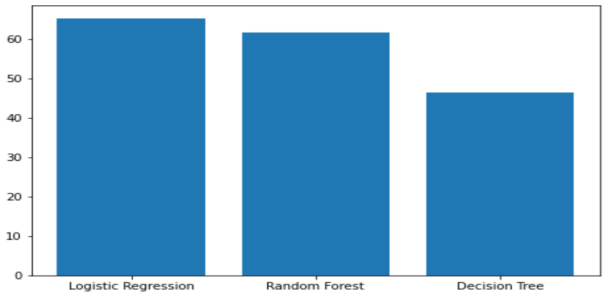

Figure 4: Accuracy Score comparison

5 PROPOSED METHOD

5.1 Preprocessing

•Original tweets are converted to plain English and splitting of data using the regular expressions For example, " ohh...he is angryð got4 senior..., @user" Will get converted to 'Ohh' 'he' 'is' 'angry' 'got4' 'senior' 'user'.
•The next step involves, Removal of stopwords from processed data after regularization.
•The Word count of each word in each document created i.e vectorization of each word for all documents was made.
•Generation of Tf-IDF vectors using L2 normalization and sublinear TF-scaling using TF-IDF transformer class.

5.2 Models used

Emotion Detection from tweets is a classification problem, there are different classification techniques of Machine learning. Having a non-linear separation dataset, we have used different traditional ML methods. We are going to use Machine learning models as Machine learning models were performing better in baseline models.
We proposed to train our model using a Decision Tree, Logistic Regression and Random Forest algorithms. With the pre-processing steps using TF-IDF vectors, accuracy was improved.

5.3 Results

After creating word vectors using TF-IDF accuracies was improved a lot, To measure the performance of each model we have taken the accuracy parameter. We will further use models such as KNN

Model	Accuracy
Decision Tree	83.9831
Logistic Regression	89.5636

Figure 5: Proposed Models Comparison

and Random Forest and we will test accuracies on them.

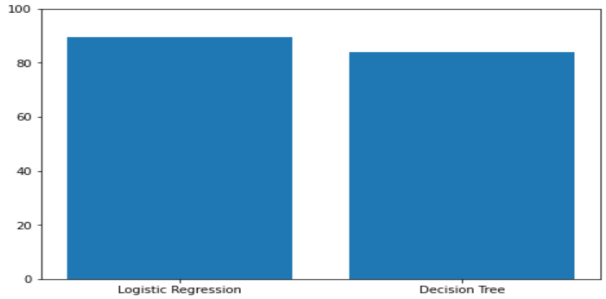

Figure 6: Proposed Models

6 REFERENCES

[1] Saif M. Mohammad, Svetlana Kiritchenko, Xiaodan Zhu 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets .

[2] Jasy Liew Suet Yan, Howard R. Turtle 2016. Exploring Fine-Grained Emotion Detection in Tweets.

[3] Maria Krommyda, Anastatios Rigos, Kostas Bouklas, Angelos Amditis 2020. Emotion detection in Twitter posts: a rule-based algorithm for annotated data acquisition.

[4] Xing Fang, Justin Zhan 2015. Sentiment analysis using product review data.

[5] Jaishree Ranganathan, Nikhil Hedge, Allen S. Irudayaraj, Angelina A. Tzacheva 2018. Automatic Detection of Emotions in Twitter Data - A Scalable Decision Tree Classification Method