Team Members:

Sri Vishnu. M : srivishnu1103@gmail.com (Team Leader)

Sathis Kumar. G : g.sathiskumar22@gmail.com

Sanjay Kumar. R : rsanjaykumar1319@gmail.com

Mohammed Noorulla :noorullasnoorul123@gmail.com

Fake News Detection

Fake News Detection in Python:

In this machine learning project, we build a classifier that detects whether the news is fake or not.

This is a binary classification problem. We preprocess the text data from our dataset using TF-IDF Vectorizer. We apply the Multinomial Naive Bayes algorithm to the preprocessed text and train and evaluate our model on the dataset.

In this project, we have used various natural language processing techniques and machine learning algorithms to classify fake news articles using sci-kit libraries from python.

Dataset used:

The data source used for this project is LIAR dataset which contains 3 files with .tsv format for test, train and validation. Below is some description about the data files used for this project.

LIAR: A BENCHMARK DATASET FOR FAKE NEWS DETECTION

William Yang Wang, "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection, to appear in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), short paper, Vancouver, BC, Canada, July 30-August 4, ACL.

the original dataset contained 13 variables/columns for train, test and validation sets as follows:

- Column 1: the ID of the statement ([ID].json).
- Column 2: the label. (Label class contains: True, Mostly-true, Half-true, Barely-true, FALSE, Pants-fire)

- Column 3: the statement.
- Column 4: the subject(s).
- Column 5: the speaker.
- Column 6: the speaker's job title.
- Column 7: the state info.
- Column 8: the party affiliation.
- Column 9-13: the total credit history count, including the current statement.
- 9: barely true counts.
- 10: false counts.
- 11: half true counts.
- 12: mostly true counts.
- 13: pants on fire counts.
- Column 14: the context (venue / location of the speech or statement).

To make things simple we have chosen only 2 variables from this original dataset for this classification. The other variables can be added later to add some more complexity and enhance the features.

Below are the columns used to create 3 datasets that have been in used in this project

- Column 1: Statement (News headline or text).
- Column 2: Label (Label class contains: True, False)

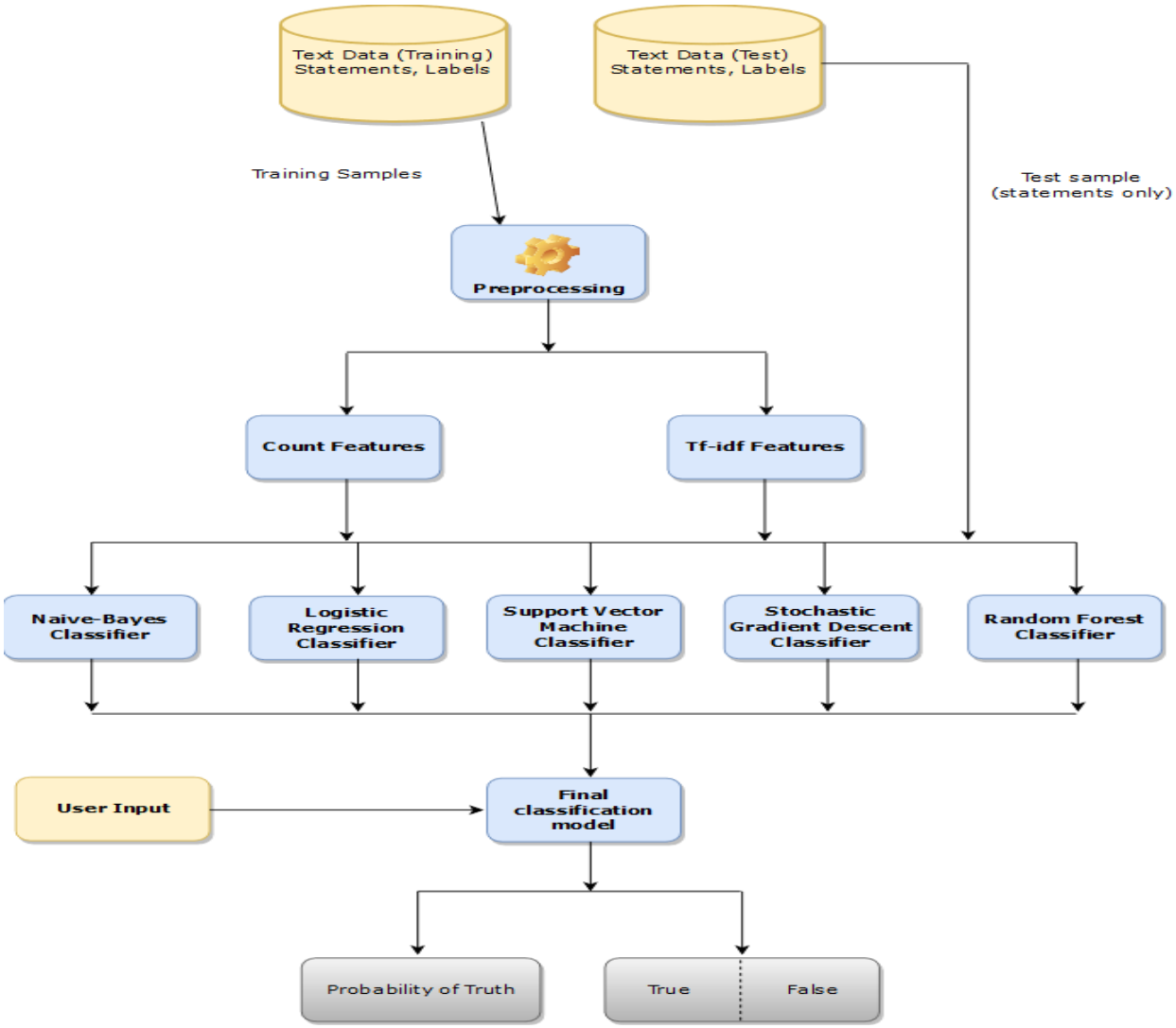
You will see that newly created dataset has only 2 classes as compared to 6 from original classes. Below is method used for reducing the number of classes.

- Original -- New
- True -- True
- Mostly-true -- True
- Half-true -- True

- Barely-true -- False
- False -- False
- Pants-fire -- False

The dataset used for this project were in csv format named train.csv, test.csv and valid.csv and can be found in repo. The original datasets are in "liar" folder in tsv format.

Below is the Process Flow of the project:



Installing and steps to run the software:

If you have chosen to install python (and already setup PATH variable for `python.exe`) then follow instructions:

Open the command prompt and change the directory to project folder as mentioned in above by running below command

```
cd C:/your cloned project folder path goes here/
```

run below command

```
python.exe C:/your cloned project folder path goes here/
```

After hitting the enter, program will ask for an input which will be a piece of information or a news headline that you want to verify. Once you paste or type news headline, then press enter.

Once you hit the enter, program will take user input (news headline) and will be used by model to classify in one of categories of "True" and "False". Along with classifying the news headline, model will also provide a probability of truth associated with it.