NEWS CLASSIFIER

Members: Mahsa Razavi – Mahsa Anvarian

Student ID: 94523144 – 95521054

Professor: Dr. Pilevar

Project github link: https://github.com/mahsawz/AI-NewsClassifier

Iran University of Science and Technology

INTRODUCTION

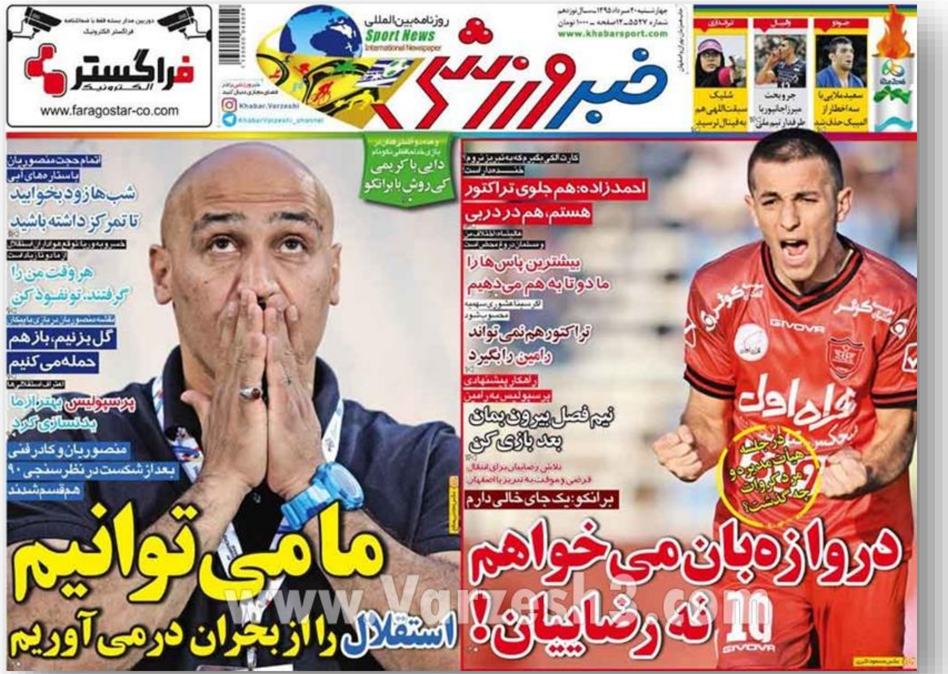
- **One of the widely used natural language processing task in different business problems is "Text Classification".**
- **Automatically classify the text documents into one or more defined categories.**

➢ Understanding audience sentiment from social media,

➢ Detection of spam and non-spam emails,

➢ Categorization of news articles into defined topics.

NEWS CLASSIFICATION



فراگستر الکترونیک
www.faragostar-co.com

روزنامه بین المللی
Sport News
International Newspaper
خبر ورزشی
www.khabarsport.com

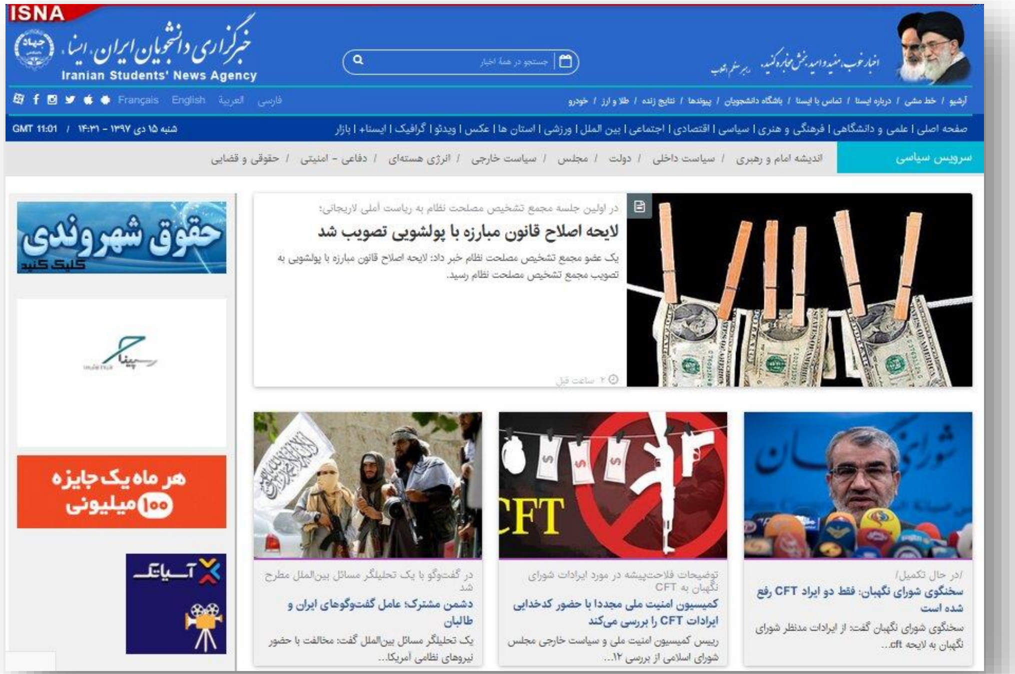
سعید ملایی با سه اخطار از المپیک حذف شد
جروبحث میرزاجانپور با طرفدار تیم ملی
شلیک سبقت اللهی هم به فینال نرسید

دایی با کریمی کی روش با برانکو

اتمام حجت منصوریان با ستاره های آبی
شب ها زود بخوابید تا تمرکز داشته باشید
هر وقت من را گرفتند، تو نفوذ کن
گل بزنیم، باز هم حمله می کنیم
پرسپولیس بهتر از ما بدنسازی کرد
منصوریان و کادر فنی بعد از شکست در نظرسنجی ۹۰ هم قسم شدند
ما می توانیم
استقلال را از بحران درمی آوریم

احمدزاده: هم جلوی تراکتور هستم، هم در دربی
بیشترین پاس ها را ما دو تا به هم می دهیم
تراکتور هم نمی تواند رامین را بگیرد
راهکار پیشنهادی پرسپولیس به رامین
نیم فصل بیرون بمان بعد بازی کن
برانکو: یک جای خالی دارم
در جلسه هیات مدیره و مزد کروات چه گذشت؟
دروازه بان می خواهم نه رضاییان!

vs



ISNA
خبرگزاری دانشجویان ایران، ایسنا
Iranian Students' News Agency

جستجو در همهٔ اخبار

اخبار خوب، مفید و امید بخش مخابره کنید. رهبر معظم انقلاب

فارسی | العربية | English | Français

آرشیو / خط مشی / درباره ایسنا / تماس با ایسنا / باشگاه دانشجویان / پیوندها / نتایج زنده / طلا و ارز / خودرو

صفحه اصلی | علمی و دانشگاهی | فرهنگی و هنری | سیاسی | اقتصادی | اجتماعی | بین الملل | ورزشی | استان ها | عکس | ویدئو | گرافیک | ایسنا+ | بازار

شنبه ۱۵ دی ۱۳۹۷ - ۱۴:۳۱ / GMT 11:01

سرویس سیاسی
اندیشه امام و رهبری / سیاست داخلی / دولت / مجلس / سیاست خارجی / انرژی هسته‌ای / دفاعی - امنیتی / حقوقی و قضایی

در اولین جلسه مجمع تشخیص مصلحت نظام به ریاست آملی لاریجانی:
لایحه اصلاح قانون مبارزه با پولشویی تصویب شد
یک عضو مجمع تشخیص مصلحت نظام خبر داد: لایحه اصلاح قانون مبارزه با پولشویی به تصویب مجمع تشخیص مصلحت نظام رسید.
۲ ساعت قبل

حقوق شهروندی
کلیک کنید

هر ماه یک جایزه ۱۰۰ میلیونی

آسیاتک

/در حال تکمیل/
سخنگوی شورای نگهبان: فقط دو ایراد CFT رفع شده است
سخنگوی شورای نگهبان گفت: از ایرادات مدنظر شورای نگهبان به لایحه cft...

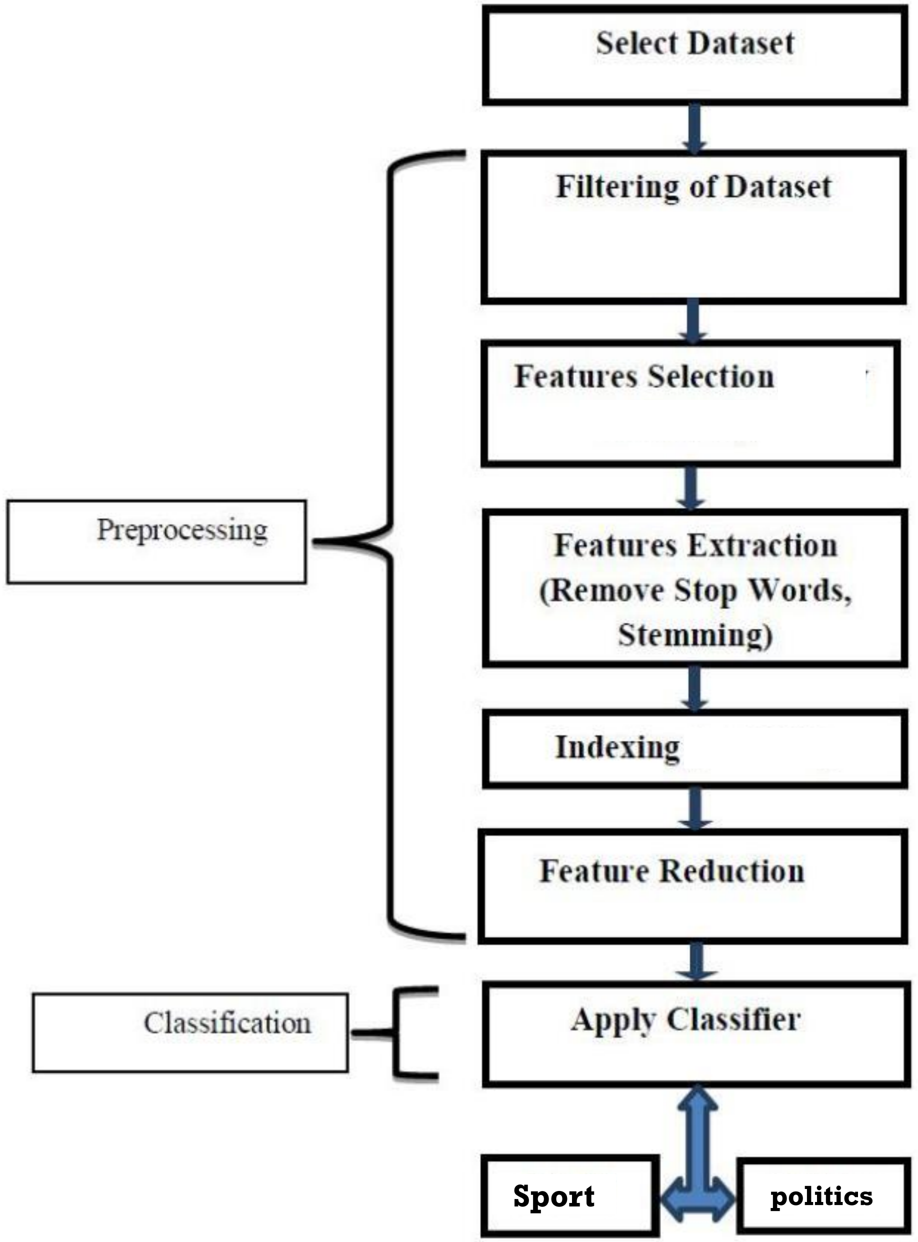
توضیحات فلاحت‌پیشه در مورد ایرادات شورای نگهبان به CFT
کمیسیون امنیت ملی مجددا با حضور کدخدایی ایرادات CFT را بررسی می‌کند
رییس کمیسیون امنیت ملی و سیاست خارجی مجلس شورای اسلامی از بررسی ۱۲...

در گفت‌وگو با یک تحلیلگر مسائل بین‌الملل مطرح شد
دشمن مشترک؛ عامل گفت‌وگوهای ایران و طالبان
یک تحلیلگر مسائل بین‌الملل گفت: مخالفت با حضور نیروهای نظامی آمریکا...

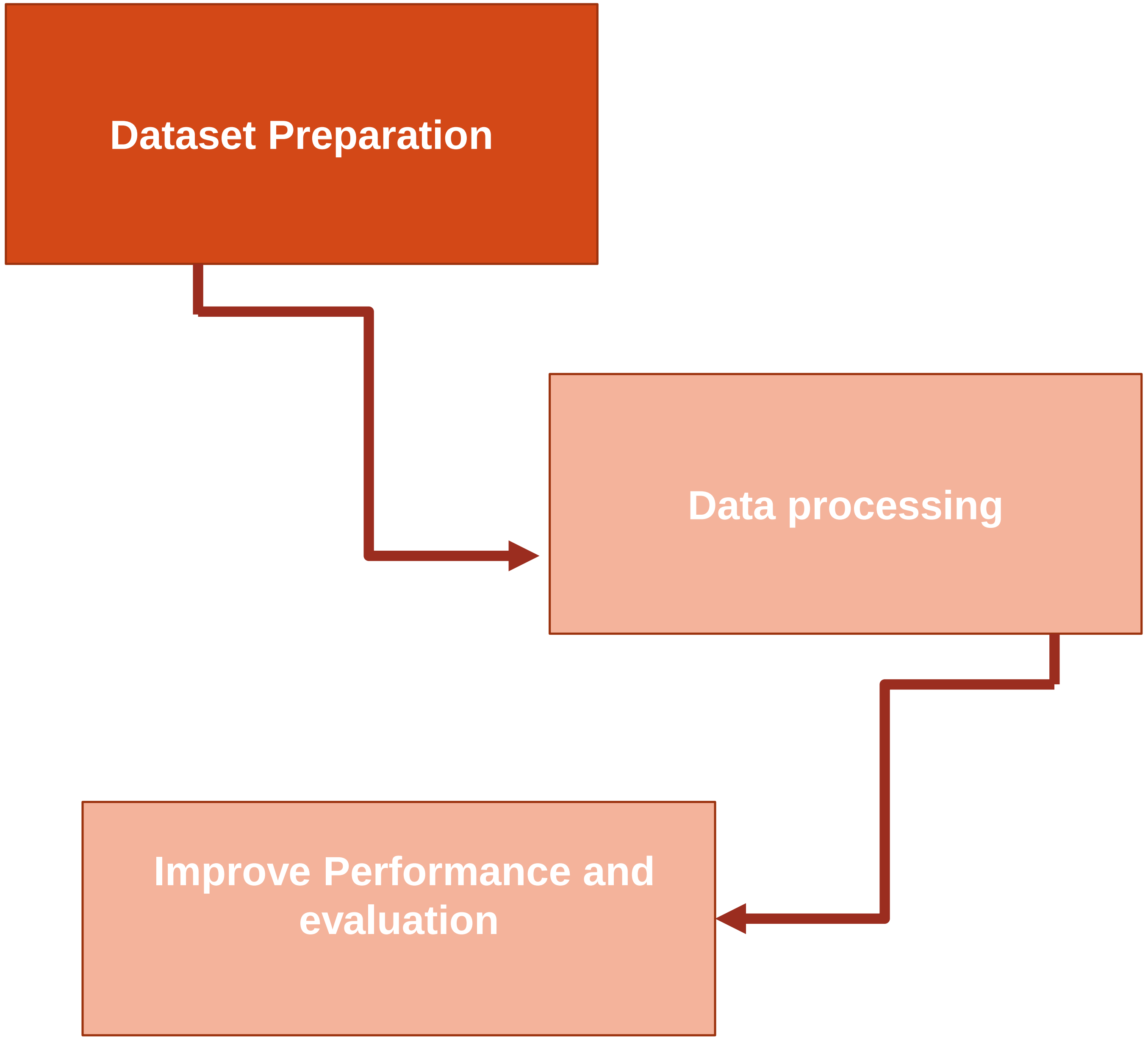
sport

politics

Main Steps


Select Dataset
Filtering of Dataset
Features Selection
Features Extraction (Remove Stop Words, Stemming)
Indexing
Feature Reduction
Apply Classifier
Preprocessing
Classification
Sport
politics

MAIN STEPS


Dataset Preparation
Data processing
Improve Performance and evaluation

1. DATASET PREPARATION

- **BBC News Dataset**

The dataset consists of text reviews and their labels which can be downloaded at this link.

Class	Words
Sport	89001
Politics	99335

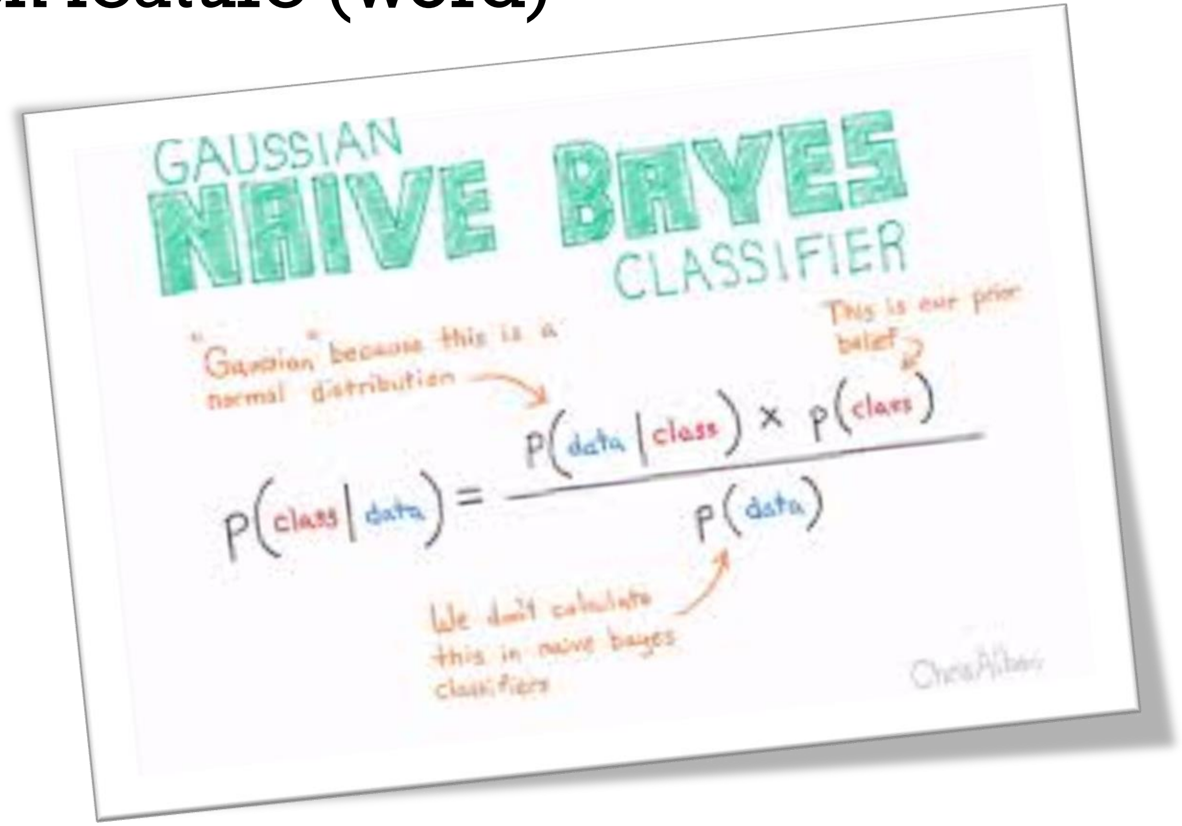
1. DATASET PREPARATION

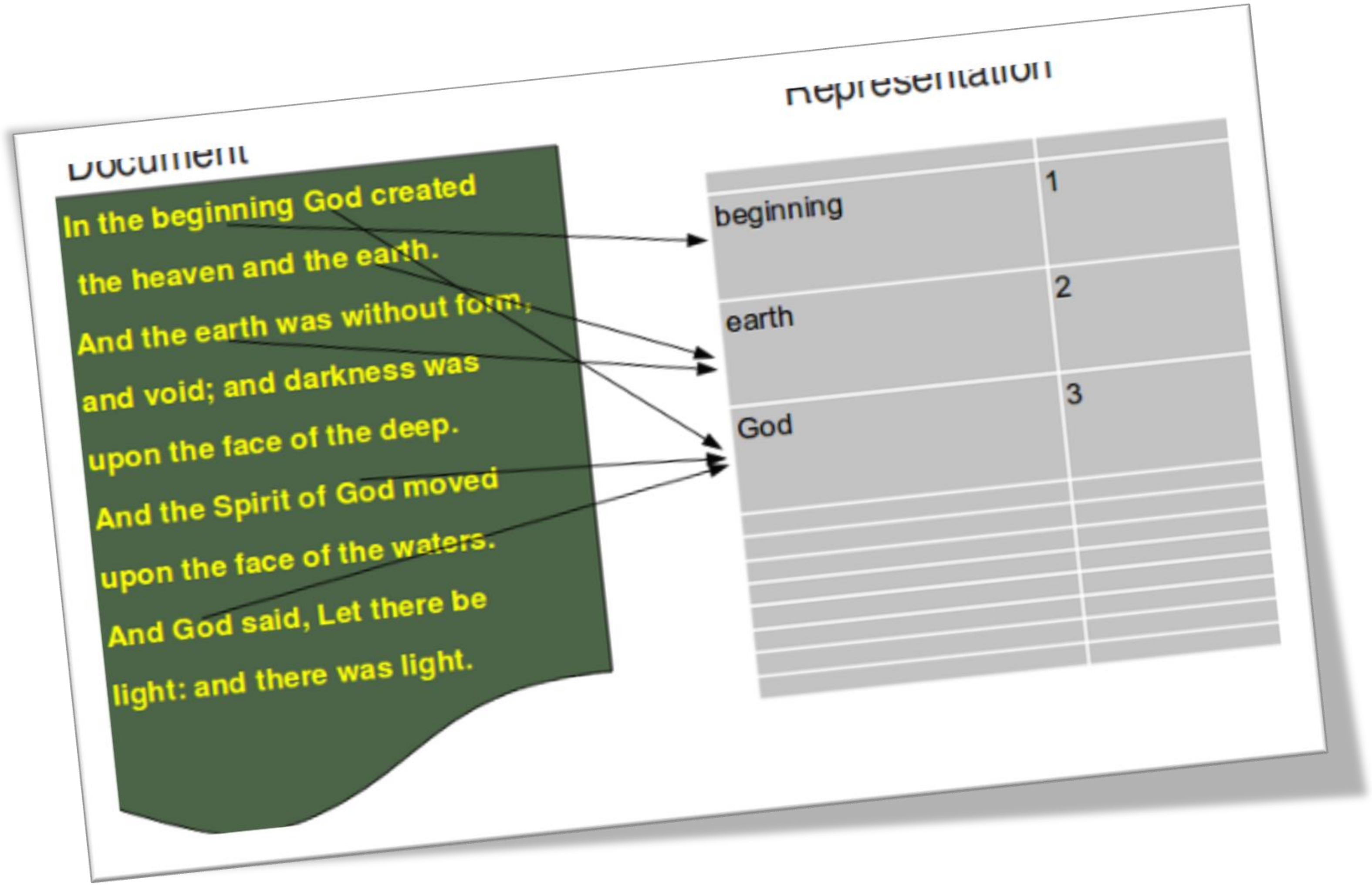
- **Text Cleaning**

- Lower Case
- Removing Digits
- Removing Punctuation
- Removal of Stop Words
- Tokenization
- Stemming
- Lemmatization
- …

2. NAÏVE BAYES

- Compute the probability of a label in condition of given features
- We need to specify how each feature (word) depends on the class

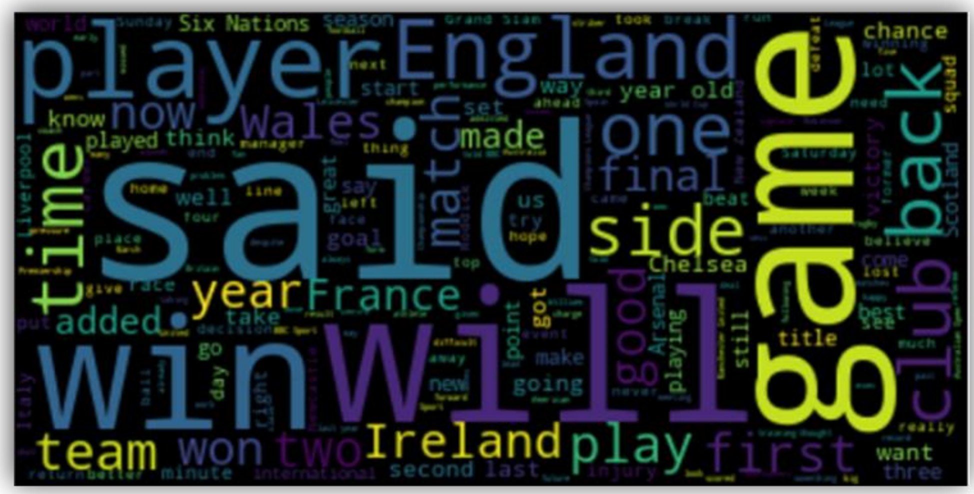

GAUSSIAN
NAIVE BAYES
CLASSIFIER
"Gaussian" because this is a normal distribution
This is our prior belief
p(class | data) = p(data | class) × p(class) / p(data)
We don't calculate this in naive bayes classifiers
ChrisAlbon


Document
Representation
In the beginning God created
the heaven and the earth.
And the earth was without form,
and void; and darkness was
upon the face of the deep.
And the Spirit of God moved
upon the face of the waters.
And God said, Let there be
light: and there was light.
beginning
1
earth
2
God
3

Bag
of
Words

3. EVALUATING RESULTS

- Word Clouds



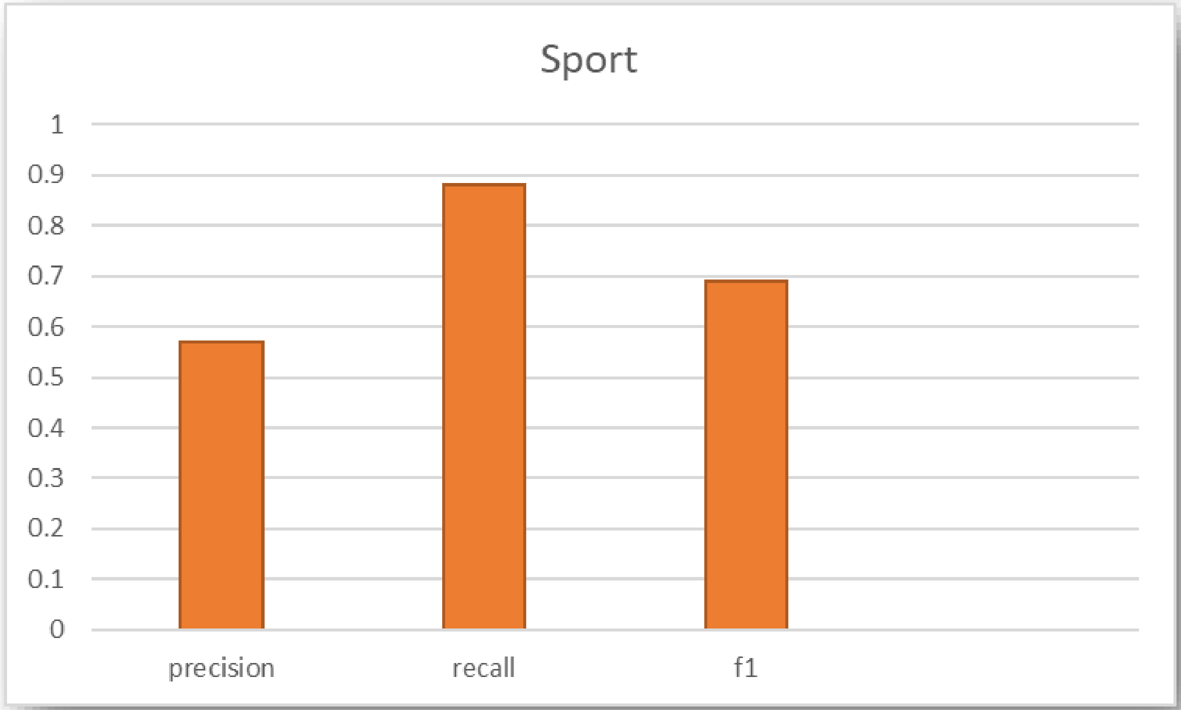
said
will
government
party
Mr Blair
Labour
people
say
new
UK
plan
one
year
Britain
public
election
Mr Brown
Tony Blair
issue
report
prime minister
Tories
Conservative
Michael Howard
Lib Dem

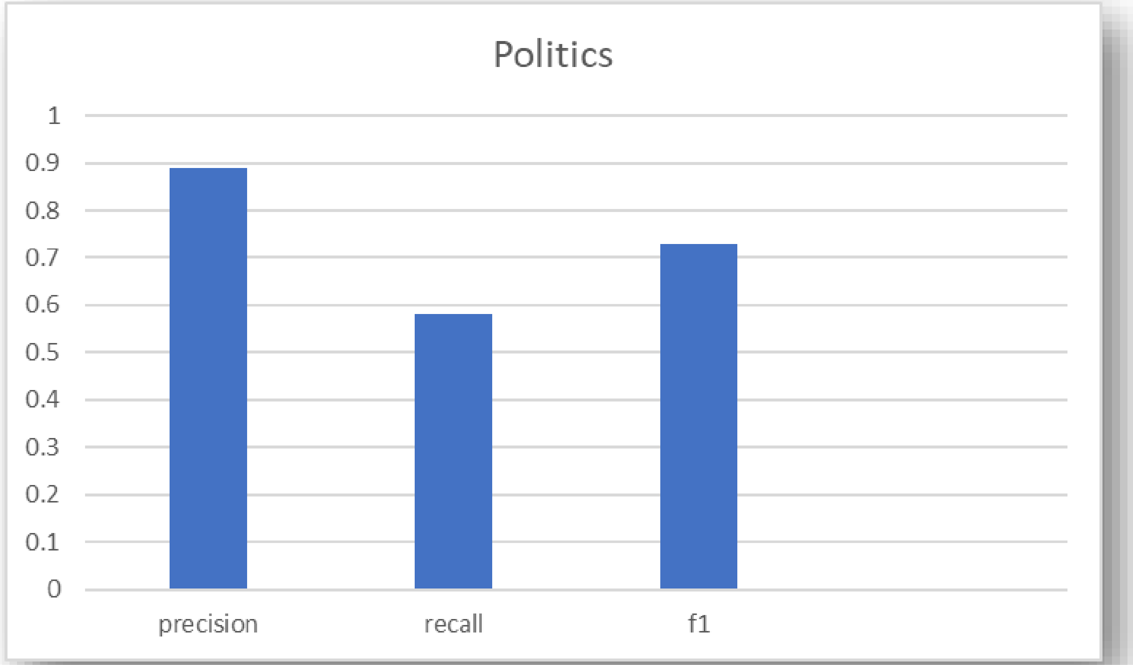
3. EVALUATING RESULTS

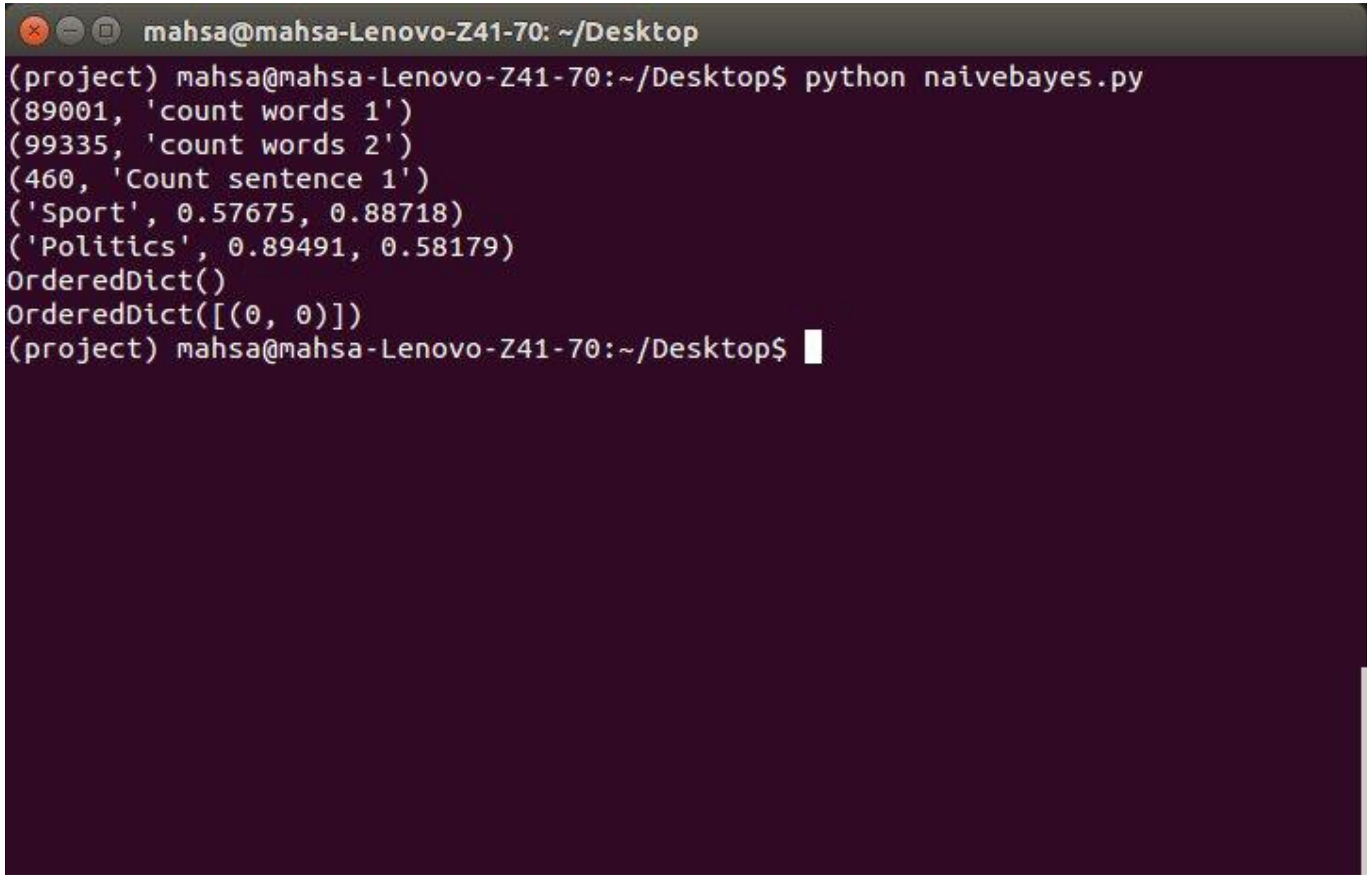
Test data : 10% of dataset devoted for testing

- Precision
- Recall
- F-score
- Accuracy

3. Evaluating results







```
(project) mahsa@mahsa-Lenovo-Z41-70:~/Desktop$ python naivebayes.py
(89001, 'count words 1')
(99335, 'count words 2')
(460, 'Count sentence 1')
('Sport', 0.57675, 0.88718)
('Politics', 0.89491, 0.58179)
OrderedDict()
OrderedDict([(0, 0)])
(project) mahsa@mahsa-Lenovo-Z41-70:~/Desktop$
```

Naïve Bayes Summary

Advantages:

- Fast to train (single scan through data)
- Fast to classify

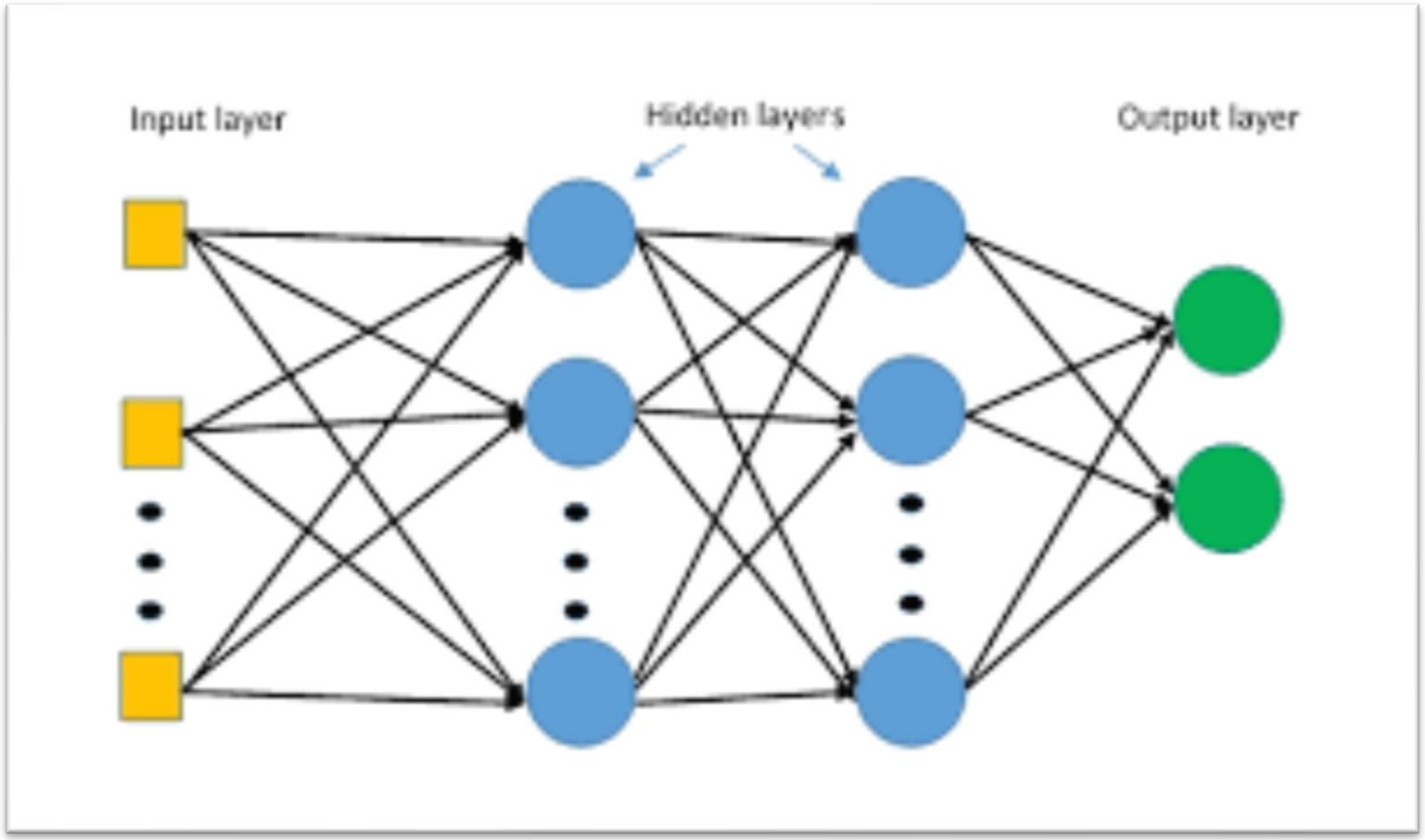
Disadvantages:

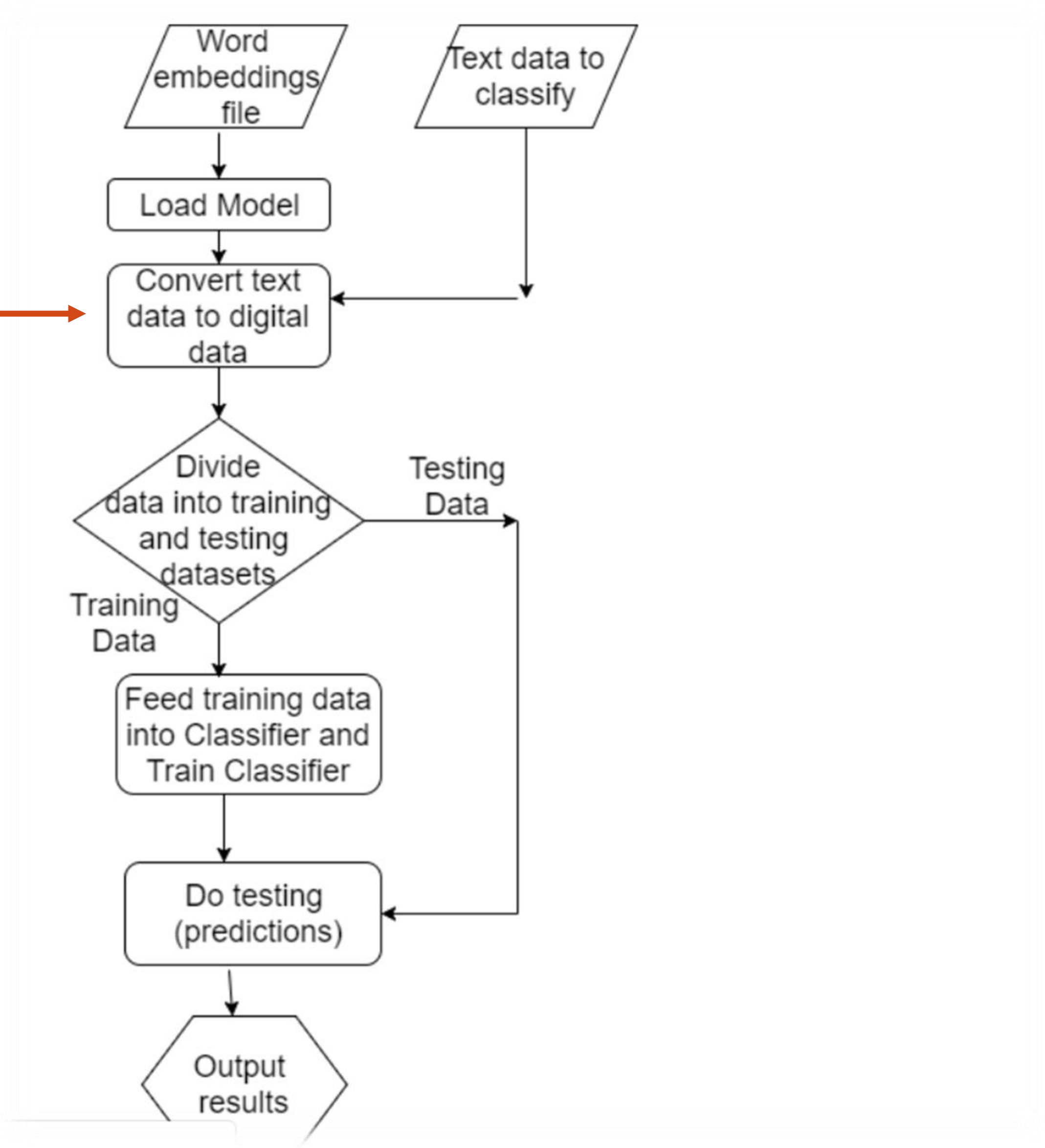
- Assumes independence of features

CONCLUSION

MLP (MULTILAYER PERCEPTRON)

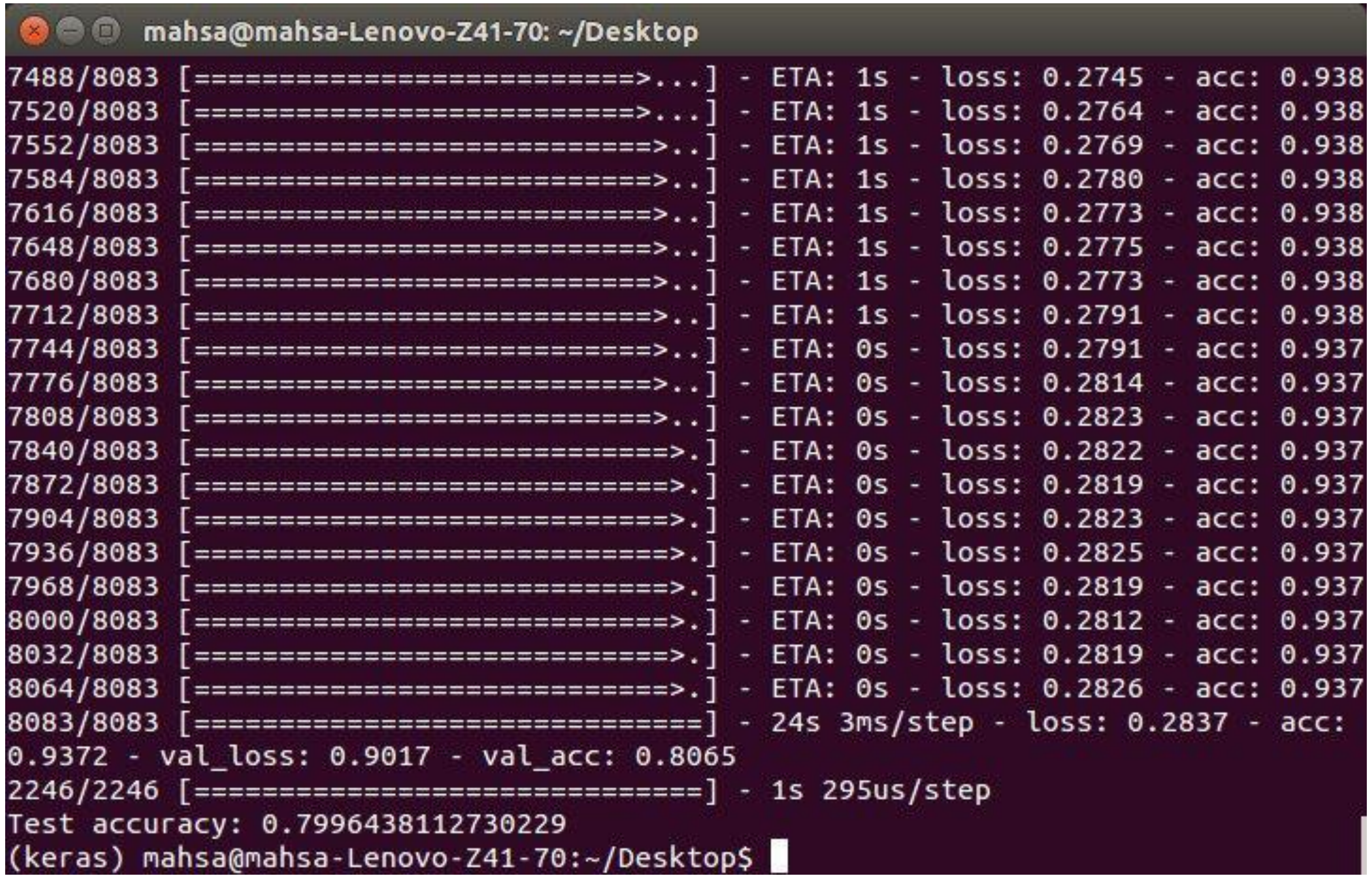
Second Approach


Input layer
Hidden layers
Output layer


Word embeddings file
Text data to classify
Load Model
Convert text data to digital data
Divide data into training and testing datasets
Testing Data
Training Data
Feed training data into Classifier and Train Classifier
Do testing (predictions)
Output results

Main Steps

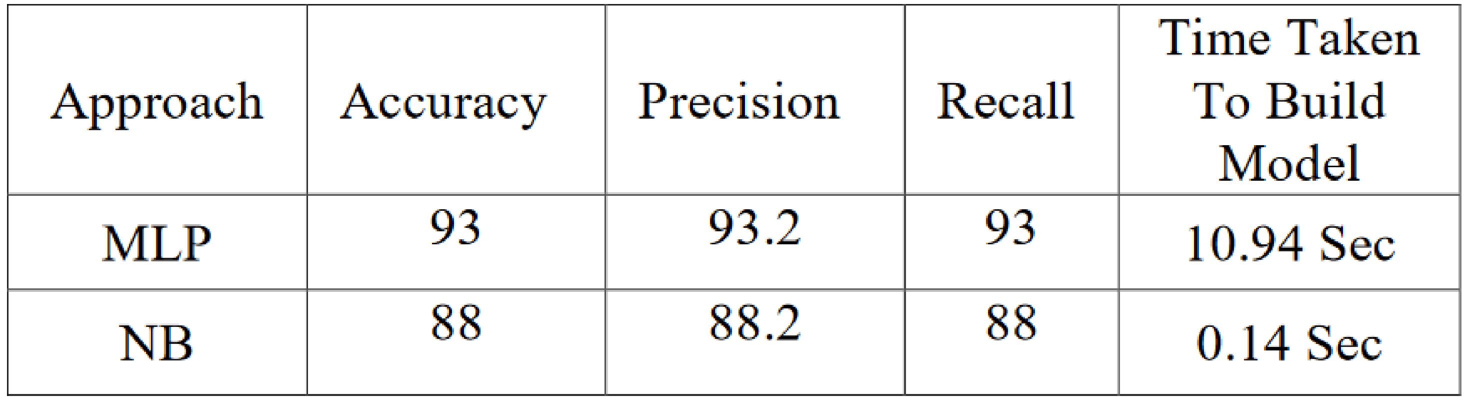
Total Accuracy 0.799


mahsa@mahsa-Lenovo-Z41-70: ~/Desktop

```
7488/8083 [==========================>...] - ETA: 1s - loss: 0.2745 - acc: 0.938
7520/8083 [==========================>...] - ETA: 1s - loss: 0.2764 - acc: 0.938
7552/8083 [===========================>..] - ETA: 1s - loss: 0.2769 - acc: 0.938
7584/8083 [===========================>..] - ETA: 1s - loss: 0.2780 - acc: 0.938
7616/8083 [===========================>..] - ETA: 1s - loss: 0.2773 - acc: 0.938
7648/8083 [===========================>..] - ETA: 1s - loss: 0.2775 - acc: 0.938
7680/8083 [===========================>..] - ETA: 1s - loss: 0.2773 - acc: 0.938
7712/8083 [===========================>..] - ETA: 1s - loss: 0.2791 - acc: 0.938
7744/8083 [===========================>..] - ETA: 0s - loss: 0.2791 - acc: 0.937
7776/8083 [===========================>..] - ETA: 0s - loss: 0.2814 - acc: 0.937
7808/8083 [===========================>..] - ETA: 0s - loss: 0.2823 - acc: 0.937
7840/8083 [============================>.] - ETA: 0s - loss: 0.2822 - acc: 0.937
7872/8083 [============================>.] - ETA: 0s - loss: 0.2819 - acc: 0.937
7904/8083 [============================>.] - ETA: 0s - loss: 0.2823 - acc: 0.937
7936/8083 [============================>.] - ETA: 0s - loss: 0.2825 - acc: 0.937
7968/8083 [============================>.] - ETA: 0s - loss: 0.2819 - acc: 0.937
8000/8083 [============================>.] - ETA: 0s - loss: 0.2812 - acc: 0.937
8032/8083 [============================>.] - ETA: 0s - loss: 0.2819 - acc: 0.937
8064/8083 [============================>.] - ETA: 0s - loss: 0.2826 - acc: 0.937
8083/8083 [==============================] - 24s 3ms/step - loss: 0.2837 - acc:
0.9372 - val_loss: 0.9017 - val_acc: 0.8065
2246/2246 [==============================] - 1s 295us/step
Test accuracy: 0.7996438112730229
(keras) mahsa@mahsa-Lenovo-Z41-70:~/Desktop$
```

MLP VS NB ?

- Acciracy
- Precision
- Recall
- Time taken to build model



Approach	Accuracy	Precision	Recall	Time Taken To Build Model
MLP	93	93.2	93	10.94 Sec
NB	88	88.2	88	0.14 Sec

Performance measurement of both classifiers

REFERENCES

- https://pdfs.semanticscholar.org/d945/29ae612b76287c7dcdccd5bf09f3c5e772af.pdf
- https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/
- https://www.saedsayad.com/naive_bayesian.htm
- https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/
- https://medium.com/@minbaekim/text-mining-preprocess-and-naive-bayes-classifier-da0000f633b2
- https://www.python-course.eu/text_classification_python.php
- https://www.datacamp.com/community/tutorials/wordcloud-python
- http://ai.intelligentonlinetools.com/ml/fasttext-word-embeddings-text-classification-python-mlp/