Reddit Flair Detection

Nivedita Singhal - 02701032017
Sunaina Chaudhary - 05301032017
Kiran Chawla - 05401032017
Shivika Gupta - 07301032017
Ena Motwani - 07801032017

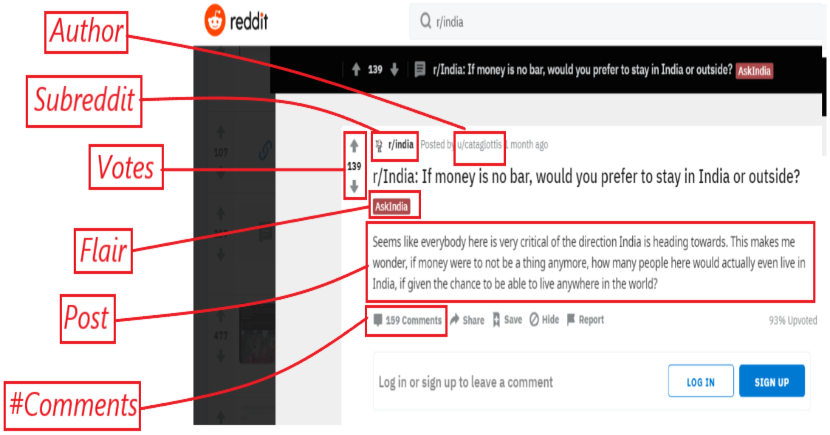
Indira Gandhi Delhi Technical University for Women, New Delhi,110006, India
{niveditasinghal7,sunaina17ch,kiranchawla09,shivikagupta.599,enamotwani98
}@gmail.com

Abstract. Reddit, a social networking website that provides a platform for its users to create posts, discussion groups etc., generates a huge amount of data every day. This data needs to be organized and analyzed in order to label the data and divide it into specific categories. The members of Reddit, that are people who use Reddit, can submit their views on this platform, for example, posts, links, images, and many more in order to express themselves and to reach out to the world. Posts on this social networking platform are organized with the help of categories which are defined by Reddit and are known as "subreddits", few of them to mention are- science, food, India, Europe, video games, etc.
Through this study, an effort has been made to design a system that, given a link, can detect the flair (category) of a Reddit post, is made. The data is collected using the Reddit Application Program Interface. This requires the word embedding, that is, words that have a meaning similar to each other are represented analogously. Word embedding is done using techniques like Word2Vec, CountVectorizer, TF-IDF Vectorizer. For the prediction of the flairs/categories, various machine learning algorithms are used such as K Nearest Neighbour, Decision Tree, Naive Bayes, Linear Regression, etc. The results of this study illustrate the importance of the proposed work.

Keywords: Machine Learning, Reddit, Flair Classification, Natural Language Processing

1 Introduction

[3] Reddit is an American social networking platform. It was founded by Steve Huffman and Alexis Ohania in 2005. As of 2020, it is the seventh most visited website in the United States. Fig.1 represents an annotated depiction of Reddit post.



r/India: If money is no bar, would you prefer to stay in India or outside?

Seems like everybody here is very critical of the direction India is heading towards. This makes me wonder, if money were to not be a thing anymore, how many people here would actually even live in India, if given the chance to be able to live anywhere in the world?

Log in or sign up to leave a comment

Fig. 1. A Reddit website post, with important elements annotated

Subreddits are the communities around which Reddit is organized. The name of a subreddit starts with ”r/” or ”/r/”. Every subreddit has volunteer community members that create and enforce the rules specific to a subreddit. They can remove content from their specific subreddit and can refrain certain users from submitting content.

A user of Reddit is referred to as a Redditor. One must create an account on Reddit to become a Reddit user. There is no compulsion on the user to register with their real name. The username starts with ”u/” or ”/u/”. A Redditor can perform any of the following actions:

Submit a new post: A user can submit a post that may be text, image, video, a link, or a title without any content to a subreddit.
Comment: A user can either directly comment on a post, or any other comment.
Upvote or Downvote: Content is downvoted if it is not relevant to the subreddit it was posted to, and upvoted if it is relevant. A user can upvote or downvote a post or a comment.

Redditors share their opinions, perspectives, insights, and outlooks on various subjects by posting on the website or by commenting on the posts of other members. Users can up-vote the content if they like it. These actions generate a lot of data every day. This study aims to build a system that, after providing a link, can detect the flair (category) of a Reddit post among the subreddits. The categories can be Food, AskIndia, Technology, etc.

1.1 Problem Statement

Reddit Flair Detection

Flairs are tags or icons that one can sometimes see next to a Reddit user's name or link post title in subreddits that support it. The content uploaded by Reddit members is organized by subject, and in categories defined by members called "subreddits", these categories may include music, books, games, food, fitness, etc. In this project, the main goal is to develop a system that when provided a link to a Reddit post, can predict the flair (category) of a Reddit post.

1.2 Objective

In this section, the objective of this study is presented.

Data Acquisition: The aim is to collect the posts from the r/india subreddit making sure to have enough posts belonging to each flair on the subreddit. In this study, the objective is to collect as much data associated with the post as possible (comments, link, timestamp, user handle, comments, etc.)

Flair Detection: Training of a machine learning model to learn the flair prediction from the features of the posts is attempted in this part. Multiple models using different sets of features (title, comment, link) are trained and report the test set accuracy on each one of them. As an end result, Prediction of the flair of the post based on classified flairs on Reddit like AskIndia, Food, Business, Political, and many more are made.

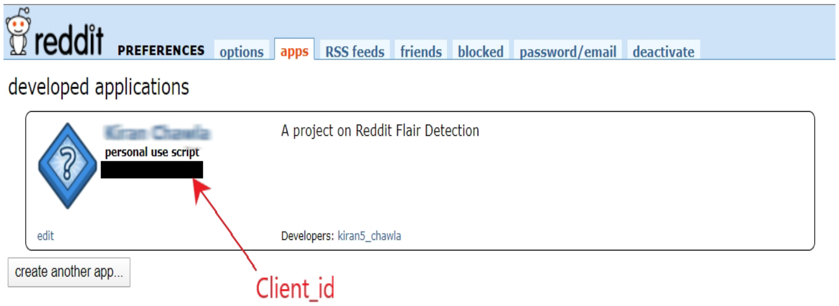
Deployment through Web Application: As a final step, deployment of the code and the trained model as a web application with a user interface where the user will input a link to a Reddit post and the model would be used to predict the flair for that post is made.

2 Methodology

In this section, the methodology used in this study is presented. First, the description of the dataset used is explained.

2.1 Dataset Description

The fetching of the data on Reddit is a challenge itself. The Application Program Interface (API) of Reddit can be accessed by anyone. In the Reddit API, the developers need to create the application to start collecting data by mentioning the purpose of making the application and specifying how the data would be used. The developers get the authentication from Reddit which includes the client_id and client_secret. This client_id and client_secret is used by the praw library in Python to fetch the data from the Reddit API. The API interface is as shown in Fig.2 and Fig.3.



developed applications

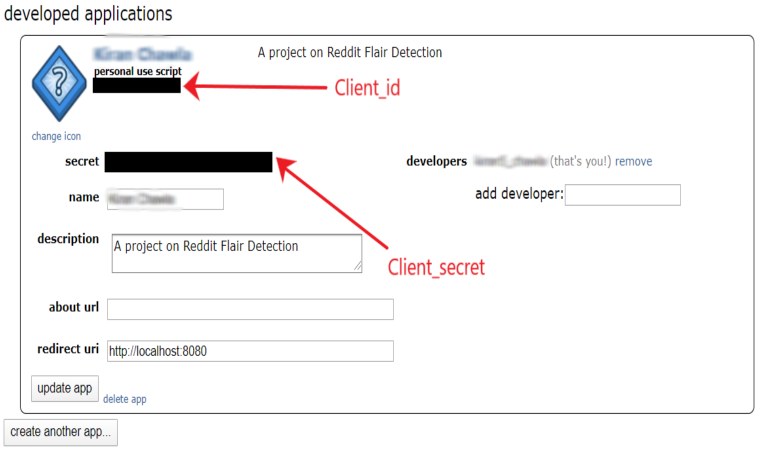
[redacted]
personal use script
[redacted]

A project on Reddit Flair Detection

edit

Developers: kiran5_chawla

Fig. 2. Creating app on Reddit API


developed applications

[redacted]
personal use script
[redacted]

A project on Reddit Flair Detection

secret [redacted]

developers [redacted] (that's you!) remove

add developer:

name [redacted]

description A project on Reddit Flair Detection

about url

redirect uri http://localhost:8080

Fig. 3. Reddit API

But the major drawback of using the Reddit API for collecting the posts is the limitation imposed by Reddit on collecting large historical data. Reddit allows a restricted amount of data to be collected by the users at a time.
In this study, data is collected using the Reddit API applying the praw library (Python Reddit API Wrapper) in Python. The flairs, from subreddit r/india, that are considered in our study are:

1. AskIndia
2. AMA
3. Reddiquette
4. Business/Finance
5. Food
6. Non-Political
7. Photography
8. Policy/Economy
9. Politics
10. Scheduled
11. Science/Technology
12. Sports

The flair distribution is shown in Fig. 4. The flair named [R]eddiqette is comparatively newer and hence has fewer number of posts associated with it. All other flairs have a proportionate number of posts for each.

Table 1 represents the key entities of the dataset. The dataset in this study is a labeled dataset, as only the posts related to the above-mentioned flairs are fetched from the Reddit API.

Table 1. Details of Data Set.

Details	**Count**
Number of Instances	4987
Number of attributes	12
Label Information	Label Present

Dataset comprises of following data attributes:

1. Flair
2. Title
3. Created_at
4. Url_address
5. Comments
6. #Comments
7. Body
8. Author
9. Word_Count
10. Char_Count
11. Hashtags
12. Timestamp
13. Label

Table 2 describes attributes of data along with the explanation of the attributes which are further used in this study.

Table 2. Details of Data Attributes.

Data Attributes	Brief Explanation	Type
Title	The Title Of the Post On Reddit	String
Created_at	When was the post created	Date
#Upvotes	Number of votes i.e. likes on the Post	Numeric
Url_address	URL of the post	String
Comments	Comments on "the post"	String
#Comments	Number of comments on that post	Numeric
Body	Body of the post i.e content	String
Author	Who wrote the post	String
Word_count	Number of words in the post	Numeric
Char_count	Number of characters in the post	Numeric
Hashtags	Number of Hashtags on the post	Numeric
Timestamp	Time of Creation	Time
Label	Flair	Categorical

2.2 Data Exploration

Data exploration is an extremely beneficial practice in order to make big and small data effortless to understand. This subsection contains the plots for various relations among the data attributes so as to understand the data better.

Fig. 5 shows the titles of the posts with maximum number of hashtags.On a similar note, Fig. 6 illustrates the titles of the posts with maximum number of comments. The correlation among the data attributes is shown in Fig. 7. The correlation between the attributes is strong if it is closer to absolute(1) which is represented by the lighter sections in the correlation plot. Fig. 8 illustrates the distribution of number of comments on the posts.

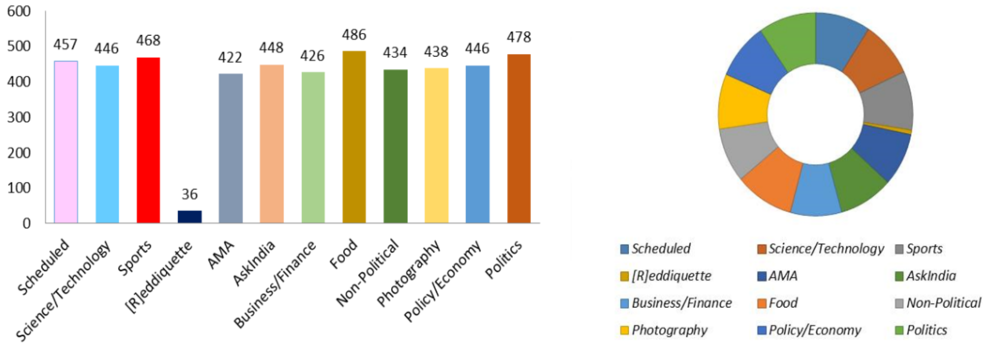


Fig. 4. Flair Distribution

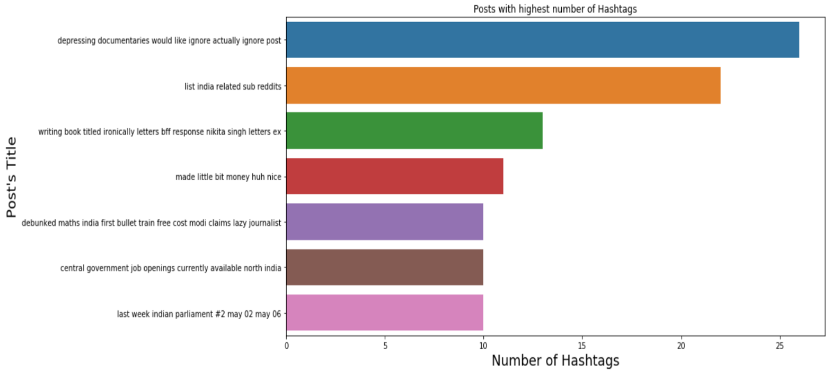


Fig. 5. Posts with highest number of Hashtags

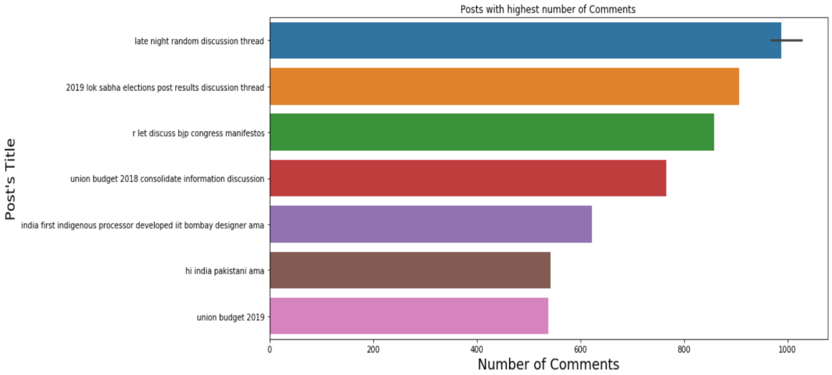


Fig. 6. Posts with highest number of Comments

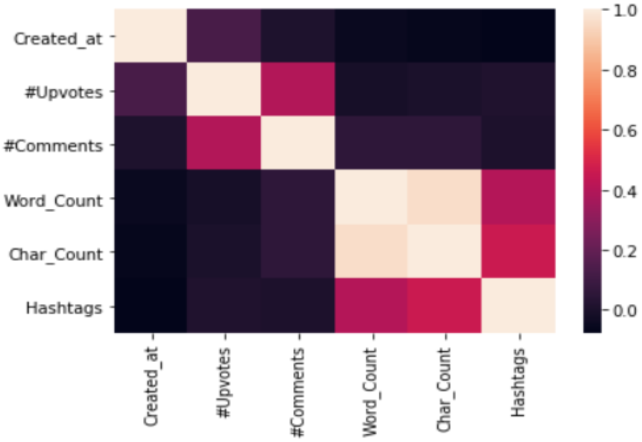


Fig. 7. Correlation

Fig. 8 depicts the distribution of comments for the posts in the dataset. It is evident that most of the posts have atmost twenty comments in the dataset. It can be inferred that an average of two posts per date and time combination are there in the data. Fig. 9 illustrates the number of posts of a particular date and time, in the collected data. Fig. 10 depicts the response of the Redditors on the

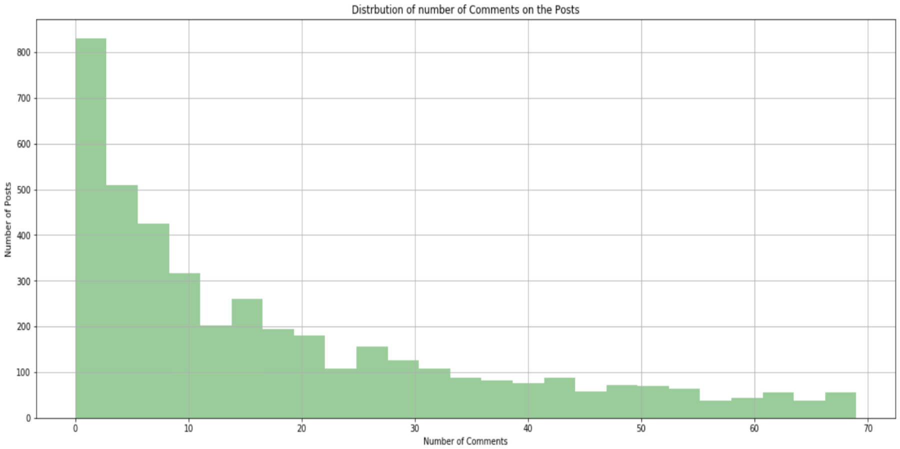


Fig. 8. Distrbution of number of Comments on the Posts

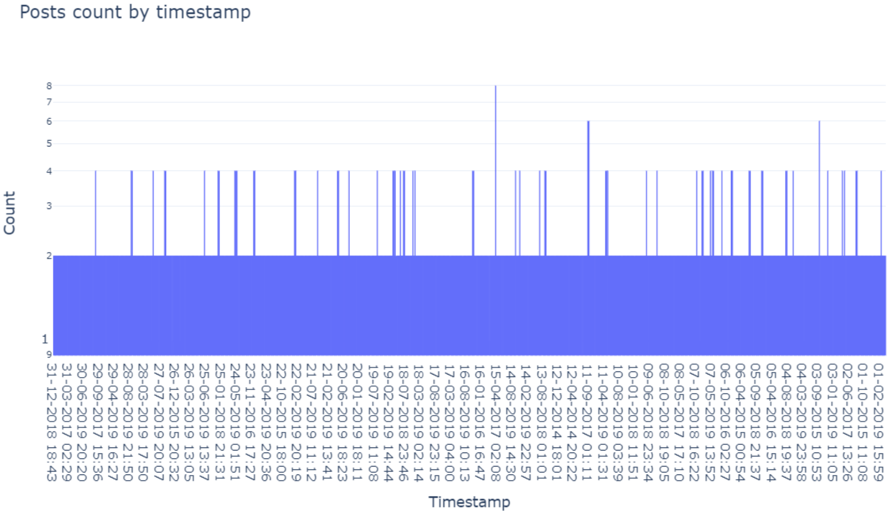


Fig. 9. Posts count versus Timestamp

post belonging to every flair. The upvotes symbolize the number of people who liked the content of the post.

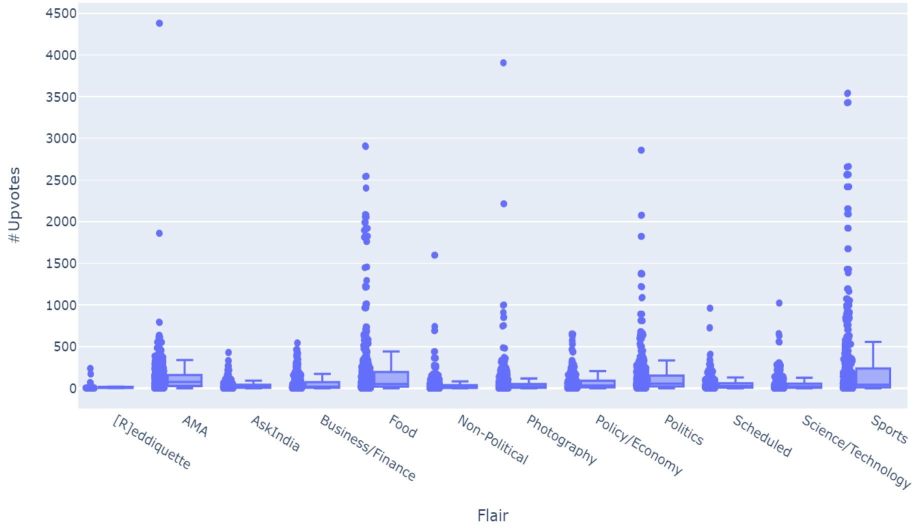


Fig. 10. Flair versus Upvotes

2.3 Data Pre-processing

Data Files Combining-

The data collected using the Reddit API is not ready to be used as the data set as it contains various unwanted texts, characters, etc. It is the need to assemble the collected data. As an example, the comments are bundled together as a single string.

Data Cleaning-

The data cleaning includes the removal of stop words, null spaces, hashtags, bad symbols, URL heads (like https://www), rare words, etc. Data Cleaning also includes converting the upper case letters into the lower case to improvise the results.
The symbols : [/()@,;] are replaced by space (‘ ’).
The symbols : [0-9a-z #+] are replaced by space (‘ ’).
The libraries used are:

1. Nltk[2]
   (a) Stop words
   (b) punkt
2. Beautiful Soap
3. Re

Stopwords are commonly used words such as “the”, “a”, “an”, “in”. These words are to be ignored while working with Natural Language Processing.
re— Regular expression operations
It is a module that facilitates regular expression matching. The re module allows the user to find if a particular set of characters complement the regular expression.

”Beautiful Soap [5] is a python library for pulling data out of HTML and XML files.” This library allows the user to work with various ”parsing” strategies, amending the parse tree and many more such features.

2.4 Data Preparation

Feature Computation

Direct Features: These features are already present in the data set as attributes, so no computation is required as such. The features like Title, URL, Comments, #Comments, Flair, etc are the direct features.

Indirect Features: The features that are developed from direct features. These are the features that are developed by intuition, trials, etc. For example, the number of words or characters in the posts, number of hashtags in the post, etc.

Word Embedding Techniques:
These techniques are most frequently used to get the input text(raw data) in an adjusted format which can be used by various machine learning models and applications like Natural Language Processing, to convert the words into vectors. There are various techniques, depending on the storage of words in vectors, with their own advantages and disadvantages. Few to mention are:

1. CountVectorizer
2. TF-IDF Vectorizer[4]
3. Word2Vec [6]

2.3.1 CountVectorizer

This is one of the most basic word embedding technique that counts the occurrences of words in the document and uses this as the feature extraction model.

The words selected become the features and the number of times each word has occurred in the particular document is made as the feature value. In this technique, the words are represented as columns and the documents are the rows in the matrix of the features.

2.3.2 TF-IDF Vectorizer

"The term TF stands for the term frequency and IDF for Inverse Document Frequency." The concept here is that the word that is frequent in all the documents is not a significant word to be considered in feature extraction model.
This technique functions in a manner, that the word is to be considered as significant only if it is frequent within the document and the frequency of appearing of the word in the number of documents should be low. The TF-IDF score is calculated for each term and the more is the score the more the word is rare and hence significant for use.

2.3.3 Word2Vec

This is a word embedding technique that uses a two-layer neural network model and converts the words into vectors of 32 or more dimensions that can be given as input to the various algorithms of machine learning applied. The CountVectorizer and the TF-IDF Vectorizer do not store the semantic information whereas the Word2Vec preserves the semantic information and relation between different words.
Types of Word2Vec -

1. Continuous Bag Of Words
2. Skip Gram

The continuous bag of words model predicts the current word from the source words.
The skip-gram model predicts the source word from the target word and is opposite of the continuous bag of words.

Example of Continuous Bag Of Words and Skip-gram -

Continuous Bag Of Words :

1. I like to make videos for YouTube (Example Sentence)
2. Target Word - Videos (Output of the neural network)
3. Window Size - 2
4. Context Words - to, make, for, YouTube

Skip-gram : Take the target word and predict the context words

1. Input- Pair with the target word and context word

2. Target Word - Videos (Above Example Considered)
3. Window Size - 2
4. Input- (Videos, YouTube) each pair is formed in this way and passed onto the neural network as input for the model.

The output is a label i.e. either 0 or 1 depending upon the target is in pair with the correct context then 1 otherwise 0 is the output.

2.5 Attribute Selection

The attributes chosen for the analysis do not always contribute to increasing accuracy. Choosing an attribute for analysis can adversely affect the results as well in some cases. That is, the accuracy might increase or decrease by choosing or rejecting any attribute. It is mandatory to choose the attribute combinations wisely in further analysis. The plots are drawn corresponding to various attributes and the flairs, so as to foresee if the attribute is profitable to choose or not. Some of the graphical representations are shown in Fig.11 and Fig. 12 Thus, the results are calculated for different attribute combinations.

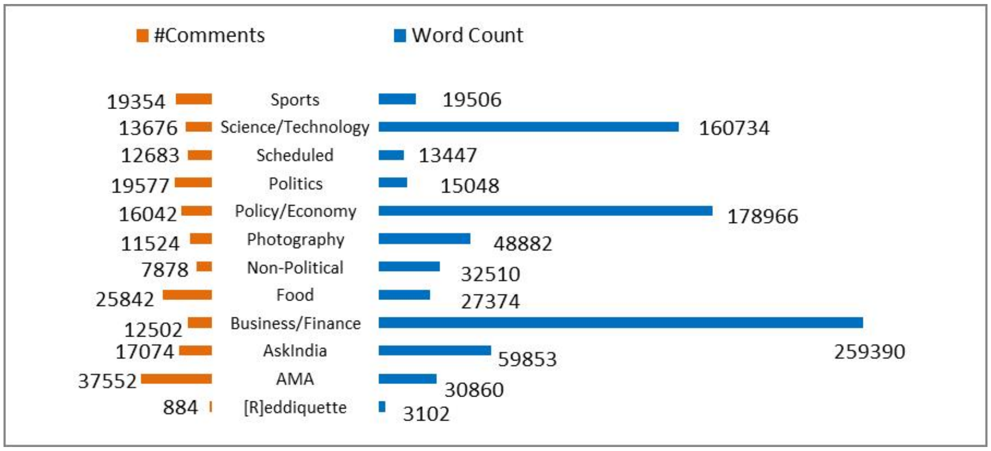


Fig. 11. Flair versus Word Count versus Number of Comments

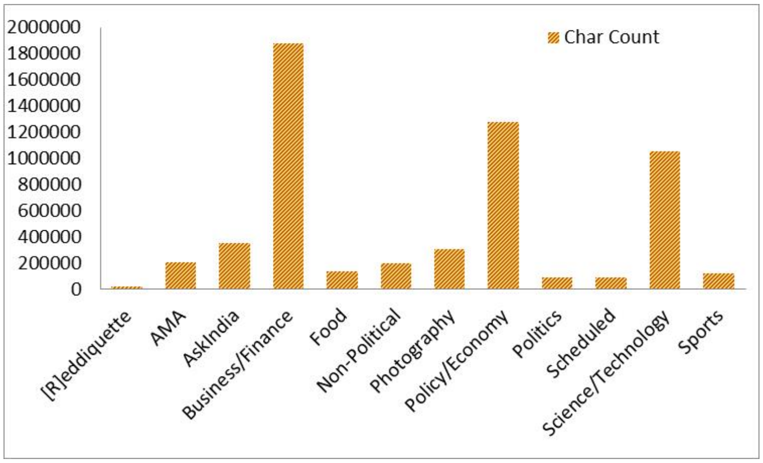


Fig. 12. Flair versus Character Count

2.6 Proposed Approach

The steps for Flair detection of a Reddit post range from collecting data from Reddit, processing the data, applying machine learning algorithms to flair detection of a post.The proposed methodology is illustrated by Fig.13.

3 Machine Learning Algorithms

The data is prepared for the analysis by passing it through data cleaning, preprocessing, and attribute selection phases. Various machine learning algorithms are applied to reach out to the results with maximum accuracy. After applying the machine learning algorithms and finding the most appropriate results, the model is finalized.
In all the algorithms the k-fold cross-validation technique is used to estimate the skill of the model, with the value of k as 10.

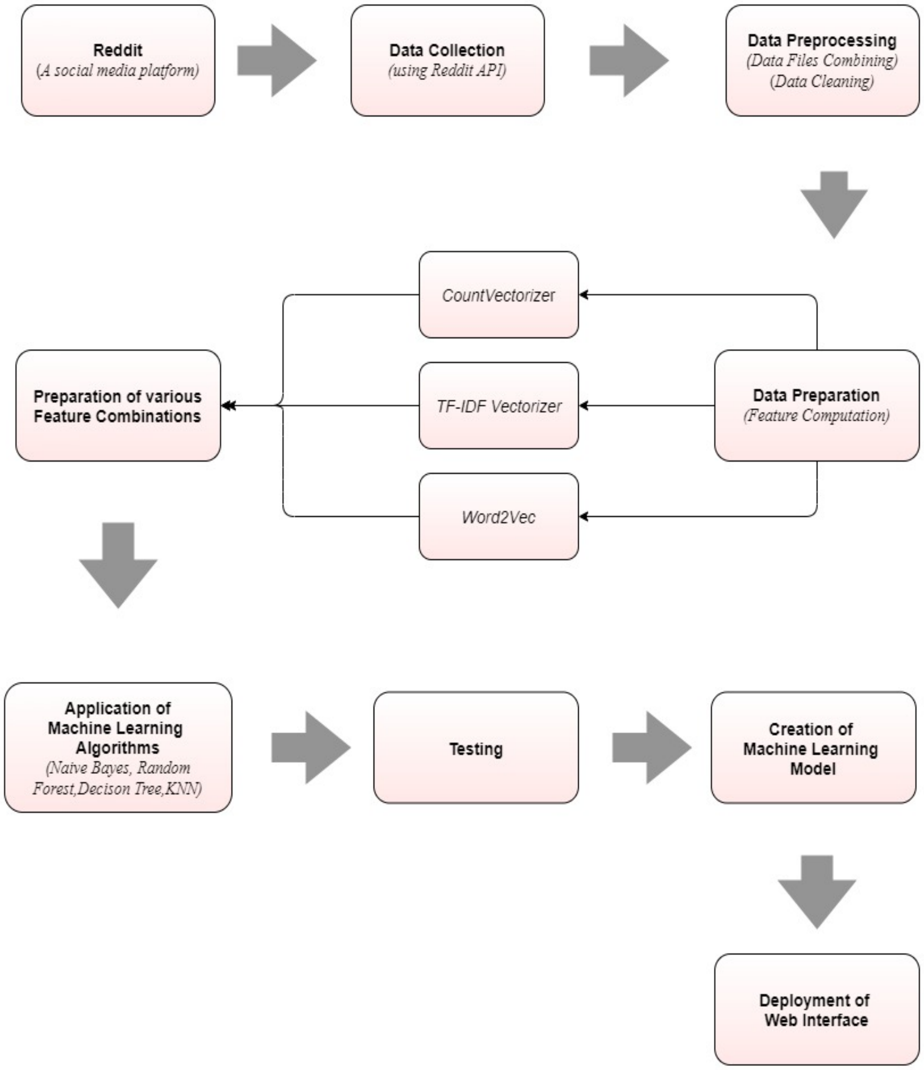


Fig. 13. Proposed Approach

Algorithm 1 ”Decision Tree”

```
procedure D(e)cision Tree[7]
Train the model using the labeled training data set created using Reddit API praw. Decision Tree is a ”recursive greedy algorithm” that at each step determines a point(attribute) of difference/split such that the data points are as segregated as much as possible with each step.
At each step, it uses the concept of ”information gain” (i.e. the amount of information gained by splitting data on basis of a particular attribute)or ”Gini’s Index”(for binary splits) to determine the spitting attribute.
The recursive procedure of tree generation terminates when no data point is left to classified or there is no attribute left on which split can be performed or all data points beyond a level belong to the same class.
As a consequence a tree is generated in which each internal node is an attribute, branches are the values of the node attribute, the leaf signifies the class label. Each path from the root to leaf is a distinct set of attribute values that result in a particular class label. The tree is used to generate if-then rules that are used to classify new data points.

end procedure
```

Algorithm 2 ”KNN”

```
procedure K(N)earest Neighbours [8]
  1. Once data has been acquired by using praw, the value of parameter K is chosen.
  2. For the data entry for which class label here the Reddit flair is to be determined,
     do the following:
       – Determine the distance of the current data point from all other data points
         present in the labeled set.
       – Sort according to distance
       – Consider the ” K nearest neighbors” i.e consider the first k closest data
         points.
       – The label for the current data point is the majority of the K closest data
         point labels.
end procedure
```

Algorithm 3 ”Random Forest”

```
procedure R(a)ndom Forest [1]
It generates multiple decision trees for given training data. Each tree is based on a random sample(say 15 percent of training data is selected randomly for each tree) of the data meant for training, as a result different decision trees are generated. This prevents overfitting in the long run as well. While classifying the given data point classification label is determined by the use of each of these trees. Each tree will provide a label based on its structure and consequent ”if-then rules” out of all the labels generated from the various trees the label of the data point is said to be the label that is predicted by the ”maximum number” of decision trees
end procedure
```

Algorithm 4 ”Logistic Regression”

```
procedure L(o)gistic Regression [9]
    It is a classification technique inspired by linear regression and hence the name. The aim here is to use logistic regression to classify Reddit posts into flairs(class labels). The method makes use of the ”sigmoid function” as it’s the logistic function, which ranges between 0 and 1. The decision boundary obtained as a means to classify can be non-linear unlike linear regression. Gradient Descent can be used to further optimize the model by minimizing the cost function of logistic regression. Logistic Regression can be understood to depict the probability (of data point) belonging to a particular class. A technique could be to determine the probability(of data point) of belonging to each class and the final label would be the class for which probability is maximum.
end procedure
```

Algorithm 5 ”Naive Bayes ”

```
procedure N(a)ive Bayes [10]
    This Algorithm is based on the ”Bayes Theorem”.It is termed as Naive as it assumes that each attribute/feature is independent of the other attribute values, this is often not the case in real-life scenarios. Despite the assumption, it performs equally well as other algorithms and at times even better.
    It is based on the calculation of P(C/X): the probability that the class label is C given the attributes /features of a data point are X. P(c—x) is the posterior probability of class (c, target) given predictor (x, attributes).
    P(c) is the ”prior probability” of class.
    P(x/c) is the equivalent to the number of data points with features/attribute values X that belong to class C
    P(x) is the ”prior probability” of the predictor.
    Then, ”P(c/x)=P(x/c)*P(c)/P(x)”.
    1. For the data entry for which class label here the Reddit flair is to be determined, do the following:
        – For each class label C possible the P(C/X) is determined for the given X(data point features), using P(x/c), P(x), and P(c).
        – The label is then set as the class for which P(c/x) is maximum.
    For continuous values ”Gaussian distribution” is used.
end procedure
```

Algorithm 6 ”Stochastic Gradient Descent”

```
procedure S(t)ochastic Gradient Descent(SDG)
    Gradient Descent is used to find values of parameters such that the corresponding cost function is minimum. Gradient Descent is an iterative technique for optimization that makes use of the entire data set in each iteration. Stochastic Gradient Descent is also an optimization technique, the variation between the two methods is that instead of the entire data set being used in each iteration, a sample of the data set is used. This sample is shuffled and selected randomly. The process used in SDG involves a much nosier path than gradient descent but SDG is computationally more efficient than gradient descent as the sample size is much smaller than the size of the entire data set. SDG was applied to the data set created using Reddit API Praw in an attempt to explore the performance of the algorithm against the rest. SDG supports linear Support Vector Machines(SVM) Support Vector Machine plots the data point in an n-dimensional space in our case n=12 and determines the hyperplane that best differentiates the class and has the maximum margins. In multi-class classification like ours i.e 12 possible labels, it uses one against all approaches.
end procedure
```

4 Results

The results of the study are illustrated in this section.

4.1 Analysis

On the dataset, different data preparation techniques(Word Embedding) like Count Vectorizer, Word2Vec, and TF-IDF are applied. For each data preparation technique, different sets of attributes are selected and multiple algorithms like KNN, Naive Bayes, Decision Tree, Random Forest, Logistic Regression and SDG have been applied to each selection.
The following plots in Fig. 14, Fig. 15 and Fig. 16 summarize the results of CountVectorizer, Word2Vec, TF-ID Vectorizer respectively, that have been obtained by applying different data preparation techniques, selecting different sets of attributes, and by applying different algorithms. The results are illustrated in Table 3.

Table 3. Maximum accuracy results corresponding to each attribute(s) combination.

Vectorizer	Attribute(s) Combination	Accuracy	Algorithm
CountVectorizer	Body	0.5528	Decision Tree
CountVectorizer	Body+Comments+URL	0.5386	Decision Tree
CountVectorizer	Title+Body+Comments	0.9178	Decision Tree
CountVectorizer	Title+URL+Comments	0.9117	SDG
CountVectorizer	Title+Body+URL	0.5656	Decision Tree
CountVectorizer	Title+Comments	0.8908	Random Forest
CountVectorizer	Title+Body	0.5711	Decision Tree
CountVectorizer	Title+URL	0.9285	SDG
CountVectorizer	Body+Comments	0.5406	Decision Tree
CountVectorizer	Body+URL	0.5542	Decision Tree
CountVectorizer	Comments+URL	0.8778	Logistic Regression
Word2Vec	Body	0.538026	Decision Tree
Word2Vec	Body+Comments+URL	0.521179	Random Forest
Word2Vec	Title+Body+Comments	0.520777	Decision Tree
Word2Vec	Title+URL+Comments	0.844329	Random Forest
Word2Vec	Title+Body+URL	0.543844	Random Forest
Word2Vec	Title+Comments	0.844938	Decision Tree
Word2Vec	Title+Body	0.542038	Random Forest
Word2Vec	Title+URL	0.882247	Decision Tree
Word2Vec	Body+Comments	0.519571	Decision Tree
Word2Vec	Body+URL	0.540031	Decision Tree
Word2Vec	Comments+URL	0.842526	Random Forest
TF-IDF	Body	0.560696	Random Forest
TF-IDF	Body+Comments+URL	0.544249	Decision Tree
TF-IDF	Title+Body+Comments	0.551064	Decision Tree
TF-IDF	Title+URL+Comments	0.895889	SDG
TF-IDF	Title+Body+URL	0.56408	Decision Tree
TF-IDF	Title+Comments	0.897292	Random Forest
TF-IDF	Title+Body	0.571125	Decision Tree
TF-IDF	Title+URL	0.931801	SDG
TF-IDF	Body+Comments	0.544048	Decision Tree
TF-IDF	Body+URL	0.561699	Decision Tree
TF-IDF	Comments+URL	0.883256	SDG

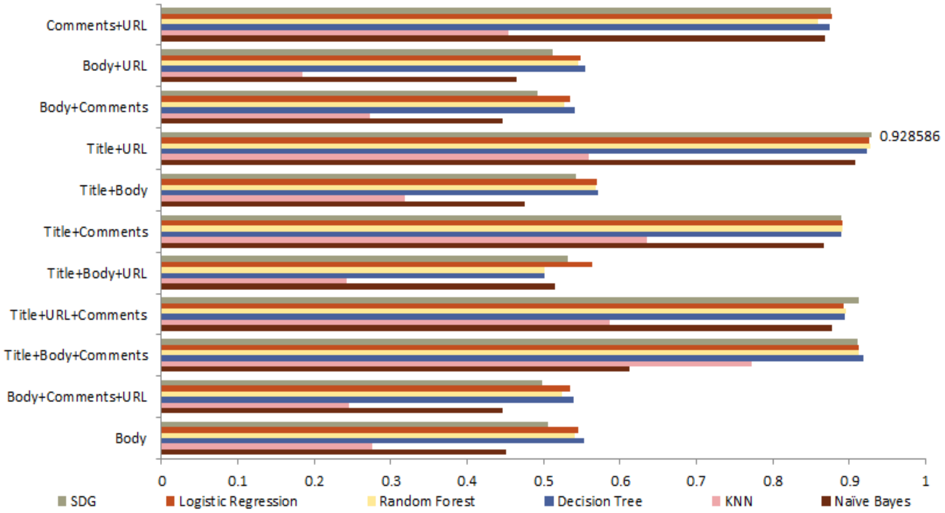


Fig. 14. Results: Count Vectorizer

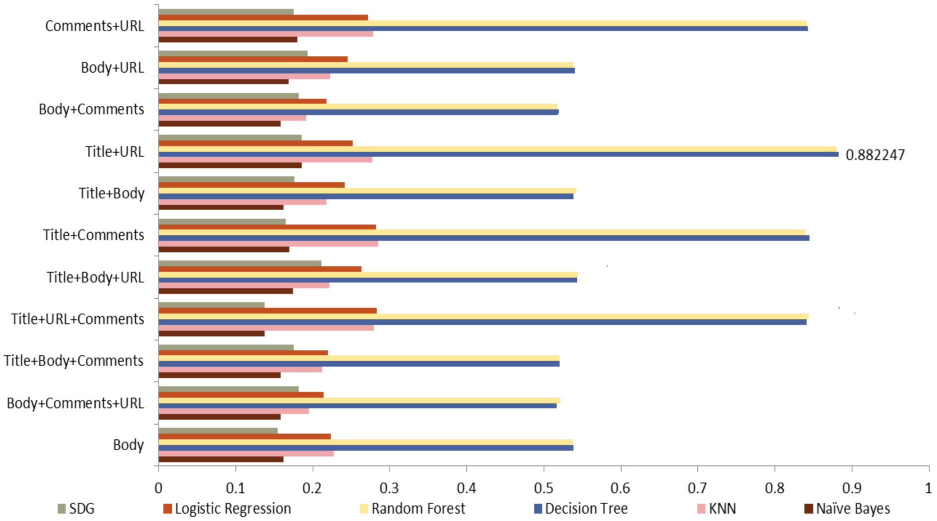


Fig. 15. Results: Word2Vec

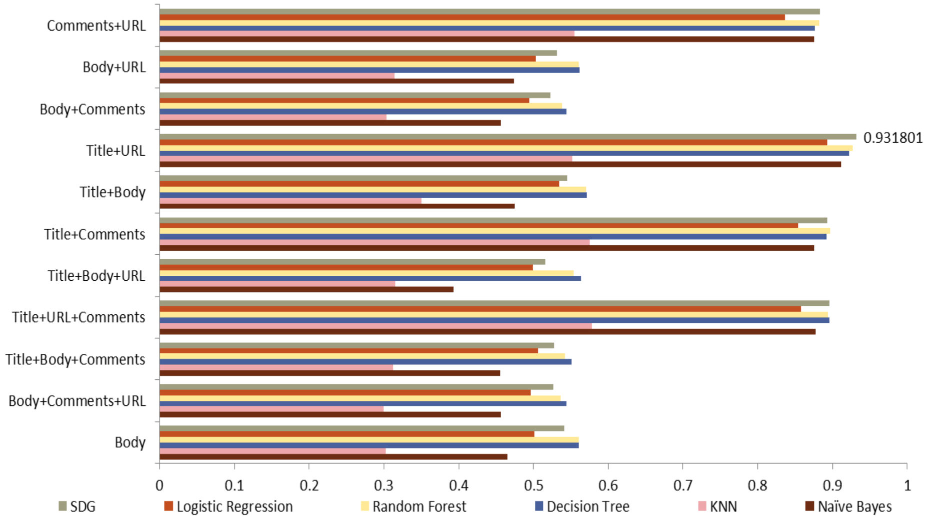


Fig. 16. Results: TF-IDF Vectorizer

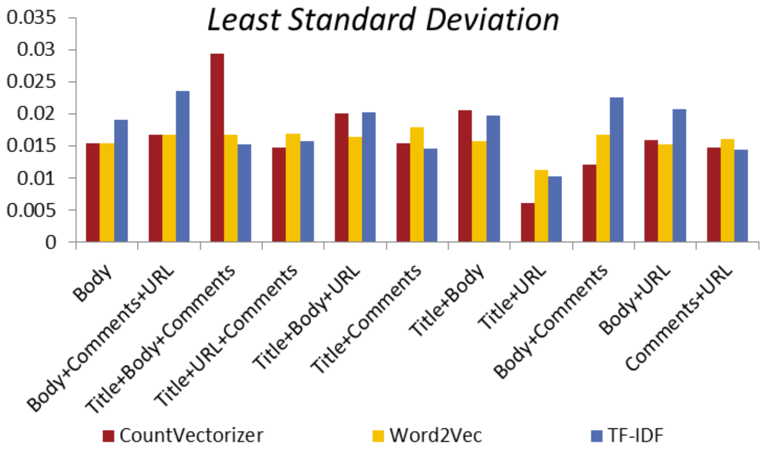
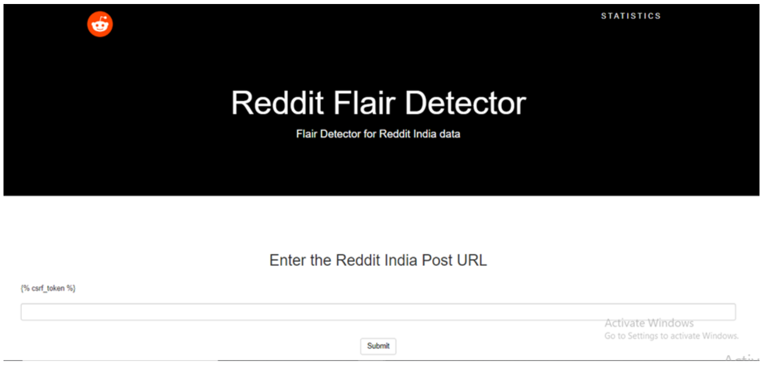


Fig. 17. Least Standard Deviation per attribute combination

Fig. 17 shows that the least standard deviation achieved for each attribute combination in this study.

4.2 Deployment and development of web interface

Deployment of the code and the trained model as a web application is done using Django, HTML, CSS, and JavaScript. The aim is to develop a user interface where the user will input a link to a Reddit post and the model would be used to predict the flair for that post.”Django is a high-level Python Web framework that encourages rapid and clean development, and helps us in linking output produced by the model to the frontend code”. Development of frontend, HTML is used for designing the webpage, CSS, and Bootstrap for styling the web application and JavaScript for making the page dynamic i.e. to enable user interaction. The user enters the URL of a Reddit post whose flair he/she wishes to predict in the input button provided on the webpage and clicks submit.



Reddit Flair Detector

Flair Detector for Reddit India data

Enter the Reddit India Post URL

{% csrf_token %}

Fig. 18. Frontend of the Reddit Flair Detector

After the user clicks the submit button, the input URL is sent to the backend using a POST request. Here, we used POST request instead of GET request because using POST, the input URL won’t get displayed on the screen. HTML forms are used for taking input from the user. CSRF tokens are being used for user verification.

5 Conclusion

On observing different selections and combinations of attribute(s) (Table 3), it is observed that:

1. The best combination amongst all chosen attribute combinations is the integration of Title and URL since most of the algorithms give the best results with this combination.
2. In the case of CountVectorizer, Decision Tree gives the most justifiable results

with the combinations "Title+URL" and "Title+Body+Comments".

It can be inferred that using the TF-IDF Vectorizer, for vectorization of the text, and SDG Classifier supervised learning algorithm for the attribute combination-Title and URL. The most suited explanation for such results is that the "Title", "URL" are the attributes that contain the most relevant and admissible information about any post. Also, these two attributes add to the contribution to the accuracy by providing the relevant content of the post. The accuracy for this result is 93.18% with the standard deviation of 0.0119.
These results are obtained by applying numerous machine learning algorithms with various possible attribute combinations. The accuracy percentage and standard deviation are the parameters that are used to measure the relevance of the study.

6 Future Work

Natural Language Processing is an extremely important tool nowadays, making the tasks effortless. Different validation measures prove that the proposed method returns the considerably accurate result to solve the problem of detecting the flairs of the Reddit posts using Natural Language Processing. The current accuracy, that is, 93.18%, achieved can be the baseline for further exploration. There is a scope of improvement by providing algorithms like advanced Neural Networks in amalgamation with the concepts of Natural Language Processing in order to further enhance the model for public deployment. Additionally, various other attributes can be included in the analysis along with taking various other attribute combinations. In Natural Language Processing, the scope of improvement never ends.

References

1. Jehad Ali, Rehanullah Khan, Nasir Ahmad, and Imran Maqsood. Random forests and decision trees. 2012.
2. Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics.
3. Robin De Pril. User classification based on public reddit data. 2019.
4. Shahzad Qaiser and Ramsha Ali. Text mining: Use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181, 07 2018.
5. Leonard Richardson. Beautiful soup documentation. *April*, 2007.
6. Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.
7. S. Rasoul Safavian and David A. Landgrebe. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.*, 21:660–674, 1991.

8. Raj Shri, Sanjay Gaur, and Prof Chowdhary. Text classification using knn with different feature selection methods. 08 2018.
9. Chao ying Peng, Kuk Lida Lee, and Gary M. Ingersoll. An introduction to logistic regression analysis and reporting. 2002.
10. Harry Zhang. The optimality of naive bayes. *AA*, 1(2):3, 2004.