# *Exploratory data analysis on Reddit data: An efficient pipeline for classification of flairs*

Reshma Shaji
*Department of Electronics and Communication Engineering*
*Government Engineering College, Barton Hill*
APJ Abdul Kalam Technological University, India
reshma.tkm17ec101@gecbh.ac.in

2021 IEEE Seventh International Conference on Multimedia Big Data (BigMM) | 978-1-6654-3414-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigMM52142.2021.00018

***Abstract*— Internet has now become a platform for people to learn new things, share their opinions, and communicate with each other from anywhere. As technology is growing, the number of internet users are growing as well. With the increase in number of users, the amount of data is also enormously increasing. Social networking sites like Reddit, Facebook, Twitter have gained global popularity as a platform through which people can create individual public profiles, interact with real friends, share their interests and opinions, and post messages on any topics. Each post is tagged for filtering purposes. These tags are called flairs in the Reddit world. In this paper, a comparative data analysis using existing Machine Learning and Natural language processing techniques is provided to detect the flair of each Reddit post. Proper data analysis was done on the data using different features and a pipeline of various natural language processing techniques like Count Vectorization and TfIdf Transformation, and various machine learning techniques like K-Nearest Neighbor (KNN), Decision Tree, Support Vector Machines (SVM), and Logistic Regression was used to research on the data, and classify the flairs**

***Keywords—Reddit, Subreddit, Flair, Data Mining, Exploratory data analysis, Machine learning, Classification, K-Nearest Neighbor (KNN), Decision Tree, Support Vector Machines (SVM), Logistic Regression***

## I. INTRODUCTION

Popularity of the social networking sites are rapidly growing over the past few years. As most of these sites are free to use and user friendly, the number of people joining them are only growing. Users can connect with new people, share opinions with likeminded people, and stay in touch with old friends and colleagues. These sites allow people to post real time messages about their opinions on a variety of topics, discuss current issues, complain, and express their views of products they use in daily life. So, companies have started to study these user opinions on different products that people express on these sites, so that they can offer their products and services as per the users' requirements. Companies can find new trending topics and develop products based on that. These sites also allow different businesses to promote their new products. Manually analyzing such large user generated contents are extremely tedious and automating this process is therefore important.

Reddit is one among such popular sites. Reddit is a collection of forums where people discuss varied topics. Each topic forms a subreddit where people post messages on that topic, upvote or downvote posts and comments under them. Each post is tagged for filtering purposes. These tags are called flairs. Analyzing these messages can provide information on the trending and mostly discussed topics.

This paper focuses on exploring and analyzing the data on /r/india subreddit and insights gathered from the analysis are discussed. A comparison on the performance of different machine learning algorithms in classification of flairs using different features of the data is provided. A pipeline of various natural language processing techniques, and machine learning techniques was used to research on the data, and classify the flairs.

The rest of the paper is organized as follows. In section 2, details about the data is provided. In section 3, exploratory data analysis is briefly discussed. In section 4, we discuss about how the data was pre-processed and about the natural language techniques used. In section 5, we discuss about the different machine learning approaches and classification algorithms. In section 6, the results are presented and discussed. In section 9, we conclude and give future directions of research.

## II. DATA DESCRIPTION

Reddit is a collection of forums where people can share any content as a thread or a comment on other people's posts. Reddit is broken up into more than a million communities known as "subreddits," each of which covers a specific topic. The name of a subreddit begins with /r/, which is part of the URL that Reddit uses. For example, /r/india is a subreddit where stories from India can be created and shared by the users. For this research, posts from /r/india was used. A flair is a 'tag' that can be added to threads posted on the reddit website within a sub-reddit. They help users filter specific kind of posts based on their preferences.

The data used for this research was scraped from the Reddit India website using PRAW. PRAW is a Python wrapper for the Reddit API that helps to scrape data from subreddits. The data consists of 3000 rows of real time messages. Each row consists of a title, selftext(body) of the post, comments under the post, and its corresponding flair.

**Title:** The title of the submission.

**Selftext:** The submissions' text or an empty string if a link post.

**Comments:** Provides an instance of a forest of comments, starting with multiple top-level comments.

The flairs in /r/india website were 'AskIndia', 'Non-Political', 'Scheduled', 'Photography', 'Science/Technology', 'Politics', 'Business/Finance', 'Policy/Economy', 'Sports', and 'Food'.

**Policy/ Economy:** All posts about central or state policies are flaired as 'Policy/Economy'.

978-1-6654-3414-0/21/$31.00 ©2021 IEEE
DOI 10.1109/BigMM52142.2021.00018

**Politics:** All posts about politics and politicians in India are flaired as 'Politics'

**Science/Technology:** All submissions about sci-tech are flaired as 'Science & Technology' as long they have no policy/political aspect.

**Food:** All submissions about food are flaired as 'Food' as long they have no policy/political aspect. Also, submissions regarding recipes or images of food that you have cooked.

**Sports:** All submissions about sports are flaired as 'Sports' as long they have no policy/political aspect.

**Photography:** All submissions of photographs that are taken by you.

**Business/Finance:** All submissions about business and finance are flaired as 'Business/Finance' as long they have no policy/political aspect.

**AskIndia:** This flair is used only when a question is posted and if none of the previous flair is applicable.

**Scheduled:** For all the scheduled submissions and other weekly/biweekly/monthly scheduled submissions. For example, the scheduled discussion topic of Wednesday might be Mental Health discussion thread, and if you want to post on that.

**Non-political**: Any post that you do not feel is political in nature. When none of the above flairs are applicable this flair is used.

300 posts of each flair were specifically scraped, that counts to a total of 3000 rows of real time posts. Due to the slow processing speed of the machine used for the research, large amount of data couldn't be obtained.

## III. EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is an approach used for visualizing and studying data to uncover statistical regularities that might not be apparent otherwise [3].

**Feature:** An attribute of the data with respect to which the analysis is made.

The features that will be used for this research are title, selftext, comments, and a combination of title, selftext and comments.

Analysis was done using these features of the data to discover patterns, gather insights from the data and test the performance of the classification models. Important results were found out as a result, which will be discussed in the Results and discussions section.

## IV. NATURAL LANGUAGE PROCESSING

Roughly speaking, statistical NLP associates probabilities with the alternatives encountered in the course of analyzing an utterance or a text and accepts the most probable outcome as the correct one[5].

Language is a structured medium humans use to communicate with each other. It can be in the form of speech or text. But machines cannot understand this data. Natural language processing techniques helps to extract useful information from text data. Some of the approaches to process text data are Count Vectorization and Tf-Idf Transformation.

### *A. Data Pre-processing*

Reddit posts mostly contains opinions of different users. So, the raw data might contain undesirable text due to which results might not give efficient accuracy, and might make it hard to understand and analyze. So, proper pre-processing must be done on raw data.

The raw data was pre-processed as follows,

1) All the uppercase letters were converted to lowercase.
2) The text string (the string might be in encoding or unicode standard) was then converted to Unicode. Unicode is an abstract encoding standard that aims to list every character used by human languages and give each character its own unique code.
3) Hashtags, punctuations, symbols, and numbers were removed.
4) Stop words were removed. Stop words refer to the most common words in a language.

### *B. Count Vectorization*

Count vectorization transforms a given document into a vector on the basis of the frequency (count) of each term that occurs in the entire document. Consider a Corpus C of D documents $\{d_1, d_2 \ldots .d_n\}$ and N unique tokens extracted out of the corpus C. The size of the Count Vector matrix M will be given by n X N. Each row in the matrix M contains the frequency of tokens in document $D_i$.

### *C. Tf-Idf Transformation*

Term frequency-inverse document frequency, is a numerical statistic that is used to reflect how important a word is to a document.

#### *a) Term frequency (Tf)*

It is a measure of how frequently a term, t, appears in a document, d. It is a common term weighting scheme in information retrieval.

Tf = (Number of times t appears in d) / (Total number of terms in d)

#### *b) Inverse document frequency (Idf)*

Idf is a measure of how important a term is. Words relevant to the document appear often in the document.

Idf = log(N/n)

where, N is the total number of documents and n is the number of documents in which a term t has appeared in.

Tf-Idf score for each word in the corpus is,
Tf-Idf score= Tf*Idf
Tf-Idf score is used to weight the term frequency by its Idf values. Higher the Tf-Idf score of a word, higher is its importance.

The pipeline of Count vectorization followed by Tf-Idf transformation was done so that the term count vectors for different tasks can be separately analyzed. With the word counts computed by Count Vectorization, the Idf values and Tf-Idf scores was computed by Tf-Idf transformation.

## V. MACHINE LEARNING

Machine learning focuses on the use of data and algorithms to imitate the way humans learn, to gradually improve its accuracy. Using statistical methods, algorithms are used to make classifications or predictions. This research uses supervised learning techniques. The machine learning algorithm is trained on labeled data. This research is a multi-class classification task with 10 classes. We use a balanced data-set of 300 instances for each class and therefore the baseline probability of each class is 0.1.

The machine learning algorithms used for the classification is briefly described,

### *A. K- Nearest Neighbor (KNN)*

K-Nearest Neighbor is one of the simplest Machine Learning algorithms. KNN algorithm assumes the similarity between a particular data sample and nearby samples and assigns the new sample to the category that is most similar to the available categories. KNN algorithm first selects a number K. Euclidean distance of the K neighbors are calculated. K nearest neighbors are chosen according to the calculated Euclidean distance. Among these k neighbors, the number of the data points in each category is counted. New data points are assigned to that category for which the number of the neighbors is maximum.

### *B. Decision tree*

A decision tree is a flowchart like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). Decision tree selects the best features for the root node using the concept of information gain ratio and then builds its subtrees in a recursive manner.

### *C. Support Vector Machine*

Support Vector Machine is one of the most successful algorithms for classification. It places class-separating hyperplanes in the original or transformed feature space, and the new sample is labeled with the class label that maximizes decision function—the distance between support vectors (examples of different classes closest to the hyperplane).

### *D. Logistic regression*

Logistic regression is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable. It uses maximum likelihood estimation to evaluate the probability of the class membership

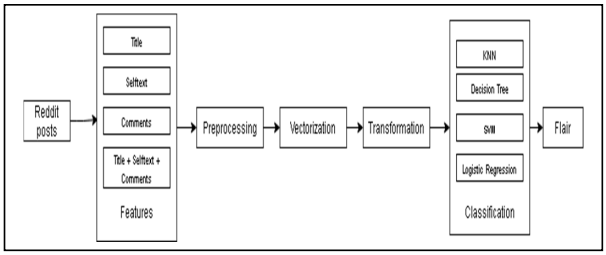


Fig. 1. Pipeline for classification of flairs

## VI. RESULTS AND DISCUSSION

### *A. Analysis of first 3000 posts scraped from Reddit India site*

Initially, first 3000 posts in the Reddit India website were scraped. The analysis on this data shows that almost 50% of the posts were on Non-political topics. The second most discussed topic was Politics, covering almost 20%, and then came Sports, but was only 6.1%. The rest of the topics covered about 3.5% to 4.5%. The least number of messages were on the scheduled flair, with a percentage of 2.6. This is because scheduled flair only contains posts of the scheduled days on the weekly/biweekly/monthly scheduled discussion topic.

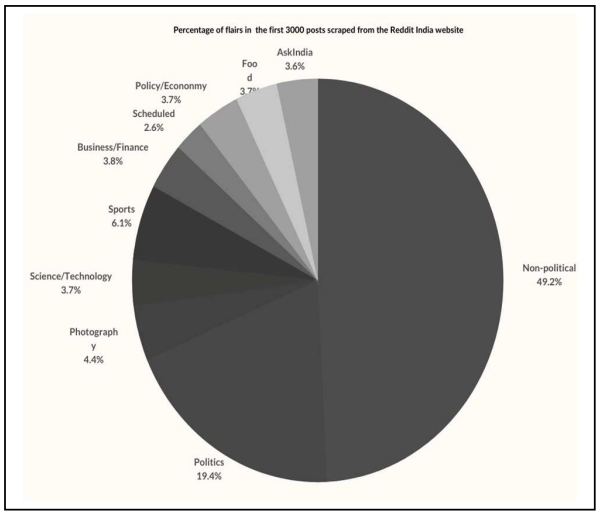


Fig. 2. Percentage of flairs in the first 3000 posts scraped from Reddit India website

Since this data is not balanced, it wasn't used for further research. New dataset containing 3000 posts was scraped again, that contains 300 posts of each flair, to keep the dataset balanced.

### *B. Performance of the classifiers*

The performance of the classifiers was evaluated based on accuracy. Table 1,2,3 and 4 compares the performance of the four models each using title, selftext, comments and title+selftext+comments of the posts for classification, respectively.

#### *a) Accuracy*

Accuracy represents the number of correctly classified data instances over the total number of data instances.

Accuracy score = (TP+TN) / (TP+TN+FP+FN)
where TP, FN, FP and TN refer respectively to the number of true positive instances, the number of false negative instances, the number of false positive instances and the number of true negative instances.

Since the dataset used is balanced and every flair to be classified is equally important, using accuracy score as the evaluation metric is better.

TABLE I. TITLE AS FEATURE

| Algorithm | Accuracy score |
|---|---|
| KNN | 0.543216 |
| Decision Tree | 0.597342 |
| SVM | 0.778345 |
| Logistic Regression | 0.765832 |

When the title was used for classification, the result shows that SVM performs the best with an accuracy score of 0.778 and KNN performs the worst. Models with title used for classification performing well might be because of the fact that title consists of all the keywords to be expected in the selftext. Thus, titles will contain lesser noise.

TABLE II. SELFTEXT AS FEATURE

| Algorithm | Accuracy score |
|---|---|
| KNN | 0.257847 |
| Decision Tree | 0.283754 |
| SVM | 0.390625 |
| Logistic Regression | 0.425184 |

When the selftext part of the data was used to classify the flairs, it was found that Logistic Regression model performs the best with an accuracy score of 0.425 and KNN the worst. Using selftext for classification provides bad results overall. Models using selftext as feature doesn't perform well, might be because the body of the message might contain more noise and the weighted terms might not be having enough variation to capture critical patterns.

TABLE III. COMMENTS AS FEATURE

| Algorithm | Accuracy score |
|---|---|
| KNN | 0.518658 |
| Decision Tree | 0.532973 |
| SVM | 0.754675 |
| Logistic Regression | 0.761362 |

When the comments under the posts was used to classify the flairs, it was found that Logistic Regression model performs the best with an accuracy score of 0.761 and KNN the worst. Models with comments as feature also performs good.

TABLE IV. TITLE+SELFTEXT+COMMENTS AS FEATURE

| Algorithm | Accuracy score |
|---|---|
| KNN | 0.517375 |
| Decision Tree | 0.648649 |
| SVM | 0.783784 |
| Logistic Regression | **0.787645** |

When the title, selftext and comments were used to classify the flairs, results show that Logistic Regression model performed the best with an accuracy of 78.8% and KNN the worst. Logistic regression model using the combination of title, selftext and comments for classification performs the best in the overall analysis. The results show that KNN is not a good model for multi-class text classification.

## VII. CONCLUSION

This paper presents an analysis on Reddit data and comparative study of existing techniques of machine learning and natural language processing to process the data and classify the flairs. Analysis on the first 3000 posts in the Reddit India website shows that almost 50% of the posts were on Non-political topics and second mostly discussed topic was politics. Research results show that machine learning methods, such as Logistic Regression and SVM provides the highest accuracy, while KNN the worst on every case. Thus, we can conclude that KNN is not an efficient algorithm for multi-class text classification. Effects of various features on each of the classifiers were also studied, showing that using title, selftext and comments of the posts for classification provided the best results. Using title of the posts to classify the flair also provided good results. This reveals that less noisier the data and more the content, and higher the accuracy because machine learning models detect specific words to identify the sentiment and classify the flairs. In future work, we will explore even richer linguistic analysis.

## REFERENCES


[1] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011Workshop on Languages in Social Media, 2011.

[2] Vishal A. Kharde, S S Sonavane, "Sentiment Analysis of twitter data: A survey of techniques", IJCA, 2016

[3] Warren R. Greiff, The use of Exploratory Data Analysis in Information Retrieval Research, Advances in Information Retrieval, pp 37-72, INRE Vol. 7.

[4] Wen Zhang, Takeshido Yoshida, Xijin Tang, A comparative study of TF*IDF, LSI and multi-words for text classification, Expert Systems with applications, pp 2758-2765, Vol 38

[5] Ruslan Mitkov, Oxford Handbook of computational Lingusistics, first edition,2005

[6] Bojana R.Andjelkovic Cirkovic, Machine learning approach for breast cancer prognosis prediction, Computational Modeling in Bioengineering and Bioinformatics, 2020.

[7] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[8] Jake Widman, What is Reddit, digitaltrends.com, 2021

[9] R. Watermeyer, Social Networking Sites, Encyclopedia of applied ethics (second edition), 2012

[10] Gilbert Tanner, Scraping Reddit data, towardsdatascience.com, 2019

[11] https://praw.readthedocs.io/en/latest/code_overview/models/submission.html#submission.