Internship Report

on

Natural Language Processing (NLP) Data Analysis for low resource Indo-European languages using Python

Submitted in partial fulfilment of the requirements for the award of the degree

of

Bachelor of Technology

in

COMPUTER SCIENCE AND ENGINEERING

by

Mr. Abhishek Kumar (170101003)

Mr. Himanshu Ranjan (170101017)

Mr. Suraj Kumar (170101052)

भारतीय सूचना प्रौद्योगिकी संस्थान भागलपुर
Indian Institute of Information Technology
Bhagalpur

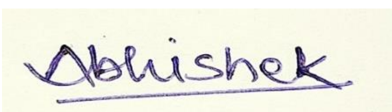
Department of Computer Science and Engineering

Indian Institute of Information Technology Bhagalpur

June, 2021

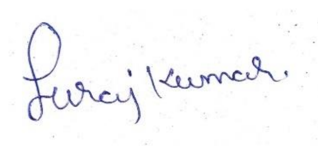
DECLARATION

We hereby declare that the work reported in this project on the topic "***Natural Language Processing (NLP) Data Analysis for low resource Indo-European languages using Python***" has been carried out by us independently in the **Department of Computer Science and Engineering, IIIT Bhagalpur** under the guidance of **Mr. Ajay Kumar Mishra**, Mentor, Yscholar Technology LLP. We also declare that this work has not formed the basis for the award of any other Degree, Diploma, or similar title of any university or institution.



Mr. Abhishek Kumar
(170101003)

Mr. Himanshu Ranjan
(170101017)



Mr. Suraj Kumar
(170101052)

YSCHOLAR TECHNOLOGY LLP

9061, SOBHA DREAM ACRES, PANATHUR ROAD, SOBHA
DREAM BALAGERE VILLAGE BANGALORE Bangalore KA 560087

INTERNSHIP COMPLETION LETTER

To whomsover this may concern

This is to certify that **Mr. Abhishek Kumar** a student of IIIT Bhagalpur successfully completed the 4 month internship at Yscholar Technology LLP. in Software development from Jan 4th 2021-April 30th 2021.

During the internship he worked on **NLP Data analysis for low resource Indo-European languages** using Python programming language.

During the internship we found **Mr. Abhishek Kumar** to be sincere, hard-working and a quick learner. We wish him all the best in his future endeavors.

Sincerely,

Vijay Mishra, Director
Yscholar Technology LLP
9061, SOBHA DREAM ACRES, PANATHUR ROAD, SOBHA DREAM BALAGERE VILLAGE BANGALORE Bangalore KA 560087
Tel: (+91) 998-674-6745

Date Issued: May 1st, 2021

YSCHOLAR

YSCHOLAR TECHNOLOGY LLP

9061, SOBHA DREAM ACRES, PANATHUR ROAD, SOBHA DREAM BALAGERE VILLAGE BANGALORE Bangalore KA 560087

INTERNSHIP COMPLETION LETTER

To whomsover this may concern

This is to certify that **Mr. Himanshu Ranjan** a student of IIIT Bhagalpur successfully completed the 4 month internship at Yscholar Technology LLP. in Software development from Jan 4th 2021-April 30th 2021.

During the internship he worked on **NLP Data analysis for low resource Indo-European languages** using Python programming language.

During the internship we found **Mr. Himanshu Ranjan** to be sincere, hard-working and a quick learner. We wish him all the best in his future endeavors.

Sincerely,

Vijay Mishra

Vijay Mishra, Director
Yscholar Technology LLP
9061, SOBHA DREAM ACRES, PANATHUR ROAD, SOBHA DREAM BALAGERE VILLAGE BANGALORE Bangalore KA 560087
Tel: (+91) 998-674-6745

Date Issued: May 1st, 2021

YSCHOLAR

YSCHOLAR TECHNOLOGY LLP

9061, SOBHA DREAM ACRES, PANATHUR ROAD, SOBHA
DREAM BALAGERE VILLAGE BANGALORE Bangalore KA 560087

INTERNSHIP COMPLETION LETTER

To whomsover this may concern

This is to certify that **Mr. Suraj Kumar** a student of IIIT Bhagalpur successfully completed the 4 month internship at Yscholar Technology LLP. in Software development from Jan 4th 2021-April 30th 2021.

During the internship he worked on **NLP Data analysis for low resource Indo-European languages** using Python programming language.

During the internship we found **Mr. Suraj Kumar** to be sincere, hard-working and a quick learner. We wish him all the best in his future endeavors.

Sincerely,

Vijay Mishra

Vijay Mishra, Director
Yscholar Technology LLP
9061, SOBHA DREAM ACRES, PANATHUR ROAD, SOBHA DREAM BALAGERE VILLAGE BANGALORE Bangalore KA 560087
Tel: (+91) 998-674-6745

Date Issued: May 1st, 2021

Tel. +91 998-674-6745
yscholar11@gmail.com

भारतीय सूचना प्रौद्योगिकी संस्थान भागलपुर

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY BHAGALPUR

An Institute of National Importance Under Act of Parliament

CERTIFICATE

This is to certify that the project entitled "***Natural Language Processing (NLP) Data Analysis for low resource Indo-European languages using Python***" is carried out by

Mr. Abhishek Kumar (170101003)

Mr. Himanshu Ranjan (170101017)

Mr. Suraj Kumar (170101052)

, B. Tech. students of IIIT Bhagalpur. This project has been submitted in partial fulfilment for the award of "***Bachelor of Technology"*** degree in ***Computer Science and Engineering*** at ***Indian Institute of Information Technology Bhagalpur***.

No part of this project has been submitted for the award of any previous degree to the best of my knowledge.

(Head)
Dr. Pradeep Kumar Biswal
(Assistant Professor, CSE, IIIT Bhagalpur)

Acknowledgement

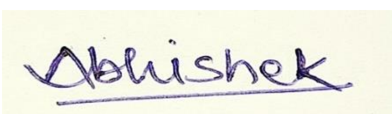
It is with great pleasure that we express our cordial thanks and indebtedness to our admirable Guide, *Mr. Ajay Kumar Mishra*, Mentor, Yscholar Technology LLP. His vast knowledge, expert supervision, and enthusiasm continuously challenged and motivated us to achieve our goal. We will be eternally grateful to him for allowing us the opportunity to work on this project.

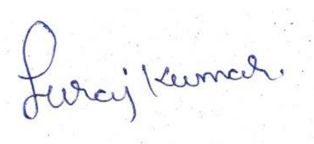
We express our sincere gratitude to *Dr. Pradeep Kumar Biswal*, Assistant Professor and Head of Department, Computer Science and Engineering and *Dr. Rupam Bhattacharyya*, Faculty Advisor, Computer Science and Engineering, for their valuable help and suggestions and for providing us all relevant facilities that helped us to complete this work in time.

During the course of this Internship report preparation, we have received a lot of support, encouragement, advice, and assistance from many people and to this end, we are deeply grateful to them all.

We have great pleasure in expressing our sincere gratitude and thanks to the *Prof. Arvind Choubey*, *Director*, Indian Institute of Information technology Bhagalpur, and all the faculty members of the Department of Computer Science and Engineering, IIIT Bhagalpur for the constant encouragement for innovation and hard work.

Finally, the present work certainly would not have been possible without the help of our friends, and also the blessings of our parents.





Mr. Abhishek Kumar
(170101003)

Mr. Himanshu Ranjan
(170101017)

Mr. Suraj Kumar
(170101052)

June 2021

About Yscholar Technology LLP

Directors:

Vijay Mishra and Priyata Pandey

About:

Yscholar Technology LLP is a very early-stage company working on building next generation EdTech products.

To give you a brief summary of what we do. We are specialists in text analytics and NLP. We are developing products to improve knowledge representation and knowledge acquisition. Our input is primarily texts from books corpora, articles and other open source as well as proprietary text corpus and we do text modelling and are building products which leverage them. We also focus on India's specific needs and culture and applying our techniques to understanding India's historical literature and texts.

Table of Contents

1 Introduction

Yscholar Technology LLP is a very early-stage company working on building next generation EdTech products.

To give you a brief summary of what we do. We are specialists in text analytics and NLP. We are developing products to improve knowledge representation and knowledge acquisition. Our input is primarily texts from books corpora, articles and other open source as well as proprietary text corpus and we do text modelling and are building products which leverage them. We also focus on India's specific needs and culture and applying our techniques to understanding India's historical literature and texts.

We got an internship opportunity Yscholar Technology LLP as a Back End Software Developer intern. The position was fully remote and began on Jan. 4th, 2021. We successfully completed our 16-week internship session during the academic session 2020-2021 B. Tech 8th semester. We were given a task "**Natural Language Processing (NLP) Data Analysis for low resource Indo-European languages using Python**" which we completed successfully.

Industry Type: EdTech

Founded: 2020

HeadQuarters: Banglore, Karnataka

2 Project Plan

The project has been done in four phases, namely:

- Survey Existing Data
- Collect available data
- Pre-process the data
- Analyze the data

The time spent on these phases is 10%, 10%, 30% and 30%. The rest 20% was spent on incremental development of this project and testing.

The project followed the Incremental model.

**Incremental Model is a process of software development where requirements are broken down into multiple standalone modules of a software development cycle.

3 Requirement Analysis

3.1 Software Configuration:

- This software package is developed using
    1. PyCharm
    2. Python
    3. Git
- Operating System:
    1. Windows 7 | Windows 8 | Windows 10

3.2 Hardware Configuration:

- Processor: Core i5, 2.4GHz
- Hard Disk: 150 GB
- RAM: 2GB
- Resolution: 1280 X 700

4 Installation Guide and Tutorials

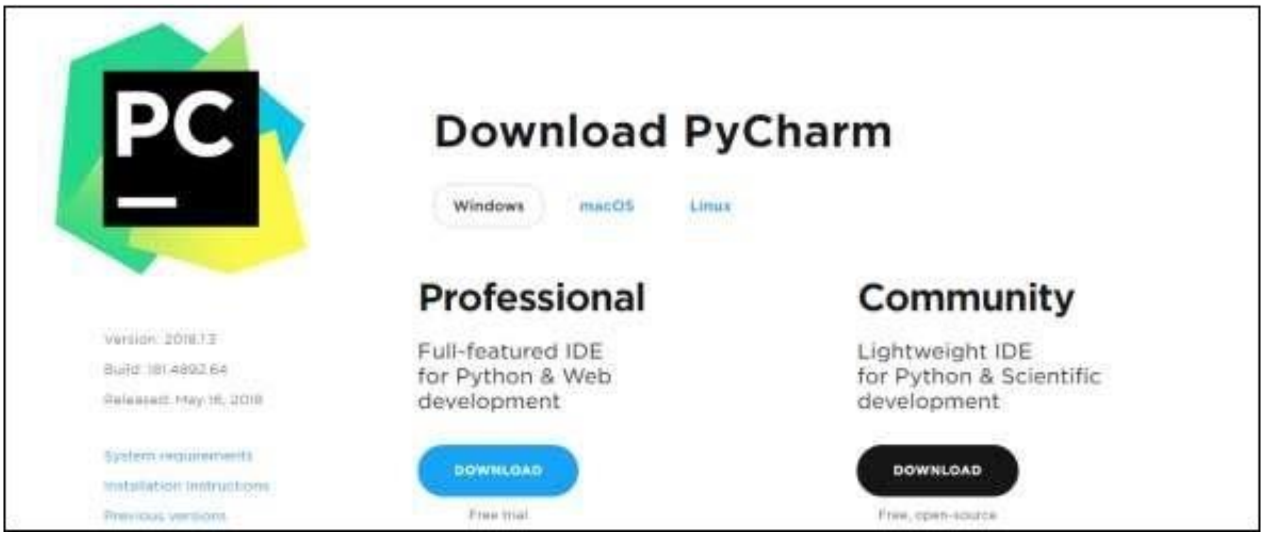
4.1 Installation Guide for PyCharm.

Steps Involved

We will have to follow the steps given below to install PyCharm on your system. These steps show the installation procedure starting from downloading the PyCharm package from its official website to creating a new project.

4.1.1 Step 1

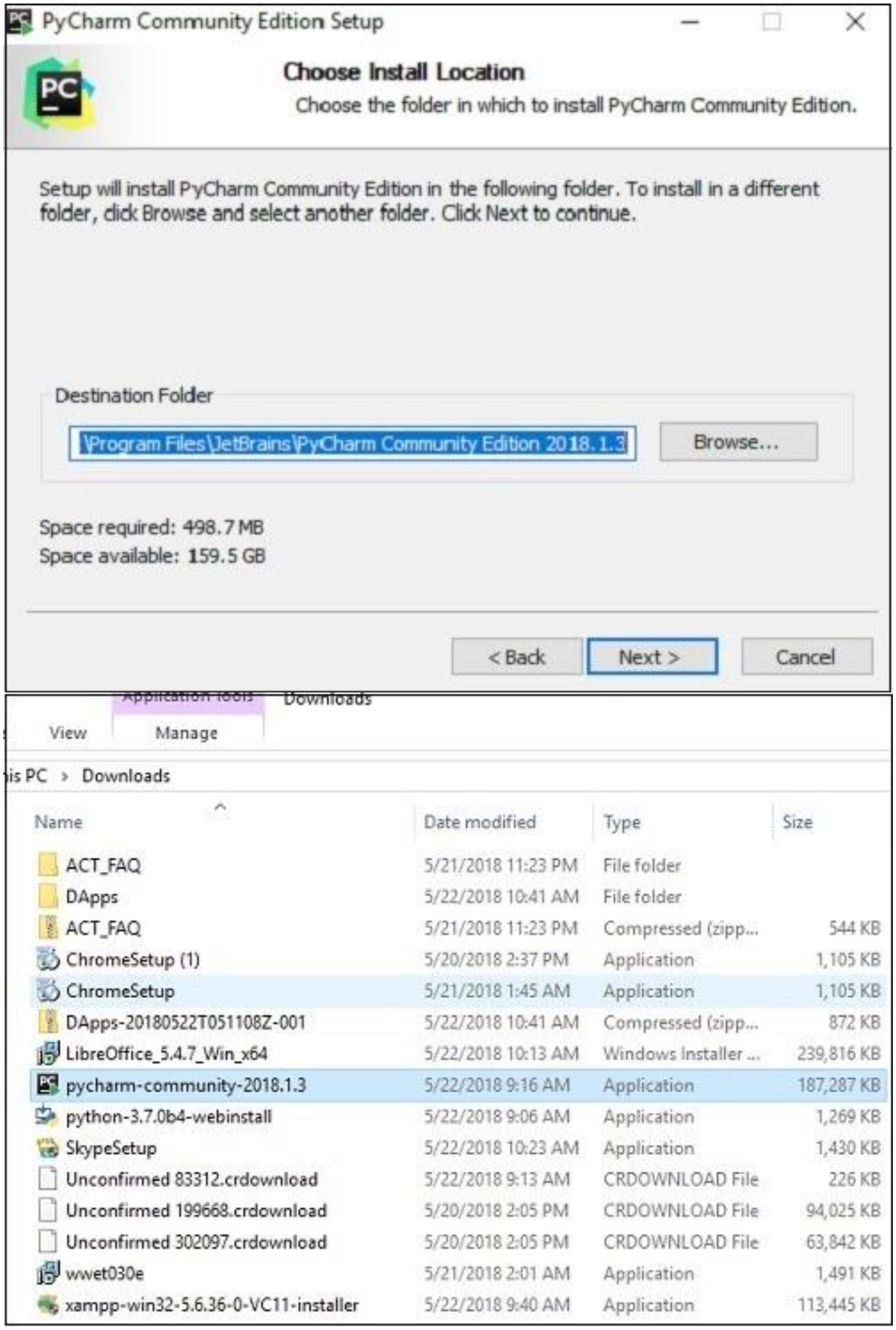
Download the required package or executable from the official website of PyCharm https://www.jetbrains.com/pycharm/download/#section=windowsHere you will observe two versions of package for Windows as shown in the screenshot given below −



Note that the professional package involves all the advanced features and comes with free trial for few days and the user has to buy a licensed key for activation beyond the trial period. Community package is for free and can be downloaded and installed as and when required. It includes all the basic features needed for installation. Note that we will continue with community package throughout this tutorial.

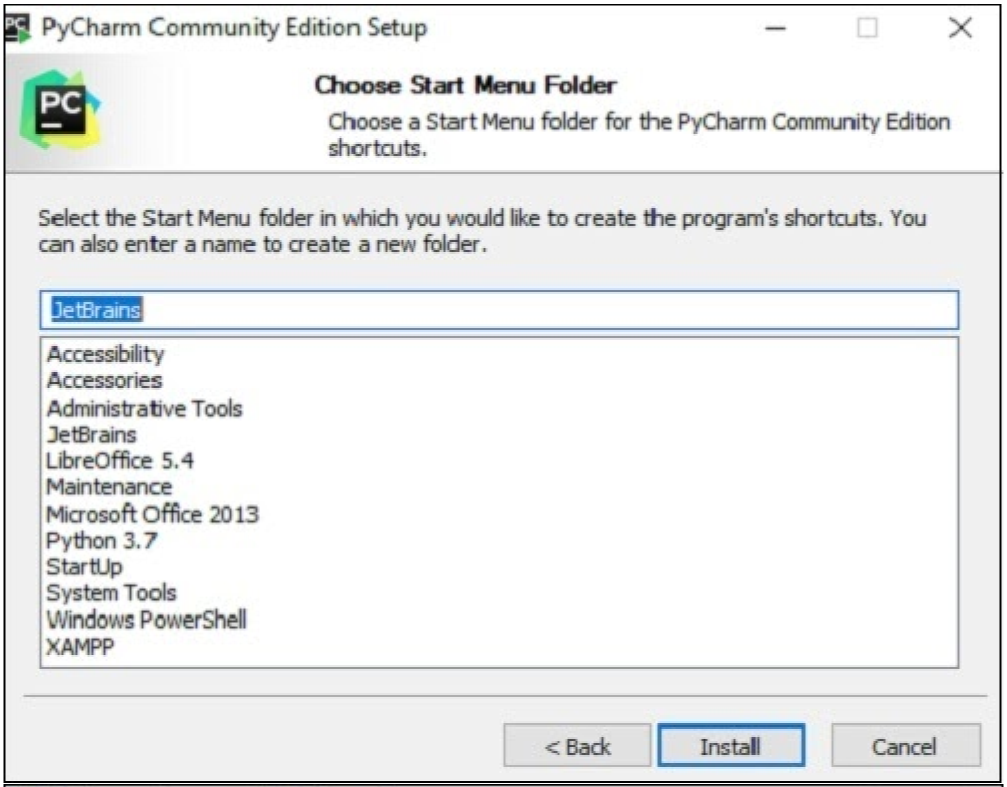
4.1.2 Step 2

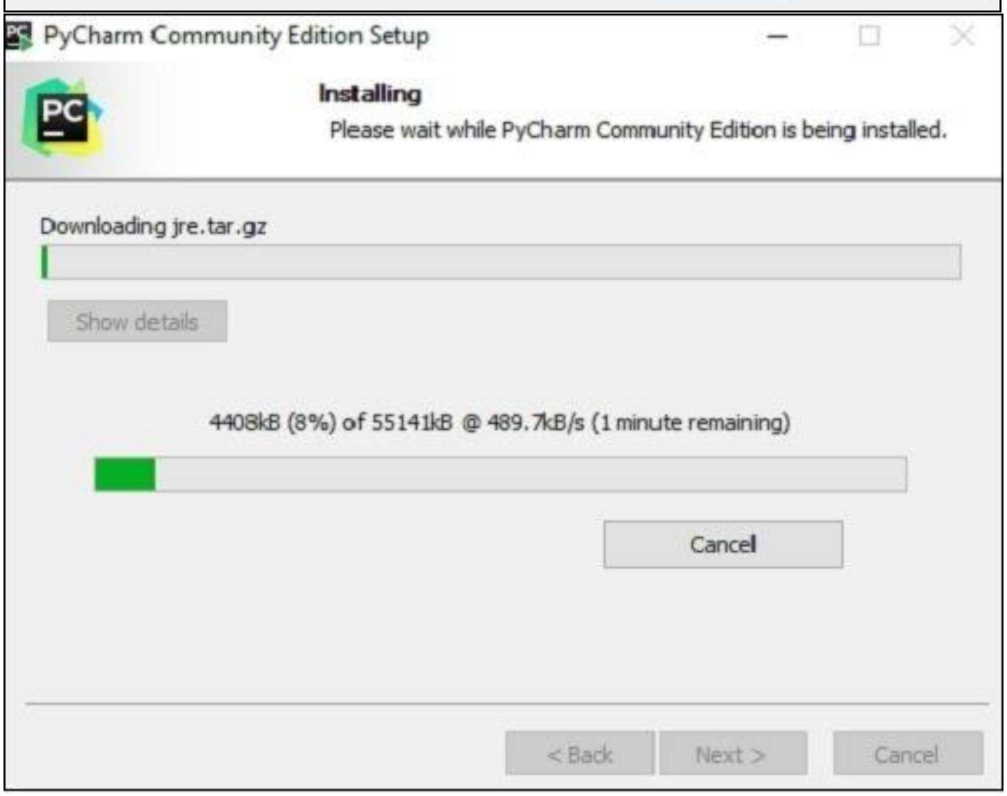
Download the community package (executable file) onto our system and mention a destination folder as shown below −



4.1.3 Step 3

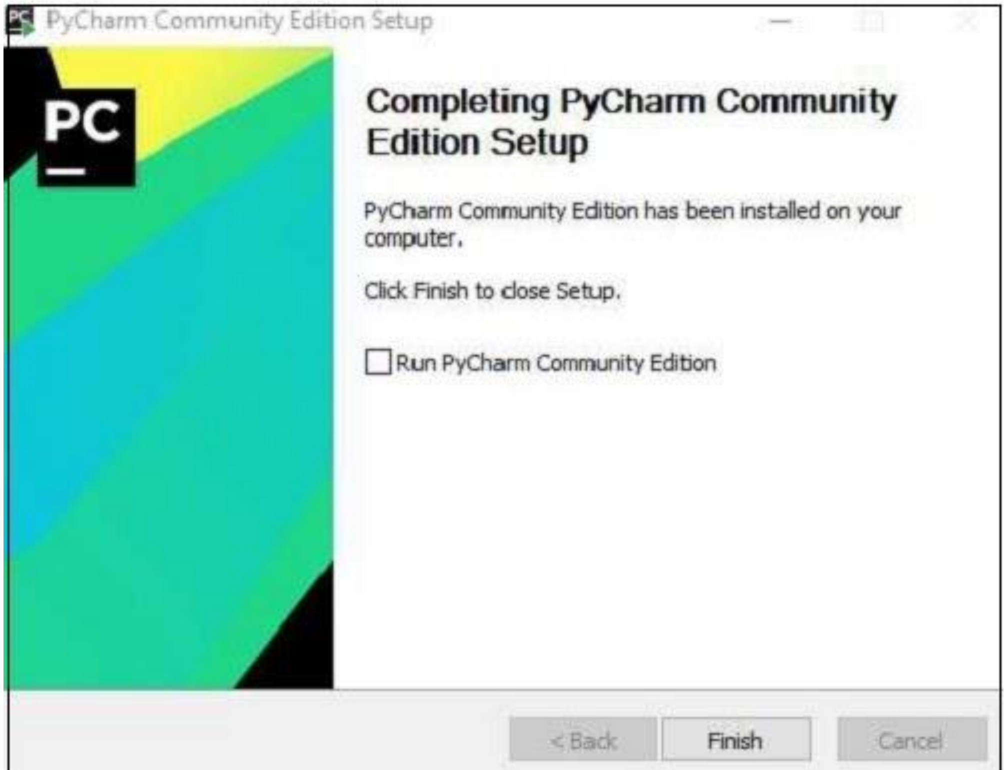
Now, begin the installation procedure similar to any other software package.

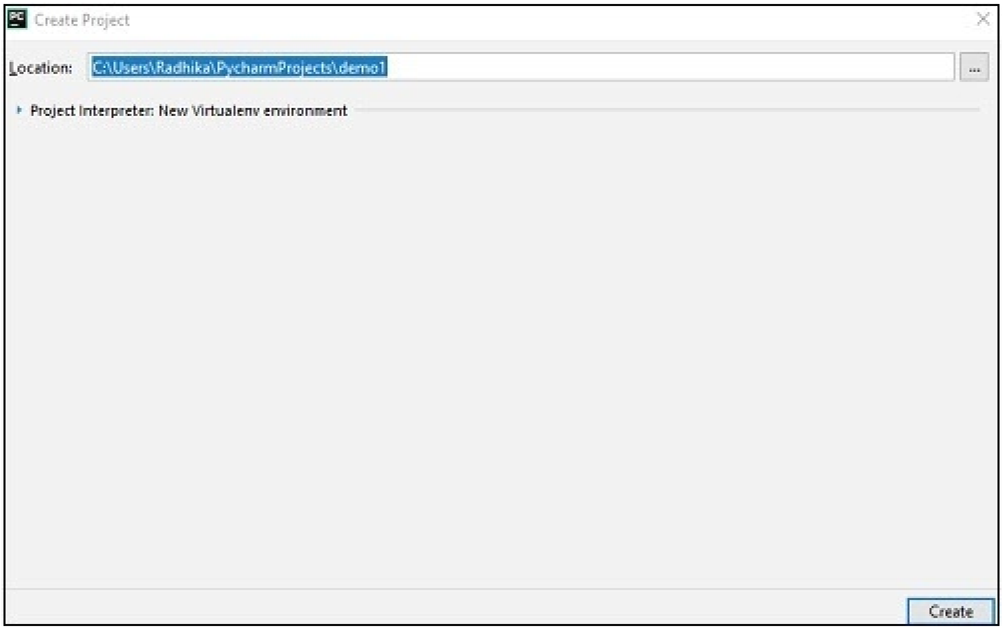




4.1.4 Step 4

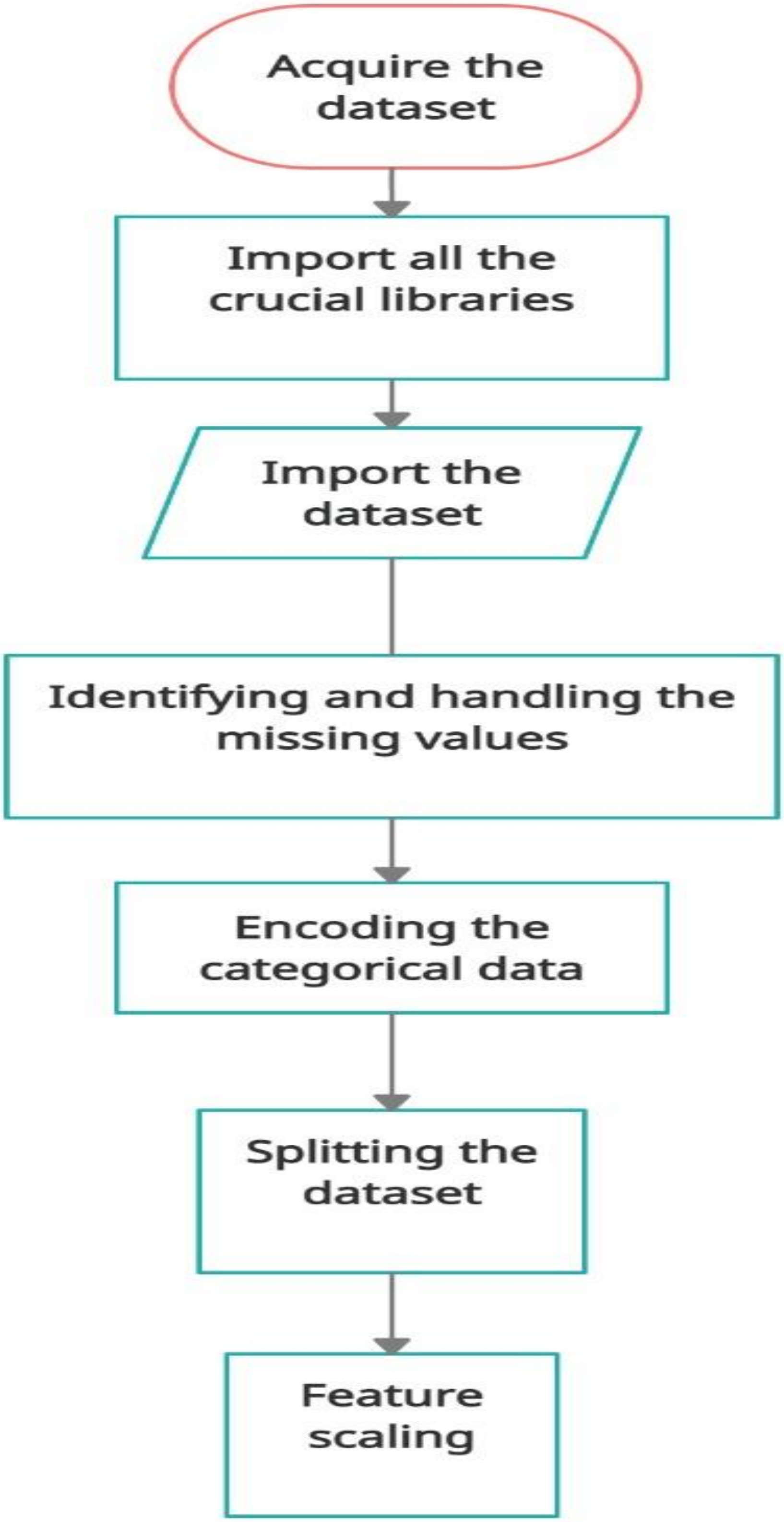
Once the installation is successful, PyCharm asks us to import settings of the existing package if any.





This helps in creating a new project of Python where we can work from the scratch. Note that unlike other IDEs, PyCharm only focusses on working with projects of Python scripting language.

5 Flow Chart


Acquire the dataset
Import all the crucial libraries
Import the dataset
Identifying and handling the missing values
Encoding the categorical data
Splitting the dataset
Feature scaling

6 Code Structure

6.1 Important Steps

There are 4 main important steps for the pre-processing of data.

- Splitting of the data set in Training and Validation sets
- Taking care of Missing values
- Taking care of Categorical Features
- Normalization of data set

Let’s have a look at all of these points.

6.2 Train Test Split:

Train Test Split is one of the important steps in Machine Learning. It is very important because your model needs to be evaluated before it has been deployed. And that evaluation needs to be done on unseen data because when it is deployed, all incoming data is unseen.

The main idea behind the train test split is to convert original data set into 2 parts

- Train
- Test

where train consists of training data and training labels and test consists of testing data and testing labels.

```
# Data Preprocessing
# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

```python
# Importing the dataset
dataset = pd.read_csv('../input/Data.csv')

# Creating Matrix of the features(independent Variables)
X = dataset.iloc[:, :-1].values

# Creating The dependent Variable Vector
y = dataset.iloc[:, 3].values

# Taking care of missing data (replacing with the mean)
from sklearn.preprocessing import Imputer
imputer = Imputer(missing_values ='NaN', strategy ="mean", axis = 0)

# Fitting the imputer object to the matrix of features X
imputer = imputer.fit(X[:, 1:3])

# Replacing the missing data by the mean of the column
X[:, 1:3] = imputer.transform(X[:, 1:3])
```

```python
# Encoding Categorical Data
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
labelencoder_X = LabelEncoder()
X[:, 0] = labelencoder_X.fit_transform(X[:, 0])

#Dummy Encoding
onehotencoder = OneHotEncoder(categorical_features = [0])
X = onehotencoder.fit_transform(X).toarray()

# Encoding Categorical data
labelencoder_y = LabelEncoder()
y = labelencoder_y.fit_transform(y)
```

```python
# Splitting the Dataset into the training Set and Test set
from sklearn.cross_validation import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

6.3 Taking Care of Missing Values:

There is a famous Machine Learning phrase which is Garbage in Garbage out.

If our data set is full of NaNs and garbage values, then surely our model will perform garbage too. So, taking care of such missing values is important.

3. Taking care of Categorical Features:

We can take care of categorical features by converting them to integers. There are 2 common ways to do so.

1. Label Encoding
2. One Hot Encoding

6.4 Normalizing the Dataset:

This brings us to the last part of data pre-processing, which is the normalization of the dataset. It is proven from certain experimentation that Machine Learning and Deep Learning Models perform way better on a normalized data set as compared to a data set that is not normalized.

The goal of normalization is to change values to a common scale without distorting the difference between the range of values.

```
# Feature Scaling(Standardisation and Normalisation)
from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)
```

6.5 Learning Outcomes

- Splitting the Dataset
- Filling in Missing values
- Dealing with Categorical Data
- Normalization of Dataset for improved results

7 Dataset Preparation

7.1 Word Similarity

Brief overview of Datasets -

This dataset contains similar words and their similarity count which lies between 0-10.

With the advent of word representations, word similarity tasks are becoming increasing popular as an evaluation metric for the quality of the representations. In this task, we present manually annotated monolingual word similarity datasets of six Indian languages - Urdu, Telugu, Marathi, Punjabi, Tamil and Gujarati. These languages are most spoken Indian languages worldwide after Hindi and Bengali. For the construction of these datasets, our approach relies on translation and re-annotation of word similarity datasets of English. We also present baseline scores for word representation models using state-of-the-art techniques for Urdu, Telugu and Marathi by evaluating them on newly created word similarity datasets.

Word representations are being increasingly popular in various areas of natural language processing like dependency parsing (Bansal et al., 2014), named entity recognition (Miller et al., 2004) and parsing (Socher et al., 2013). Word similarity task is one of the most popular benchmarks for the evaluation of word representations. Applications of word similarity range from Word Sense Dis-ambiguation (Patwardhan et al., 2005), Machine Translation Evaluation (Lavie and Denkowski,2009), Question Answering (Mohler et al., 2011), and Lexical Substitution (Diana and Navigli,2009). Word Similarity task is a computationally efficient method to evaluate the quality of word vectors. It relies on finding correlation between hu-man assigned semantic similarity (between words) and corresponding word vectors.

Datasets Size - Size of training, testing and dev sets

Size of training set-160

Size of testing set-52

Size of dev set-52

Different classes of labels and their counts

Word1-236

Word2-236

Similarity Count-236

SAMPLE DATA:

	A	B	C
1	Word1	Word2	Similarity
2	मोहब्बत	सेक्स	6.8
3	बाघ	बिल्ली	7.0
4	किताब	कागज़	7.6
5	कंप्यूटर	कीबोर्ड	7.6
6	कंप्यूटर	इंटरनेट	8.0
7	विमान	कार	6.0
8	रेलगाड़ी	कार	6.2
9	टेलीफोन	संचार	7.6
10	टेलीविजन	रेडियो	6.4
11	मीडिया	रेडियो	5.8
12	ब्रेड	मक्खन	6.6
13	खीरा	आलू	5.8
14	चिकित्सक	नर्स	7.6
15	प्रोफ़ेसर	चिकित्सक	4.6
16	छात्र	प्रोफ़ेसर	7.0
17	होशियार	छात्र	5.2
18	होशियार	बेवकूफ	6.2
19	किताब	पुस्तकालय	7.6
20	बैंक	पैसे	8.0
21	लकड़ी	जंगल	7.0
22	प्रोफ़ेसर	खीरा	0.0
23	राजा	रानी	8.2
24	बिशप	रबी	7.0
25	यरूशलेम	इजराइल	7.8
26	पवित्र	सेक्स	1.0
27	माराडोना	फ़ुटबॉल	7.8
28	फ़ुटबॉल	सॉकर	9.0
29	फ़ुटबॉल	बास्केटबाल	6.4
30	फ़ुटबॉल	टेनिस	5.8

HINDI

	A	B	C
1	Word1	Word2	Similarity
2	పులి	పిల్లి	6
3	టైగర్	టైగర్	10
4	పుస్తక	కాగితం	8
5	కంప్యూటర్	కీబోర్డు	5
6	కంప్యూటర్	ఇంటర్నెట్	6
7	విమానం	కారు	5
8	రైలు	కారు	5
9	టెలిఫోన్	కమ్యూనికేషన్స్	6
10	టెలివిజన్	రేడియో	6.125
11	మీడియా,	రేడియో	6.5
12	బ్రెడ్	బట్టర్	6
13	దోసకాయ	బంగాళాదుంప	5
14	డాక్టర్	నర్సు	7
15	ప్రొఫెసర్	డాక్టర్	4.375
16	విద్యార్థి	ప్రొఫెసర్	7
17	స్మార్ట్	విద్యార్థి	6
18	స్మార్ట్	స్టుపిడ్	3
19	సంతానోత్పత్తి	గుడ్లు	6.5
20	బుక్	లైబ్రరీ	8
21	బ్యాంకు	డబ్బు	9
22	డబ్బు	నగదు	10
23	రాజు	రాణి	7.375
24	బిషప్	రబ్బి	8
25	జెరూసలేం	ఇజ్రాయెల్	9
26	జెరూసలేం	పాలస్తీనా	9
27	పవిత్ర	సెక్స్	2
28	మారడోనా	ఫుట్బాల్	7
29	ఫుట్బాల్	సాకర్	9.5
30	ఫుట్బాల్	టెన్నిస్	3

TELUGU

	Word1	Word2	Similarity
2	புலி	பூனை	7
3	புலி	புலி	10
4	புத்தகம்	காகித	8
5	கணினி	விசைப்பலகை	8
6	கணினி	இணைய	7
7	தொலைபேசி	தொடர்பாடல்	8
8	தொலைக்காட்சி	வானொலி	7
9	ஊடக	வானொலி	8.25
10	ரொட்டி	வெண்ணெய்	7.5
11	வெள்ளரி	உருளைக்கிழங்கு	6.25
12	மருத்துவர்	செவிலியர்	7.5
13	மாணவர்	பேராசிரியர்	8.75
14	கருவுறுதல்	முட்டை	9
15	புத்தகம்	நூலகம்	9
16	வங்கி	பணம்	8.75
17	மரம்	காடு	9
18	பணம்	பணம்	10
19	பேராசிரியர்	வெள்ளரி	0
20	ராஜா	முட்டைக்கோஸ்	0
21	ராஜா	ராணி	8.625
22	ஜெருசலேம்	இஸ்ரேலின்	7.5
23	ஜெருசலேம்	பாலஸ்தீன	7
24	புனித	செக்ஸ்	0
25	மரடோனா	கால்பந்து	7
26	கால்பந்து	கால்பந்து	10
27	கால்பந்து	டென்னிஸ்	6.125
28	டென்னிஸ்	மோசடி	0
29	அரபாத்	பயங்கரவாத	7.375
30	அரபாத்	ஜாக்சன்	0

TAMIL

	Word1	Word2	Similarity
2	ਪਿਆਰ	ਸੈਕਸ	1
3	ਟਾਈਗਰ	ਟਾਈਗਰ	10
4	ਕਿਤਾਬ	ਪੇਪਰ	8
5	ਕੰਪਿਊਟਰ	ਕੀ-ਬੋਰਡ	7.375
6	ਕੰਪਿਊਟਰ	ਇੰਟਰਨੈੱਟ	6.75
7	ਜਹਾਜ਼	ਕਾਰ	6
8	ਰੇਲ ਗੱਡੀ	ਕਾਰ	4
9	ਟੈਲੀਫੋਨ	ਸੰਚਾਰ	8
10	ਮੀਡੀਆ	ਰੇਡੀਓ	7
11	ਰੋਟੀ	ਮੱਖਣ	3
12	ਖੀਰੇ	ਆਲੂ	6
13	ਡਾਕਟਰ	ਨਰਸ	8
14	ਪ੍ਰੋਫੈਸਰ	ਡਾਕਟਰ	5.375
15	ਵਿਦਿਆਰਥੀ	ਪ੍ਰੋਫੈਸਰ	7.375
16	ਕਿਤਾਬ	ਲਾਇਬਰੇਰੀ	6
17	ਲੱਕੜ	ਜੰਗਲ	7.375
18	ਪੈਸਾ	ਨਕਦ	8
19	ਰਾਜਾ	ਰਾਣੀ	9
20	ਯਰੂਸ਼ਲਮ	ਇਸਰਾਏਲ	7
21	ਮਾਰਾਡੋਨਾ	ਫੁੱਟਬਾਲ	8
22	ਫੁੱਟਬਾਲ	ਫੁੱਟਬਾਲ	10
23	ਬਾਸਕਟਬਾਲ	ਫੁੱਟਬਾਲ	6.375
24	ਫੁੱਟਬਾਲ	ਟੈਨਿਸ	6.125
25	ਟੈਨਿਸ	ਰੈਕੇਟ	7
26	ਅਰਾਫਾਤ	ਅਮਨ	5
27	ਕਾਨੂੰਨ	ਵਕੀਲ	8
28	ਫਿਲਮ	ਆਲੋਚਕ	1.125
29	ਫਿਲਮ	ਥੀਏਟਰ	8.5
30	ਫਿਜ਼ਿਕਸ	ਪ੍ਰੋਟੋਨ	4.5

PUNJABI

	A	B	C
1	Word1	Word2	Similarity
2	वाघ	मांजर	7
3	वाघ	वाघ	10
4	पुस्तक	कागद	7.375
5	संगणक	कळफलक	8
6	संगणक	इंटरनेट	8.5
7	विमान	कार	6.75
8	रेल्वे	कार	6.625
9	टेलिफोन	संवाद	9
10	दूरदर्शन	रेडिओ	7
11	मीडिया	रेडिओ	8
12	भाकरी	लोणी	6.375
13	काकडी	बटाटा	7
14	डॉक्टर	परिचारिका	6.375
15	प्राध्यापक	डॉक्टर	4
16	विद्यार्थी	प्राध्यापक	6
17	स्मार्ट	विद्यार्थी	4.375
18	स्मार्ट	मूर्ख	3
19	कस	अंडी	7
20	पुस्तक	ग्रंथालय	6.375
21	बँक	पैसा	8.5
22	लाकूड	वन	6
23	पैसा	रोख	9.375
24	प्राध्यापक	काकडी	0
25	राजा	कोबी	0
26	राजा	राणी	7.375
27	पवित्र	सेक्स	2.375
28	टेनिस	रॅकेट	6.375
29	यासिर	शांतता	6
30	अराफात	दहशतवादी	8.375

MARATHI

	A	B	C
1	Word1	Word2	Similarity
2	محبت	جنس	5
3	کتاب	کاغذ	9
4	کمپیوٹر	کی بورڈ	9
5	کمپیوٹر	انٹرنیٹ	8
6	ہوائی جہاز	گاڑی	4
7	ٹیلی ویژن	ریڈیو	8
8	روٹی	مکھن	4
9	ککڑی	آلو	3
10	ہوشیار	طالب علم	7
11	بینک	پیسہ	9
12	لکڑی	جنگل	8.5
13	پیسے	نقد	9.375
14	بادشاہ	گوبھی	0
15	بادشاہ	ملکہ	9
16	یروشلم	اسرائیل	9
17	یروشلم	فلسطینی	9
18	مقدس	جنس	0.375
19	فٹ بال	ساکر	9
20	فٹ بال	ٹینس	5
21	ٹینس	ریکیٹ	8
22	عرفات	امن	6
23	عرفات	دہشت گردی	5
24	قانون	وکیل	9
25	فلم	پاپکارن	7
26	فلم	تھیٹر	9
27	طبیعیات	پروٹون	9
28	فزکس	کیمسٹری	8
29	شراب نوشی	کیمسٹری	4
30	پیئے	گاڑی	0

URDU

7.2 Paraphrase detection

Brief overview of Datasets -

This dataset contains Paraphrase with active and passive form.

Sub Task 1: Given a pair of sentences from newspaper domain, the task is to classify them as Paraphrases (P) or Not Paraphrases (NP).

Sub Task 2: Given two sentences from newspaper domain, the task is to identify whether they are completely equivalent (E) or roughly equivalent (RE) or not equivalent (NE). This task is similar to the subtask 1, but the main difference is 3-point scale tag in paraphrases.

It contains datasets of different languages i.e., Tamil, Malayalam, Hindi and Punjabi.

Datasets Size - Size of training, testing and dev sets:

For Sub Task 1:

Size of training set-2500

Size of testing set-900

Size of dev set-900

For Sub Task 2:

Size of training set-3500

Size of testing set-1400

Size of dev set-1400

Different classes of labels and their counts:

For Sub Task 1:

Sentence1, Sentence2

Paraphrases(P) or Not Paraphrases (NP)

For Sub Task 2:

Sentence1, Sentence2

Equivalent (E) or Roughly Equivalent (RE) or Not Equivalent (NE)

Sample Data:

	A	B	C
1	29 साल के जीतू राय मैन्स 10 मीटर एयर पिस्टल टूर्नामेंट की वर्ल्ड रैंकिंग में 3rd पोजिशन पर हैं।	एयर पिस्टल टूर्नामेंट के मैन्स 10 मीटर में 29 साल के जीतू राय वर्ल्ड रैंकिंग में 3rd पोजिशन पर हैं।	P
2	सोमवार को सुबह से ऑटो चालकों की हड़ताल के चलते लोग खासे परेशान हुए, वहीं दोपहर बाद करीब सवा 3 बजे भारी बारिश भी परेशानी की सबब बन गई।	ऑटो चालकों की हड़ताल से सोमवार सुबह लोग खासे परेशान हुए और दोपहर बाद करीब 3:15 बजे भारी बारिश भी परेशानी का सबब बन गई।	P
3	ओसामा पाकिस्तानी मिलिट्री एकेडमी के पास बने घर में रह रहा था	ओसामा का ठिकाना कम्युनिटी मिलिटरी अकेडमी के निकट था।'	P
4	यह कैम्पेन देश के उत्तरी हिस्से में फैल गया और अब सोशल मीडिया पर बहस शुरु हो गई है।	कई महिलाओं ने इस पर सकारात्मक प्रतिक्रिया दी हैं।	NP
5	राहुल ने कहा टारगेट बनने से खुश हूं।	कांग्रेस वाइस प्रेसिडेंट ने मंगलवार को कहा, " टारगेट बनने से खुश हूं।"	P
6	परंपरागत रास्ते को 16 किमी तक शेड से ढंका जा चुका है। किनारे पर फेंसिंग भी की गई है, ताकि कोई गिरे नहीं और कचरा नहीं फैलाए।	किनारे पर फेंसिंग भी की गई है, ताकि कोई गिरे नहीं और कचरा नहीं फैलाए।	SP
7	स्कूल प्रिंसीपल किरणबाला नागर की यह करतूत कैमरे में कैद हो गई।	किरणबाला नागर जो की एक स्कूल प्रिंसीपल की करतूत कैमरे में कैद हो गई।	P
8	बता दें कि प्रेसिडेंट अबेदरब्बो मंसूर हादी ने अदन को यमन की अस्थाई राजधानी घोषित कर रखा है।	किसी भी आतंकी ग्रुप ने इसकी जिम्मेदारी नहीं ली है, बता दें कि प्रेसिडेंट अबेदरब्बो मंसूर हादी ने अदन को यमन की अस्थाई राजधानी घोषित कर रखा है।	SP
9	एग्रीकल्चर साइंस सेंटर के प्रोग्राम कोऑर्डिनेटर केआर साहू ने बताया कि कुछ साल पहले होमप्रकाश का नाम दिल्ली भेजकर खेती के मामले में खास काम करने 25 लाख का इनाम दिलवाने कोशिश की थी।	कुछ साल पहले होमप्रकाश का नाम दिल्ली भेजकर खेती के मामले में खास काम करने 25 लाख का इनाम दिलवाने कोशिश की थी।	SP
10	विराट कोहली पर लग सकता है एक मैच का बैन	कोहली पर लग सकता है एक मैच का बैन	P
11	ऐसा पंपिंग सेल्स की कार्यप्रणाली में गड़बड़ी आने की वजह से होता है। खराब सेल्स को निकालने पर स्वस्थ सेल उनका स्थान ले लेते हैं।	खराब सेल्स को निकालने पर स्वस्थ सेल उनका स्थान ले लेते हैं।	SP
12	प्रोजेक्ट के लिए करीब एक हजार महिलाओं को ट्रेनिंग दी जाएगी। गली में रहने वाली महिलाओं व लड़कियों से यौन उत्पीड़न जैसे मामलों पर बात करेंगी।	गली में रहने वाली महिलाओं व लड़कियों से यौन उत्पीड़न जैसे मामलों पर बात करेंगी।	SP
13	मुकेश की मौत से गांव में शोक छा गया।	गांव में मुकेश की मौत से शोक छा गया।	P
14	मानवेंद्र गाजे-बाजे के साथ लड़की दरवाजे पर पहुंच गया और पिता के पास शादी के लिए सूचना भिजवाई।	गाजे-बाजे के साथ मानवेंद्र लड़की के दरवाजे पर पहुँच और लड़की के पिता के पास शादी के लिए सूचना भिजवाई।	P
15	गिरोह ने कुछ अकाउंट नंबर भी रखे थे, जिसमें पैसे ट्रांसफर लिए जाते। गिरोह को उम्मीद थी कि वे डेढ़ हजार को परीक्षा में बुला रहे हैं और आधे लोगों से भी पैसे ले लिए तो करोड़ों का खेल हो जाएगा।	गिरोह को उम्मीद थी कि वे डेढ़ हजार को परीक्षा में बुला रहे हैं और आधे लोगों से भी पैसे ले लिए तो करोड़ों का खेल हो जाएगा।	SP
16	पूछताछ में खुलासा हुआ है कि गिरोहबाजों ने फर्जी भर्ती परीक्षा आयोजित करने के एक हफ्ते बाद हर युवक से दो दो लाख रुपए वसूल करने का प्लान बनाया था।	गिरोहबाजों ने हर युवक से दो दो लाख रुपए वसूल करने का प्लान बनाया था।	SP

HINDI

PUNJABI

	A	B	C
1	'ਆਪ' ਦੇ ਲੋਕ ਸਭਾ ਹਲਕਾ ਦੇ ਨਿਗਰਾਨ ਅਕੁੰਸ਼ ਨਾਰੰਗ ਆਗਾਮੀ ਵਿਧਾਨ ਸਭਾ ਚੋਣਾਂ ਦੌਰਾਨ ਪਾਰਟੀ ਦੀ ਟਿਕਟ 'ਤੇ ਚੋਣ ਲੜਨ ਦੇ ਚਾਹਵਾਨ ਉਮੀਦਵਾਰਾਂ ਨਾਲ ਮੀਟਿੰਗ ਕਰਨ ਲਈ ਪੁੱਜੇ ਸਨ।	ਰੋਹ ਵਿਚ ਆਏ ਹਮਾਇਤੀਆਂ ਵੱਲੋਂ ਸੁੱਚਾ ਸਿੰਘ ਛੋਟੇਪੁਰ ਨੂੰ ਅਹੁਦੇ ਤੋਂ ਹਟਾਉਣ ਕਰਕੇ ਅਰਵਿੰਦ ਕੇਜਰੀਵਾਲ ਅਤੇ ਦੁਰਗੇਸ਼ ਪਾਠਕ ਖਿਲਾਫ਼ ਨਾਅਰੇਬਾਜ਼ੀ ਕੀਤੀ ਗਈ।	NP
2	'ਆਪ' ਦੇ ਲੋਕ ਸਭਾ ਹਲਕਾ ਦੇ ਨਿਗਰਾਨ ਅਕੁੰਸ਼ ਨਾਰੰਗ ਆਗਾਮੀ ਵਿਧਾਨ ਸਭਾ ਚੋਣਾਂ ਦੌਰਾਨ ਪਾਰਟੀ ਦੀ ਟਿਕਟ 'ਤੇ ਚੋਣ ਲੜਨ ਦੇ ਚਾਹਵਾਨ ਉਮੀਦਵਾਰਾਂ ਨਾਲ ਮੀਟਿੰਗ ਕਰਨ ਲਈ ਪੁੱਜੇ ਸਨ।	ਅਕੁੰਸ਼ ਨਾਰੰਗ ਜੋ ਕਿ 'ਆਪ' ਦੇ ਲੋਕ ਸਭਾ ਹਲਕਾ ਦੇ ਨਿਗਰਾਨ ਹਨ ਨੇ ਆਗਾਮੀ ਵਿਧਾਨ ਸਭਾ ਚੋਣਾਂ ਦੌਰਾਨ ਪਾਰਟੀ ਦੀ ਟਿਕਟ 'ਤੇ ਚੋਣ ਲੜਨ ਦੇ ਚਾਹਵਾਨ ਉਮੀਦਵਾਰਾਂ ਨਾਲ ਮੀਟਿੰਗ ਕਰਨ ਲਈ ਪੁੱਜੇ ਸਨ।	P
3	ਸ੍ਰੀਮਤੀ ਗਿੱਲ ਨੇ ਆਖਿਆ ਕਿ ਪਾਰਟੀ ਜਿਹੜੀ ਜ਼ਿੰਮੇਵਾਰੀ ਸੌਂਪੇਗੀ ਉਸ ਨੂੰ ਨਿਭਾਇਆ ਜਾਵੇਗਾ|	ਪਾਲ ਸਿੰਘ ਦੇ ਵਾਰਸਾਂ ਨੇ ਉਸਦੀ ਮੌਤ ਤੋਂ ਬਾਦ ਪੰਜਾਬ ਦੇ ਮੁੱਖ ਮੰਤਰੀ ਤੋਂ ਆਰਥਿਕ ਮੱਦਦ ਦੀ ਮੰਗ ਕੀਤੀ ਸੀ।	NP
4	ਸ੍ਰੀਮਤੀ ਗਿੱਲ ਨੇ ਆਖਿਆ ਕਿ ਪਾਰਟੀ ਜਿਹੜੀ ਜ਼ਿੰਮੇਵਾਰੀ ਸੌਂਪੇਗੀ ਉਸ ਨੂੰ ਨਿਭਾਇਆ ਜਾਵੇਗਾ|	ਪਾਰਟੀ ਜਿਹੜੀ ਜ਼ਿੰਮੇਵਾਰੀ ਸੌਂਪੇਗੀ ਉਸ ਨੂੰ ਨਿਭਾਇਆ ਜਾਵੇਗਾ ਸ੍ਰੀਮਤੀ ਗਿੱਲ ਨੇ ਆਖਿਆ	P
5	ਪਾਲ ਸਿੰਘ ਵਾਸੀ ਦੀਪ ਸਿੰਘ ਵਾਲਾ 4 ਅਕਤੂਬਰ 2014 ਨੂੰ ਆੜ੍ਹਤੀਆਂ ਦੇ ਕਰਜ਼ੇ ਤੋਂ ਤੰਗ ਆ ਕੇ ਖੁਦਕੁਸ਼ੀ ਕਰ ਗਿਆ ਸੀ।	ਪਾਲ ਸਿੰਘ ਦੇ ਪਰਿਵਾਰ ਨੂੰ ਰਾਹਤ ਦੇਣ ਲਈ 27 ਮਈ ਤੱਕ ਸਿਵਲ ਸਰਜਨ ਫਰੀਦਕੋਟ ਨੇ ਡਿਪਟੀ ਕਮਿਸ਼ਨਰ ਦੇ ਦਫ਼ਤਰ ਫਾਇਲ ਭੇਜ ਦਿਤੀ ਸੀ	NP
6	ਪਾਲ ਸਿੰਘ ਵਾਸੀ ਦੀਪ ਸਿੰਘ ਵਾਲਾ 4 ਅਕਤੂਬਰ 2014 ਨੂੰ ਆੜ੍ਹਤੀਆਂ ਦੇ ਕਰਜ਼ੇ ਤੋਂ ਤੰਗ ਆ ਕੇ ਖੁਦਕੁਸ਼ੀ ਕਰ ਗਿਆ ਸੀ।	4 ਅਕਤੂਬਰ 2014 ਨੂੰ ਪਾਲ ਸਿੰਘ ਵਾਸੀ ਦੀਪ ਸਿੰਘ ਵਾਲਾ ਆੜ੍ਹਤੀਆਂ ਦੇ ਕਰਜ਼ੇ ਤੋਂ ਤੰਗ ਆ ਕੇ ਖੁਦਕੁਸ਼ੀ ਕਰ ਗਿਆ ਸੀ।	P
	ਫਤਹਿਗੜ੍ਹ ਸਾਹਿਬ ਤੋਂ ਪਾਰਟੀ ਹਾਈਕਮਾਂਡ ਸਟਾਰ ਪ੍ਰਚਾਰਕ ਗੁੱਲ ਪਨਾਗ ਨੂੰ ਉਮੀਦਵਾਰ ਵਜੋਂ ਉਤਾਰਨ ਦੀ	ਪਾਰਟੀ ਜਿਹੜੀ ਜ਼ਿੰਮੇਵਾਰੀ ਸੌਂਪੇਗੀ ਉਸ ਨੂੰ ਨਿਭਾਇਆ ਜਾਵੇਗਾ	NP

MALYALAM

	A	B	C
1	പാരീസിൽ വച്ച് നവംബർ മൂന്നിന് നടന്ന ആക്രമണത്തിൽ നൂറ്റിമുപ്പതിലേറെ പേരാണ് കൊല്ലപ്പെട്ടത്.	നവംബർ മൂന്നിന് നടന്ന ആക്രമണപരമ്പരയിൽ നൂറ്റിമുപ്പതിലേറെ പേരാണ് പാരീസിൽ കൊല്ലപ്പെട്ടത്.	P
2	തിരുവനന്തപുരം ആനയറ ഒരുവാതിൽകോട്ട സ്വദേശിനി വർക്കലയിൽ നഴ്സിംഗ് വിദ്യാർത്ഥിനിയാണ് .	വർക്കലയിൽ നഴ്സിംഗ് വിദ്യാർത്ഥിനിയാണ് തിരുവനന്തപുരം ആനയറ ഒരുവാതിൽകോട്ട സ്വദേശിനിയായ പത്തൊമ്പതുകാരി.	P
3	`തിരുവനന്തപുരം സർക്കാർ എയ്ഡഡ് മേഖലയിൽ നാല് പുതിയ കോളേജുകൾ തുടങ്ങാൻ അനുമതി നൽകി സർക്കാർ ഉത്തരവായി	നാല് പുതിയ കോളേജുകൾ തിരുവനന്തപുരം സർക്കാർ എയ്ഡഡ് മേഖലയിൽ തുടങ്ങാൻ അനുവാദം നൽകി കേരള സർക്കാർ	P
4	അംഗങ്ങളായ എക്സ്പെഡിഷൻ നാല്പത്തിയാർകമാൻഡർ സ്കോട്ട് കെല്ലി, ഫ്ലൈറ്റ് എൻജിനീയർ ടിം കോപ്ര, മറ്റൊരു ഫ്ലൈറ്റ് എൻജിനീയർ ടിം പീകെ എന്നിവരാണ് ബഹിരാകാശത്തു നിന്ന് ഭൂമിയിലേക്ക് പുതുവർഷ ആശംസ നൽകിയിരിക്കുന്നത്.	ഒരു വീഡിയൊയിലൂടെയാണ് ഇന്റർനാഷണൽ സ്പെയ്സ് സ്റ്റേഷൻ അംഗങ്ങൾ ആശംസ അറിയിച്ചിരിക്കുന്നത്.	NP
5	അംഗപരിമിതർക്ക് സഹായകമാകുന്ന കൃത്രിമക്കൈകൾ ധാരാളമുണ്ട്.	അതെല്ലാം കുറഞ്ഞ ചെലവിൽ ലഭ്യമാകുന്നില്ല എന്നതാണ് നമ്മളെയെല്ലാം അലട്ടുന്ന പ്രധാന പ്രശ്നം.	NP
6	അഖിലേഷ് യാദവ് സർക്കാർ ഏർപ്പെടുത്തിയ റാണി ലക്ഷ്മിഭായ് പുരസ്കാര ജേതാവുകൂടിയായ അപർണ ഇപ്പോൾ അലാസകയിലെ മൗണ്ട് മികിൻലെ കയറാനുള്ള	അലഹാബാദിലെ ജില്ലാ മജിസ്ട്രേറ്റായ ഭർത്താവ് സഞ്ജയ് കുമാർ അപർണയുടെ യാത്രകൾക്ക് പൂർണ പിന്തുണയുമായി കൂടെയുണ്ട്.	NP
7	അഗ്നിക്കിരയായ വസ്ത്ര വ്യാപാരശാല അനേകം നാശനഷ്ടങ്ങൾ ഉണ്ടാകി	വസ്ത്ര വ്യാപാരശാല്യ്ക്ക് തീപിടിച്ചു വൻ നാശനഷ്ടം	P
8	അങ്ങനെ വീട്ടിലിരുന്ന് വരച്ചുണ്ടാക്കിയ രണ്ടു മെഷീനുകളുടെ മാതൃകയുമായാണ് ഈ സഹോദരിമാർ	അന്നുമുതൽ പഠനത്തിനു ശേഷം കിട്ടുന്ന സമയം മുഴുവനും ഉപകരണങ്ങൾ രൂപകൽപ്പന ചെയ്യാനാണ് ഉപയോഗിച്ചിരുന്നത്.	NP

TAMIL

	A	B	C
1	அணையின் நீர்மட்டம், 136 அடிக்	*முல்லை பெரியாறு அணையின் ந	NP
2	மத்திய அமைச்சக அதிகாரிகள்	.ஆவணங்களை திருடி வெளிநாடு	SP
3	பன்றிக் காய்ச்சலை (எச்1என்1 வ	.இதுவரையில் பன்றிக் காய்ச்சலுச்	SP
4	ஐந்து மீனவர்களையும் பாதுகாக்	.தி.முக.,வில் தந்தைக்கும், தனயனு	NP
5	ஒடிசா மாநிலம் பூரி ஜெகந்நாதர்	100 மணல் சிற்ப தேர்கள் அமைத்த	SP
6	ரூபாய் நோட்டில், முன்னாள் ரிசர்	14.60 கோடி ரூபாய் நோட்டுகள், ரிச	NP
7	சர்வதேச பள்ளிகள் விளை யாட்(	148 இந்திய தடகள வீரர்கள் பத்திர	SP
8	ஹரியாணாவில் இடஒதுக்கீடு கே	15 நாட்கள் கெடு விதித்து இடஒதுக்	SP
9	16-வது மக்களவையில் 2	16-வது மக்களவையில் 2	P
10	உ.பி.,யின் மதுரா பகுதியில் கட்ட	17 ஆண்டுகளாக இழப்பீடு கிடைக்	SP
11	1857ல் வேலூர்ப்புரட்சியின்	வேலூர்ப்புரட்சியிலிருந்தே	P
12	வைகை அணையில் தற்போது, 1	1958ல் வைகை அணை கட்டப்பட்ட	NP
13	பா.ஜ., பொறுப்பேற்ற ஓராண்டில்	1975ல் இந்த குடும்பத்தில் இருந்து	NP
14	சிறையில் இருக்கும் சில பயங்க	1999 டிசம்பர் 24 ம் தேதி இந்திய வி	NP
15	மும்பை கடற்பகுதியில் ரோந்துட்	2 கடற்படை படகுகள் தீயில் எரிந்த	SP
16	நீதிபதி வேலுமணி வழக்கை விச	2 பேரின் முன்ஜாமீன் மனு ரத்து செ	NP
17	மீனவர்கள் இருவரை சுட்டுக்கொ	2012-ம் ஆண்டு பிப்ரவரி மாதம் இர	NP
18	தி.மு.க., பொருளாளர் ஸ்டாலினுக	2014ல் நடைபெற்ற லோக்சபா தேர்	NP
19	சட்டசபை நாகரீகத்தை காத்து, ப	2016 ஆம் ஆண்டு சட்டப்பேரவைத்	NP
20	வாக்களிக்க பணம் கொடுத்ததா	232 தொகுதிகளிலும் தேர்தலை ரத்	SP
21	பரபரப்பாக எதிர்பார்க்கப்பட்ட, 2	2ஜி வழக்கில் குற்றம் சாட்டப்பட்டு	NP
22	திமுக முன்னாள் மத்திய அமைச்	2ஜி வழக்கில் ராசாவிற்கு எதிராக	NP
23	மின் கம்பத்தில் ஆம்னி பஸ் மோ	3 பேர் பஸ் மோதி பலியானார்கள்.	SP
24	பாகிஸ்தானைச் சேர்ந்த ஒருவரு	35 குழந்தைகள் இருந்தும் 100 குழந்	SP
25	மகளிர் ஹாக்கி அணியின்	ஹாக்கி அணியை ஒலிம்பிக்	P
26	டெல்லியில் மத்திய அமைச்சரவ	4 ஆண்டில் ஒரு கோடி பேருக்கு திற	SP
27	சுற்றுப்பயணத்தை முடித்து	5 நாடுகளில் சுற்றுப்பயணத்தை	P
28	5 பிரிட்டன் ராக்கெட்டுகளுடன் பி	5 பிரிட்டன் ராக்கெட்டுகளுடன் பி.எ	SP
29	சமீபத்தில் நடந்து முடிந்த 5 மாநி	5 மாநில சட்டபேரவைத்தேர்தலில்	SP
30	இந்தியா வர உள்ள பாகிஸ்தான்	5000 கோடி ரூபாய் கடன் வாங்க உ	SP
31	நாட்டிங்காம் டெஸ்டில் ஆஸ்திரே	5வது டெஸ்ட் போட்டி முடிந்ததும் அ	NP
32	புர்ஹான் வானி சுட்டுக் கொல்ல	6-வது நாளாக காஷ்மீரில் இயல்பு வ	SP
33	பத்துலட்சத்திற்கு மேல்வருமான	7 லட்சம் நுகர்வோருக்கு சமையல்	SP

7.3 Language Identification

Brief overview of Datasets -

This task was aimed at identifying 5 closely-related languages of Indo-Aryan language family – Hindi (also known as Khari Boli), Braj Bhasha, Awadhi, Bhojpuri and Magahi. These languages form part of a continuum starting from Western Uttar Pradesh (Hindi and Braj Bhasha) to Eastern Uttar Pradesh (Awadhi and Bhojpuri) and the neighboring Eastern state of Bihar (Bhojpuri and Magahi). For this task, participants were provided with a dataset of approximately 15,000 sentences in each language, mainly from the domain of literature, published over the web as well as in print. It is the first dataset that is being made available for these languages (except Hindi) and it will not only be useful for automatic identification of languages and developing NLP applications but will also help in gaining insights into the proximity level of these languages (which are hypothesized to form part of a continuum and lot of times mistaken as varieties of Hindi, especially outside the scholarly linguistic circles).

This dataset will be used to identify the language of a sentence.

In this dataset every sentence is assigned their language.

The language IDs are the ISO codes of the respective languages and should be read as below -

AWA = Awadhi

BRA = Braj

BHO = Bhojpuri

MAG = Magahi

HIN = Hindi

Datasets Size - Size of training, testing and dev sets:

Size of training set-70351

Size of testing set-9692

Size of dev set-10329

Different classes of labels and their counts:

Sentence

Language: Hindi (HIN), Awadhi (AWA), Bhojpuri (BHO), Braj (BRA) and Magahi (MAG)

Sample Data:

	A	B
1	**Sentence**	**Language**
2	तभी बारिश हुई थी जिसका गीलापन इन मूर्तियों को इन तस्वीरों में एक अलग रूप देता है .	HIN
3	कहते हुए लफ्ज़ बेसुध करते गए कुछ इस तरह /कि इतना नशा तो होता न किसी असली जाम से .	HIN
4	चिट्ठी में ऊ हमरा के होली पर बोलवले रहली.	BHO
5	अब इंग्लेण्ड वाले भी जान गए साथ ही हैरान और परेशान भी हैं ,हमारे यहाँ आकर खुर्शीद जैसे लोग पढ़ गए जो अपने देश के आम आदमी के लिए ही असंविधानिक भाषा का इस्तेमाल करते हैं .	HIN
6	ज्ञान गुण गावन को कलाविक सावन को ।	BRA
7	हमरे ध्यान मां सती मइया केरी छवि उभरी ।	AWA
8	बूंदा-बांदी अब थमिगै रहै ।	AWA
9	कबीर त पुरा क पुरा भारतीय लोक जीवन मे रचल बसल बाटै ।	BHO
10	एक तरफ एक के बाद ने पीढ़ी के एंटी -बायोटिक्स आ रहें हैं दूसरी तरफ जीवाणु अपना रूप विधान तेज़ी से बदल लेता है म्युतेट हो जाता है .	HIN
11	ठंडयाई करसियान में भरि भरि कें छेदन में है कें कूल्लान में दई जाय रही ।	BRA
12	- आईँ हो भोलवा, तुँ आज गिरहतवा से गारी-गुप्ता काहे ला करलहीं हे ।	MAG
13	हाय राम ।	AWA
14	तब ऐसा ही प्रतीत होता है हम हारी हुई लड़ाई ही लड़ रहें हैं .	HIN
15	ऊ कहलक कि तूं गारी देइत हें ?	MAG
16	कहीं कभी किसी दिन हम दोनों मिल जाएंजीवन के किसी राह पर तब क्या हम मिल पायेंगे ठीक पहले की तरह ?	HIN
17	ब्रह्मचारी जी ब्रजभूमि की बा परम्परा के साहित्य सेवी है जब घर - चर में ल्हौरे-ल्हौरे बालकन कूं हजारो - हजारो कवित्त सवैया कण्ठस्थ करवाय दिये जाते हे ।	BRA
18	मान ल कि होइए गइल त का करबू ?	BHO
19	भविष्य में हमकू इनते भौत आसा है ।	BRA
20	जबरदस्ती आपरेशन कराबे को हट करके बैठ गयौ ।	BRA
21	इसमें चमड़ी से काफी तरल निकलके उड़ जाता है तथा तरह तरह के एलर्जी पैदा करने वाले तत्व इस सुरक्षा कवच में सेंध लगाके अन्दर देखिल होने लगते हैं .	HIN
22	दूसरे छोर पर जो भविष्य की संभावनाओं में जीते हैं ,दिवा -स्वप्न -जीवी हैं,सीमाओं को दरकिनार कर सिर्फ संभावनाओं में जीतें हैं ,मैं ये कर दूंगा ,वो कर दूंगा ,ये करूंगा वो करूंगा वह वर्तमान को भी जी नहीं पाते .	HIN
23	ई बात सच हकइ श्रीमान ।	MAG
24	सुभग सरोवर लसत नीर, निरमल सुख कारी ।	BRA
25	हँ अगर रजनेतन में 2-4 फीसदी जननेता के गुन बा त उ कुछ सकारात्मक नियम, योजनन पर विचार त क सकेला पर समाज, देस के असली बिकास जनते करेले।	BHO
26	दोसरा बधार के लोग के त हाँफ उखड़ जाई, साँस टंगा जाई।	BHO
27	उस दौरान अपराधियों का तो सफाया हो गया, लेकिन ज्यादातर ऐसे अपराधी मारे गये, जिनका अपराध कहीं से भी मौत की सजा के काबिल नहीं था ।	HIN
28	तनिक देर मा सबै बिदाई के भाव मा बूड़ि गयीं ।	AWA
29	मैंनें कही " बाबा प्रात: काल के और सुन लेऔ । "	BRA
30	बल्कि हमरे खातिर खुद के खियाल रखथिन ।	MAG

7.4 Hate Speech and Offensive Content Identification in Indo-European Languages

Brief overview of Datasets -

There are two sub-tasks in each of the languages. Below is a brief description of each task.

Sub-task A: Identifying Hate, offensive and profane content

This task focuses on Hate speech and Offensive language identification offered for English, German, and Hindi. Sub-task A is coarse-grained binary classification in which participating system are required to classify tweets into two classes, namely: Hate and Offensive (HOF) and Non- Hate and offensive (NOT).

(NOT) Non-Hate-Offensive - This post does not contain any Hate speech, profane, offensive content.

(HOF) Hate and Offensive - This post contains Hate, offensive, and profane content.

Sub-task B: Discrimination between Hate, profane and offensive posts

This sub-task is a fine-grained classification offered for English, German, and Hindi. Hate-speech and offensive posts from the sub-task A are further classified into three categories:

(HATE) Hate speech: - Posts under this class contain Hate speech content.

(OFFN) Offensive: - Posts under this class contain offensive content.

(PRFN) Profane: - These posts contain profane words.

Categories Explanation:

HATE SPEECH: Describing negative attributes or deficiencies to groups of individuals because they are members of a group (e.g., all poor people are stupid). Hateful comment toward groups because of race, political opinion, sexual orientation, gender, social status, health condition or similar.

OFFENSIVE: Posts which are degrading, dehumanizing, insulting an individual, threatening with violent acts are categorized into OFFENSIVE category.

PROFANITY: Unacceptable language in the absence of insults and abuse. This typically concerns the usage of swearwords (Scheiße, Fuck etc.) and cursing (Zur Hölle! Verdammt! etc.) are categorized into this category.

Datasets Size - Size of training, testing and dev sets:

Size of English training dataset-5853

Size of English testing dataset-1154

Size of English dev dataset-939

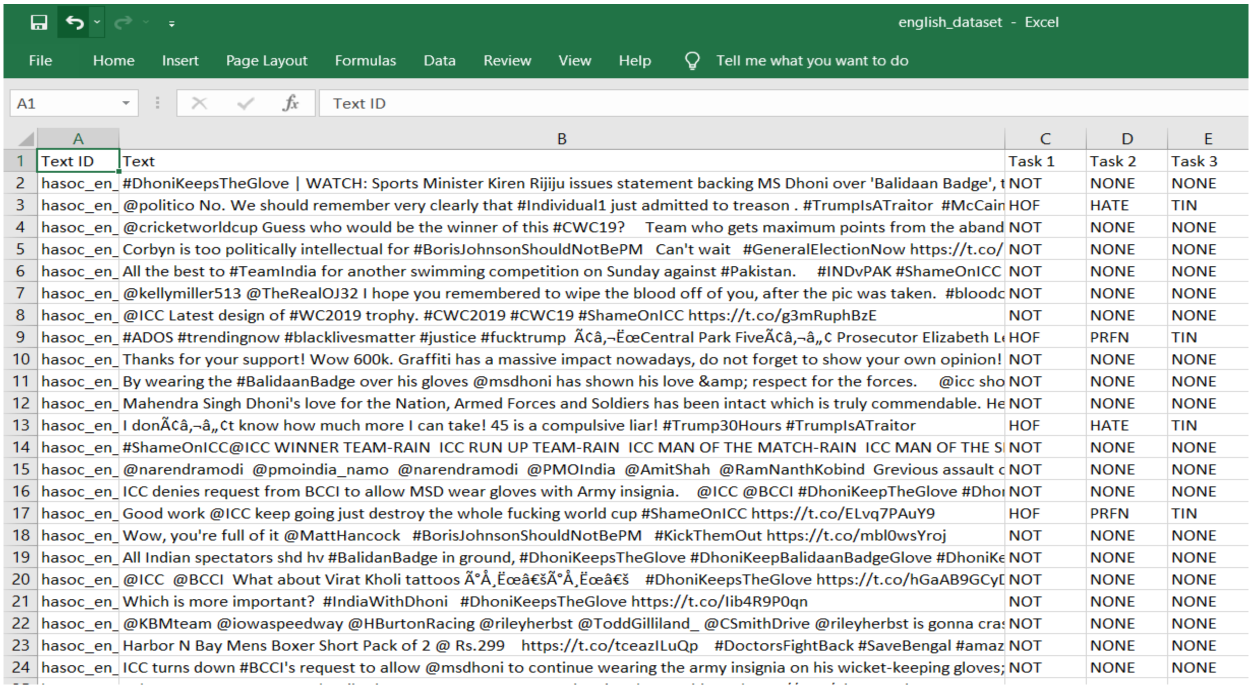
Different classes of labels and their counts:

Text ID, Text

Language: Hindi (HIN), German (GMN), English (ENG)

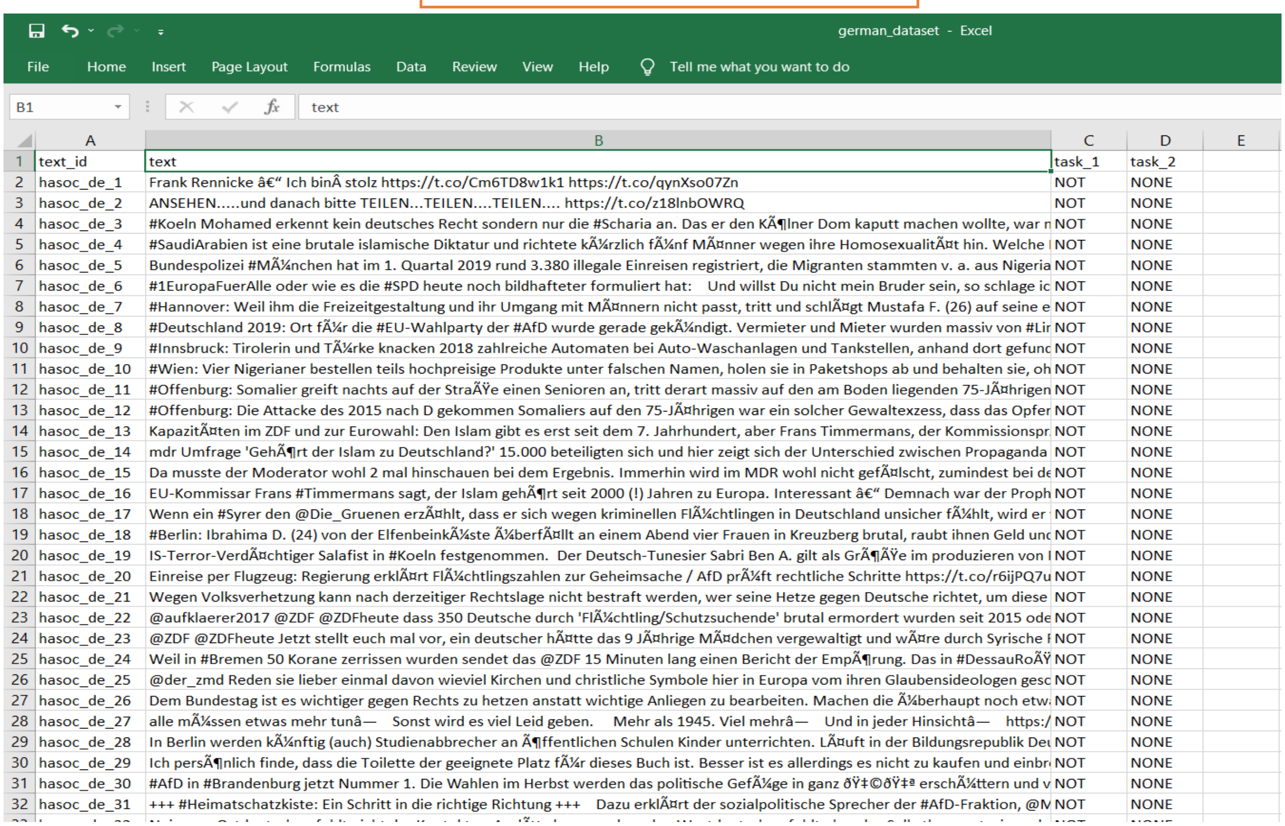
Sample Data:

English Language Based Dataset



	A	B	C	D	E
1	Text ID	Text	Task 1	Task 2	Task 3
2	hasoc_en_	#DhoniKeepsTheGlove | WATCH: Sports Minister Kiren Rijiju issues statement backing MS Dhoni over 'Balidaan Badge', t	NOT	NONE	NONE
3	hasoc_en_	@politico No. We should remember very clearly that #Individual1 just admitted to treason . #TrumpIsATraitor #McCain	HOF	HATE	TIN
4	hasoc_en_	@cricketworldcup Guess who would be the winner of this #CWC19? Team who gets maximum points from the aband	NOT	NONE	NONE
5	hasoc_en_	Corbyn is too politically intellectual for #BorisJohnsonShouldNotBePM Can't wait #GeneralElectionNow https://t.co/	NOT	NONE	NONE
6	hasoc_en_	All the best to #TeamIndia for another swimming competition on Sunday against #Pakistan. #INDvPAK #ShameOnICC	NOT	NONE	NONE
7	hasoc_en_	@kellymiller513 @TheRealOJ32 I hope you remembered to wipe the blood off of you, after the pic was taken. #bloodc	NOT	NONE	NONE
8	hasoc_en_	@ICC Latest design of #WC2019 trophy. #CWC2019 #CWC19 #ShameOnICC https://t.co/g3mRuphBzE	NOT	NONE	NONE
9	hasoc_en_	#ADOS #trendingnow #blacklivesmatter #justice #fucktrump Ã¢â‚¬ËœCentral Park FiveÃ¢â‚¬â„¢ Prosecutor Elizabeth Le	HOF	PRFN	TIN
10	hasoc_en_	Thanks for your support! Wow 600k. Graffiti has a massive impact nowadays, do not forget to show your own opinion!	NOT	NONE	NONE
11	hasoc_en_	By wearing the #BalidaanBadge over his gloves @msdhoni has shown his love & respect for the forces. @icc sho	NOT	NONE	NONE
12	hasoc_en_	Mahendra Singh Dhoni's love for the Nation, Armed Forces and Soldiers has been intact which is truly commendable. He	NOT	NONE	NONE
13	hasoc_en_	I donÃ¢â‚¬â„¢t know how much more I can take! 45 is a compulsive liar! #Trump30Hours #TrumpIsATraitor	HOF	HATE	TIN
14	hasoc_en_	#ShameOnICC@ICC WINNER TEAM-RAIN ICC RUN UP TEAM-RAIN ICC MAN OF THE MATCH-RAIN ICC MAN OF THE S	NOT	NONE	NONE
15	hasoc_en_	@narendramodi @pmoindia_namo @narendramodi @PMOIndia @AmitShah @RamNanthKobind Grevious assault c	NOT	NONE	NONE
16	hasoc_en_	ICC denies request from BCCI to allow MSD wear gloves with Army insignia. @ICC @BCCI #DhoniKeepTheGlove #Dhor	NOT	NONE	NONE
17	hasoc_en_	Good work @ICC keep going just destroy the whole fucking world cup #ShameOnICC https://t.co/ELvq7PAuY9	HOF	PRFN	TIN
18	hasoc_en_	Wow, you're full of it @MattHancock #BorisJohnsonShouldNotBePM #KickThemOut https://t.co/mbl0wsYroj	NOT	NONE	NONE
19	hasoc_en_	All Indian spectators shd hv #BalidanBadge in ground, #DhoniKeepsTheGlove #DhoniKeepBalidaanBadgeGlove #DhoniKe	NOT	NONE	NONE
20	hasoc_en_	@ICC @BCCI What about Virat Kholi tattoos Ã°Å¸Ëœâ€šÃ°Å¸Ëœâ€š #DhoniKeepsTheGlove https://t.co/hGaAB9GCyE	NOT	NONE	NONE
21	hasoc_en_	Which is more important? #IndiaWithDhoni #DhoniKeepsTheGlove https://t.co/Iib4R9P0qn	NOT	NONE	NONE
22	hasoc_en_	@KBMteam @iowaspeedway @HBurtonRacing @rileyherbst @ToddGilliland_ @CSmithDrive @rileyherbst is gonna cras	NOT	NONE	NONE
23	hasoc_en_	Harbor N Bay Mens Boxer Short Pack of 2 @ Rs.299 https://t.co/tceazILuQp #DoctorsFightBack #SaveBengal #amaz	NOT	NONE	NONE
24	hasoc_en_	ICC turns down #BCCI's request to allow @msdhoni to continue wearing the army insignia on his wicket-keeping gloves;	NOT	NONE	NONE

German Language Based Dataset



	A	B	C	D	E
1	text_id	text	task_1	task_2	
2	hasoc_de_1	Frank Rennicke â€“ Ich binÂ stolz https://t.co/Cm6TD8w1k1 https://t.co/qynXso07Zn	NOT	NONE	
3	hasoc_de_2	ANSEHEN.....und danach bitte TEILEN...TEILEN....TEILEN.... https://t.co/z18lnbOWRQ	NOT	NONE	
4	hasoc_de_3	#Koeln Mohamed erkennt kein deutsches Recht sondern nur die #Scharia an. Das er den KÃ¶lner Dom kaputt machen wollte, war n	NOT	NONE	
5	hasoc_de_4	#SaudiArabien ist eine brutale islamische Diktatur und richtete kÃ¼rzlich fÃ¼nf MÃ¤nner wegen ihre HomosexualitÃ¤t hin. Welche	NOT	NONE	
6	hasoc_de_5	Bundespolizei #MÃ¼nchen hat im 1. Quartal 2019 rund 3.380 illegale Einreisen registriert, die Migranten stammten v. a. aus Nigeria	NOT	NONE	
7	hasoc_de_6	#1EuropaFuerAlle oder wie es die #SPD heute noch bildhafteter formuliert hat: Und willst Du nicht mein Bruder sein, so schlage ic	NOT	NONE	
8	hasoc_de_7	#Hannover: Weil ihm die Freizeitgestaltung und ihr Umgang mit MÃ¤nnern nicht passt, tritt und schlÃ¤gt Mustafa F. (26) auf seine e	NOT	NONE	
9	hasoc_de_8	#Deutschland 2019: Ort fÃ¼r die #EU-Wahlparty der #AfD wurde gerade gekÃ¼ndigt. Vermieter und Mieter wurden massiv von #Lir	NOT	NONE	
10	hasoc_de_9	#Innsbruck: Tirolerin und TÃ¼rke knacken 2018 zahlreiche Automaten bei Auto-Waschanlagen und Tankstellen, anhand dort gefund	NOT	NONE	
11	hasoc_de_10	#Wien: Vier Nigerianer bestellen teils hochpreisige Produkte unter falschen Namen, holen sie in Paketshops ab und behalten sie, oh	NOT	NONE	
12	hasoc_de_11	#Offenburg: Somalier greift nachts auf der StraÃŸe einen Senioren an, tritt derart massiv auf den am Boden liegenden 75-JÃ¤hrigen	NOT	NONE	
13	hasoc_de_12	#Offenburg: Die Attacke des 2015 nach D gekommen Somaliers auf den 75-JÃ¤hrigen war ein solcher Gewaltexzess, dass das Opfer	NOT	NONE	
14	hasoc_de_13	KapazitÃ¤ten im ZDF und zur Eurowahl: Den Islam gibt es erst seit dem 7. Jahrhundert, aber Frans Timmermans, der Kommissionspr	NOT	NONE	
15	hasoc_de_14	mdr Umfrage 'GehÃ¶rt der Islam zu Deutschland?' 15.000 beteiligten sich und hier zeigt sich der Unterschied zwischen Propaganda	NOT	NONE	
16	hasoc_de_15	Da musste der Moderator wohl 2 mal hinschauen bei dem Ergebnis. Immerhin wird im MDR wohl nicht gefÃ¤lscht, zumindest bei de	NOT	NONE	
17	hasoc_de_16	EU-Kommissar Frans #Timmermans sagt, der Islam gehÃ¶rt seit 2000 (!) Jahren zu Europa. Interessant â€“ Demnach war der Proph	NOT	NONE	
18	hasoc_de_17	Wenn ein #Syrer den @Die_Gruenen erzÃ¤hlt, dass er sich wegen kriminellen FlÃ¼chtlingen in Deutschland unsicher fÃ¼hlt, wird er	NOT	NONE	
19	hasoc_de_18	#Berlin: Ibrahima D. (24) von der ElfenbeinkÃ¼ste Ã¼berfÃ¤llt an einem Abend vier Frauen in Kreuzberg brutal, raubt ihnen Geld und	NOT	NONE	
20	hasoc_de_19	IS-Terror-VerdÃ¤chtiger Salafist in #Koeln festgenommen. Der Deutsch-Tunesier Sabri Ben A. gilt als GrÃ¶ÃŸe im produzieren von I	NOT	NONE	
21	hasoc_de_20	Einreise per Flugzeug: Regierung erklÃ¤rt FlÃ¼chtlingszahlen zur Geheimsache / AfD prÃ¼ft rechtliche Schritte https://t.co/r6ijPQ7u	NOT	NONE	
22	hasoc_de_21	Wegen Volksverhetzung kann nach derzeitiger Rechtslage nicht bestraft werden, wer seine Hetze gegen Deutsche richtet, um diese	NOT	NONE	
23	hasoc_de_22	@aufklaerer2017 @ZDF @ZDFheute dass 350 Deutsche durch 'FlÃ¼chtling/Schutzsuchende' brutal ermordert wurden seit 2015 ode	NOT	NONE	
24	hasoc_de_23	@ZDF @ZDFheute Jetzt stellt euch mal vor, ein deutscher hÃ¤tte das 9 JÃ¤hrige MÃ¤dchen vergewaltigt und wÃ¤re durch Syrische I	NOT	NONE	
25	hasoc_de_24	Weil in #Bremen 50 Korane zerrissen wurden sendet das @ZDF 15 Minuten lang einen Bericht der EmpÃ¶rung. Das in #DessauRoÃŸ	NOT	NONE	
26	hasoc_de_25	@der_zmd Reden sie lieber einmal davon wieviel Kirchen und christliche Symbole hier in Europa vom ihren Glaubensideologen gesc	NOT	NONE	
27	hasoc_de_26	Dem Bundestag ist es wichtiger gegen Rechts zu hetzen anstatt wichtige Anliegen zu bearbeiten. Machen die Ã¼berhaupt noch etw	NOT	NONE	
28	hasoc_de_27	alle mÃ¼ssen etwas mehr tunâ— Sonst wird es viel Leid geben. Mehr als 1945. Viel mehrâ— Und in jeder Hinsichtâ— https:/	NOT	NONE	
29	hasoc_de_28	In Berlin werden kÃ¼nftig (auch) Studienabbrecher an Ã¶ffentlichen Schulen Kinder unterrichten. LÃ¤uft in der Bildungsrepublik Deu	NOT	NONE	
30	hasoc_de_29	Ich persÃ¶nlich finde, dass die Toilette der geeignete Platz fÃ¼r dieses Buch ist. Besser ist es allerdings es nicht zu kaufen und einbr	NOT	NONE	
31	hasoc_de_30	#AfD in #Brandenburg jetzt Nummer 1. Die Wahlen im Herbst werden das politische GefÃ¼ge in ganz ðŸ‡©ðŸ‡ª erschÃ¼ttern und v	NOT	NONE	
32	hasoc_de_31	+++ #Heimatschatzkiste: Ein Schritt in die richtige Richtung +++ Dazu erklÃ¤rt der sozialpolitische Sprecher der #AfD-Fraktion, @M	NOT	NONE	

Hindi Language Based Dataset

hasoc2019_hi_test_gold_2919 - Excel

File Home Insert Page Layout Formulas Data Review View Help Tell me what you want to do

B1 text

	A	B	C	D	E
1	text_id	text	task_1	task_2	task_3
2	hasoc_hi_5061	वक्त, इन्सान और इंग्लैंड का मौसम आपको कभी भी धोखा दे सकता है। #IndiaVsPakistan	NOT	NONE	NONE
3	hasoc_hi_2090	#कांग्रेस के इस #कमीने की #करतूत को देखिए देश की रक्षा करते #शहीद हुए मां भारती के #सपूतों का ख्याल भी नहीं रखा #दुश्मन से गले मि	HOF	OFFN	TIN
4	hasoc_hi_2960	पाकिस्तान को फेकना था फेका गया। जो हार कर भी दुश्मन को हरा दे उसी को हमने नाम दिया। @BCCI ये तुम्हारे समझ के बाहर हैं तुम बॉल	HOF	OFFN	TIN
5	hasoc_hi_864	जो शब्द तूम आज किसी और औरत के लिए यूज कर रहे वो बचाकर रखना क्योंकि यही कल कोई और कल तुम्हारी माँ बहन और बेटी के लिये	NOT	NONE	NONE
6	hasoc_hi_54	नेता जी हम समाजवादी सिपाही हमेशा आपके साथ है आपका हर लिया गया निर्णय हमे स्वीकार है निवेदन है कि कार्यकर्ताओं पे वरिष्ठ नेता ध्या	NOT	NONE	NONE
7	hasoc_hi_6768	@narendramodi @AmitShah @BJP4India @BharatKaPM @TeamDeol #AayegaTohModiHi #BJP4India #BJP #कन्हैया_कुमार ने खोदा कु	HOF	OFFN	TIN
8	hasoc_hi_163	#कुत्ते भी हो रहे हैं #किडनी और #डायबटीज जैसी #बीमारियों का शिकार, जानें वजह…	NOT	NONE	NONE
9	hasoc_hi_1947	VIVAHIT BaitION KO BHI Patrick KIRSI BHOOMI MAIN BAITON K SAMAN U.P MAIN Adhikar KOUN dilanai KI Kirpa karaiga ? Vicharniy	HOF	HATE	TIN
10	hasoc_hi_3842	10 एजेंसियों को किसी भी कंप्यूटर की निगरानी और डाटा की जांच का अधिकार दिए जाने के खिलाफ दायर याचिकाओं पर सुप्रीम कोर्ट ने सुन	HOF	HATE	TIN
11	hasoc_hi_7408	कश्मीर मे सेना पर पत्थर मारने वाले और पत्थरबाजों के हमदर्द, और, लस्सी न देने पर मर्डर करने वाले लोग पूछ रहे हैं कि बाइक चोर को क्यो	NOT	NONE	NONE
12	hasoc_hi_2771	झोंपड़ी के, कुछ दिन पहले तक तो तू उस आदतन 'बाइक चोर' तबरेज के लिये छाती पीट रहा था। हर चैनल पर रंडी रोना मचा रखा था। अब से	HOF	OFFN	TIN
13	hasoc_hi_4052	अभी सूअर 6 महीने में दिल्ली में विधानसभा चुनाव है अब हिंदू मुसलमान नहीं होगा तो कब होगा जिसने भी किया है उसे सजा मिलेगी दिल्ली पु	HOF	HATE	TIN
14	hasoc_hi_6068	कुछ कमीने #नीच ग्रह में पैदा होते है और वो पैदा होते ही अपने परिवार वालो को अपने #बाप को आस पास के लोगो को नही वरन सारी #काय	HOF	OFFN	TIN
15	hasoc_hi_715	भाई जैसे दिल्ली मे चुनाव आएगा हिन्दू-मुस्लिम की बाते जोरो से होने लगेंगी? हिदुत्व भी खतरे मे जा सकता है क्योंकि जनता स्कूल हॉस्पिटल जैस	HOF	OFFN	TIN
16	hasoc_hi_2180	Pakode talna chalu kr do... mudra scheme se paise mil jayenge.. Bua babua	NOT	NONE	NONE
17	hasoc_hi_2	बेटा धोनी के बारे में कुछ बुरा बोले ना तो क्रिकेट की कसम तुमारी गांड तोड़ देंगे	HOF	PRFN	TIN
18	hasoc_hi_5233	आदरणीय राष्ट्रीय अध्यक्ष जी से विनम्र निवेदन है उत्तर प्रदेश की बिगड़ती कानून व्यवस्था को लेकर जन आंदोलन छेड़ना चाहिए हम सब आपके	NOT	NONE	NONE
19	hasoc_hi_770	देश के #अंदरूनी_दुश्मन_राहुल_गांधी से निपटने के लिये तो,,, #भारतीय_सेना के #डॉग_यूनिट के #कुत्ते ही काफ़ी हैं... !!!	HOF	OFFN	TIN
20	hasoc_hi_6873	भाई UP में राम राज्य उबाल मार रहा है? हैवानियत और तानशाही रुके से नही रुक रही है। सरकार हिदुत्व पे वोट लेकर CM से लेकर विधायक	HOF	PRFN	TIN
21	hasoc_hi_3409	संत कबीर डिग्रीधारी नहीं थे लेकिन धार्मिक आडंबरों से मुक्त थे जबकि बहुजन समाज डिग्रीधारी व नौकरी प्राप्त फिर भी तमाम आडंबरों के की	NOT	NONE	NONE
22	hasoc_hi_447	@azkhawaja1 Beta jispar tu bhonk raha hai na wo aisi shakhsiyat hai ki jab wo choka marte the tab boundary se ball lane tumhara P	HOF	OFFN	TIN
23	hasoc_hi_6801	दिल्ली में 300 मुल्ले मंदिर तोड़ते है ओर ये मुल्लो का पिछवाड़ा चाटने वाला केजरीवाल झारखंड के चोर मुल्ले को 5लाख देता है भड़वा	HOF	HATE	TIN
24	hasoc_hi_5885	रंडी के बच्चे, बिना बाप के, मादरचोद पाकिस्तानी जात के कुत्ते, जाके देख रिकॉर्ड में कैसे मैया चोदता था तुम्हारी ये, भिखारी देश के भिखारी	HOF	PRFN	TIN
25	hasoc_hi_7187	LP Verma वर्मा जी इस समय जिसके स्पोटर है वह लोग नवाजबादी हो गये है	HOF	OFFN	UNT
26	hasoc_hi_3396	*वो अपने मोहल्ले को छोटा पाकिस्तान कहते रहे, हम चुप रहे।* *वो पाकिस्तान की मैच जीतने पर पटाखे फोड़ते रहे, हम चुप रहे।* *आज व	HOF	HATE	TIN
27	hasoc_hi_3936	दिल बड़ा कर पगली बाते तो मेरे #कमीने DOST भी बड़ी बड़ी कर लेते हैं #FRIENDSHIP	NOT	NONE	NONE
28	hasoc_hi_7459	जिन्होंने उर्दू अख़बार वन्दे मातरम् के 1931 एडिशन में हिंदू सम्पादक का नाम देखकर @AalokTweet को टार्गेट किया वो जान लें कि बँटवारे से	NOT	NONE	NONE
29	hasoc_hi_5886	ये देखो mc कांग्रेस में आते ही धर्म भी बदल लिया सीखो को चाहिए इसे लात मार कर बाहर कर दे @TajinderBagga जी सिख समुदाय से कहिये	HOF	OFFN	TIN
30	hasoc_hi_5936	इस्लाम सच्चा और अच्छा धर्म है, यही वजह है कि मै खुद अपने बच्चों को क़ुरआन की तालीम दिला रही हूँ :- अनीता नायर, लेखिका http://tee	NOT	NONE	NONE
31	hasoc_hi_7240	#Budget2019 आने में अब बस चंद दिन दूर हैं। आखिर मोदी सरकार के दूसरे कार्यकाल में क्या है आम आदमी के मन की बात, देखिए इस वीडि	NOT	NONE	NONE
32	hasoc_hi_674	ड्रामा क्यों करते हैं @NitishKumar जी? क्या आपको नहीं पता है कि भाजपा #प्रज्ञा_ठाकुर जैसे गोड्सेवादियों की शरणस्थली है! आपने सिर्फ सत	HOF	OFFN	TIN
33	hasoc_hi_1882	धमकी किसे दे रहा है बे. सूअर की औलाद.	HOF	OFFN	UNT

7.5 Sentiment Analysis

Brief overview of Datasets -

This dataset is based on Movie Review sentiment dataset for sentence level sentiment classification Review sentiment dataset for Aspect term extraction and sentiment classification Aspect term category detection and sentiment classification sentence level sentiment classification.

In this dataset every sentence is assigned in Hindi language.

Datasets Size - Size of training, testing and dev sets:

Size of Movie Review training dataset-2152

Size of testing set-643

Size of dev set-454

Different classes of labels and their counts:

id, polarity, text

Language: Hindi (HIN)

Sample Data:

	id	polarity	text
1	id	polarity	text
2	1	conflict	रेमो डिसूज़ा की फिल्म 'एबीसीडी 2' में डांस है, जोश है, देशभक्ति और अभिमान भी है, वंदे मातरम है, थोड़ा रोमांस भी है...फिल्मों में इनसे ही अलग-अलग संतुष्ट होना पड़ेगा।
3	2	neutral	दृश्ये और प्रसंग के हिसाब से इन भाव और भावनाओं से मनोरंजन होता है।
4	3	neutral	फिल्म सच्ची कहानी पर आधारित है तो कुछ सच्चाई वहां से ले लेनी थी।
5	4	positive	मुश्किल स्टेप के अच्छे डांस सिक्वेंस हैं।
6	5	negative	कहानी के अभाव में उनके व्यक्तिगत प्रयास का प्रभाव कम हो जाता है।
7	6	neutral	पास बैठे युवा दर्शक की टिप्पणी थी कि इन दिनों यूट्यूब पर ऐसे डांस देखे जा सकते हैं।
8	7	neutral	मुझे डांस के साथ कहानी भी दिखाओ।
9	8	neutral	सुरु और विन्नी मुंबई के उपनगरीय इलाके के युवक हैं।
10	9	neutral	डांस के दीवाने सुरु और विन्नी का एक ग्रुप है।
11	10	neutral	वे स्थानीय स्तुर पर 'हम किसी से कम नहीं' कंपीटिशन में हिस्सा लेते हैं।
12	11	negative	नकल के आरोप में वहां उनकी छंटाई हो जाती है।
13	12	negative	उसकी वजह से जगहंसाई भी होती है।
14	13	positive	इस तौहीन के बावजूद उनका जोश ठंडा नहीं होता।
15	14	neutral	वे मशहूर डांसर विष्णु को ट्रेनिंग के लिए राजी करते हैं (विष्णु सर को राजी करने के दृश्य में दोहराव से फिल्म खिंचती है।
16	15	neutral	बहरहाल, विषणु सर राजी तो होते हैं, लेकिन उनका कुछ और गेम प्लान है।
17	16	neutral	फिल्म में आगे वह जाहिर होता है, फिर भी ड्रामा नहीं बन पाता।
18	17	negative	'एबीसीडी 2' में ड्रामा की कमी है, इस कमी की वजह से ही सुरु और विन्नी का रोमांस भी नहीं उभर पाता।
19	18	negative	नतीजतन दोनों के बीच का रोमांटिक सॉन्ग अचानक और गैरजरूरी लगता है।
20	19	positive	प्रभु देवा फिल्मों के बेहतरीन डांसर हैं।
21	20	positive	बतौर निर्देशक उनकी कुछ फिल्में सफल भी रही हैं।
22	21	negative	किंतु खुद अभिनय करते समय वे निराश करते हैं।
23	22	negative	उनके लिए कुछ नाटकीय दृश्य रचे भी गए है, लेकिन वे उसमें परफार्म नहीं कर सके हैं।
24	23	positive	डांस के गुरू के तौर उन्हें कुछ सीन मिल सकते थे।
25	24	positive	वरुण धवन और श्रद्धा कपूर ने मुश्किल स्टेप भी सहज तरीके से अपनाए हैं।
26	25	neutral	एक्टर दृश्यों को एंजॉय करें तो इफेक्ट गहरा होता है।
27	26	neutral	दोनों अभी दो-तीन फिल्मो ही पुराने हैं।
28	27	positive	दिख रहा है कि वे हर फिल्म के साथ आगे बढ़ रहे हैं।
29	28	positive	पर्दे पर लगातार दिखने से स्वीकृति बढ़ती है और परफार्मेंस पसंद आ जाए तो स्टारडम में इजाफा होता है।
30	29	positive	सुरु और विन्नी की टीम के अन्य सदस्य भी डांस के दृश्यों में उचित योगदान करते हैं।
31	30	positive	उनका डांस सिक्वेंस नयनाभिरामी (आई कैचिंग) है।
32	31	positive	रेमो डिसूज़ा ने फिल्मा में भव्यता रखी है।

	A	B	C	D	E
1	text	aspectCate	aspectCate	_id	_polarity
2	फेसबुक का सिक्योरिटी चेकअप फीचर पॉपअप की तरह यूजर्स को दिखाइ देगा।	gui	neu	app_2	neu
3	इस पॉपअप बॉक्स में पासवर्ड चेंज करने, लॉगइन अलर्ट्स चालू करने और मौजूदा फेसबुक सेशन को पूर्	misc	neu	app_3	neu
4	आप यहीं से पुराने लॉगइन सेशंस को हमेशा के लिए हटा भी सकते हैं।	gui	neu	app_4	neu
5	ऐसे में कई बार ऐसा होता है कि आपका लॉगइन ओपन ही रह जाता है और कोई भी इसका दुरुपयोग कर सकता।			app_5	neu
6	अब नए फीचर से इस तरह की गलतियां नहीं होंगी और अकाउंट को हेक करना बेहद मुश्किल हो जाएगा।			app_6	pos
7	एंड्रॉयड 4.4.4 से ऊपर के वर्जन पर करता है काम ।	misc	neu	app_7	neu
8	Twitter periscope App एंड्रॉयड 4.4.4 किटकैट या उससे ऊपर के वर्जन पर काम करता है।	misc	neu	app_8	neu
9	कंपनी के मुताबिक, एंड्रॉयड के लिए आए पेरिस्कोप में सभी कोर फीचर्स हैं जो यूजर्स द्वारा पसंद किये गए हैं।			app_9	pos
10	पेरिस्कोप के एंड्रॉयड वर्जन में कुछ ऐसे यूनिक फीचर्स भी हैं दिए गए हैं जो आईओएस में नहीं दिए गए थे।	misc	neu	app_10	neu
11	इसमें मेटेरियल इन्स्पायर्ड डिजाइन है जो एंड्रॉयड यूजर्स को मॉडर्न लगेगा।	gui	neu	app_11	neu
12	पेरिस्कोप एप यूजर्स भी अडिशनल पुश नोटिफिकेशन सेटिंग्स अपने हिसाब से सेट कर सकते हैं जैसे फर	gui	pos	app_12	pos
13	पेरिस्कोप एप में रिज्यूम नोटिफिकेशन फीचर दिया गया है जो बहुत ही खास है।	gui	pos	app_13	pos
14	इस फीचर के तहत यूजर ब्रॉडकास्ट के जरिए देखे जाने वाले वीडियो रूकने पर फिर वहीं से देख सकते	misc	pos	app_14	pos
15	इसके अलावा इसके दूसरे फीचर्स में बिना ब्रॉडकास्टर के फाइल अपलोड किये भी रिप्पले सेव हो जाते है	misc	neu	app_15	neu
16	इस फीचर के तहत ब्रॉडकास्टर का टाइम और मोबाइल डेटा दोनों की बचत होती है।	misc	pos	app_16	pos
17	आपके स्मार्टफोन में ज्यादातर एप ऐसे ही हैं जिनसे आपको अपनी जेब का पैसा खर्च कर ही कुछ काम f	price	pos	app_17	pos
18	गूगल प्ले स्टोर पर इस एप को फ्री में अपने मोबाइल में डाउनलोड किया जा सकता है।	price	neu	app_18	neu
19	स्लिक इंटरफेस है।	gui	pos	app_19	pos
20	मैसेंजर अब ब्लैकबेरी स्मार्टफोन तक ही सीमित नहीं है।	ease_of_u	pos	app_20	pos
21	3जी पर मैसेज डिलीवरी व्हाट्सप्प और आईमैसेज की तुलना में बहुत तेज है।	ease_of_u	pos	app_21	pos
22	व्हाट्सप्प जैसा फीचर है।	misc	pos	app_22	pos
23	ऑटो करेक्टिंग के दौरान स्टॉक आईओएस कीबोर्ड अलग तरह से काम करता है।	gui	neu	app_23	neu
24	प्रतिस्पर्धियों पर भारी पड़ने के लिए वॉयस और वीडियो कॉल्स की जरूरत है।	misc	neg	app_24	neg
25	यह किसी भी साधन से प्राप्त एक क्रांतिकारी मैसेंजर नहीं, लेकिन मैसेंजर क्लाएंट से भरपूर एक इको सि	ease_of_u	neu	app_25	neu
26	यह शायद सबसे अच्छी बात है जो उन सभी यूजर्स के साथ हो सकती है जो प्लेटफॉर्म बदलना चाहते हैं ले	misc	pos	app_26	pos
27	चूंकि इसका गोइंग अच्छा था, अत: मैंने एप्पल एप्प स्टोर से बीबीएम को डाउनलोड कर लिया।	misc	pos	app_27	pos
28	जिस भी व्यक्ति ने एक ओएस के रूप में ब्लैकबेरी 10 का अनुभव नहीं किया है, इस पर बीबीएम का उप	gui	neu	app_28	neu
29	हालांकि इसके अधिकतर भाग सही स्टैंडर्ड रूप में दिखते हैं, फिर भी पूरे एप्प में साइड स्वाइप्स, एक्सेस और कार्यशैली का एक			app_29	pos
30	स्क्रीन के नीचे दिया गया टैब आपको चेट्स, कॉन्टेक्ट्स और ग्रुप्स के बीच नेविगेट करने की सुविधा देता है	misc	pos	app_30	pos
31	पूरे एप्प में, बायें और दायें हाथ पर स्थित साइड मेन्यू से अधिक से अधिक ऑप्शन्स को एक्सेस किया जा	misc	neu	app_31	neu
32	आपकी रूचि बनाये रखने के लिए, इस एप्प में काले, आसमानी, नीले और सफेद रंग के मिश्रण से युक्त प	misc	pos	app_32	pos
33	कॉन्टैक्ट लिस्ट या तो छोटे कॉन्टैक्ट इमेज के साथ हो सकती है या आपके कॉन्टैक्ट में प्रोफाइल पिक्चर के रूप में स्थित हाइलाइ			app_33	neu

7.6 Parallel Translation

Brief overview of Datasets -

The IIT Bombay English-Hindi corpus contains parallel corpus for English-Hindi as well as monolingual Hindi corpus collected from a variety of existing sources and corpora developed at the Centre for Indian Language Technology, IIT Bombay over the years. This page describes the corpus. This corpus has been used at the Workshop on Asian Language Translation Shared Task since 2016 the Hindi-to-English and English-to-Hindi languages pairs and as a pivot language pair for the Hindi-to-Japanese and Japanese-to-Hindi language pairs.

For e.g., Hindi (HIN), English (ENG)

Datasets Size - Size of training, testing and dev sets:

Size of training set-104858

Size of testing set-29692

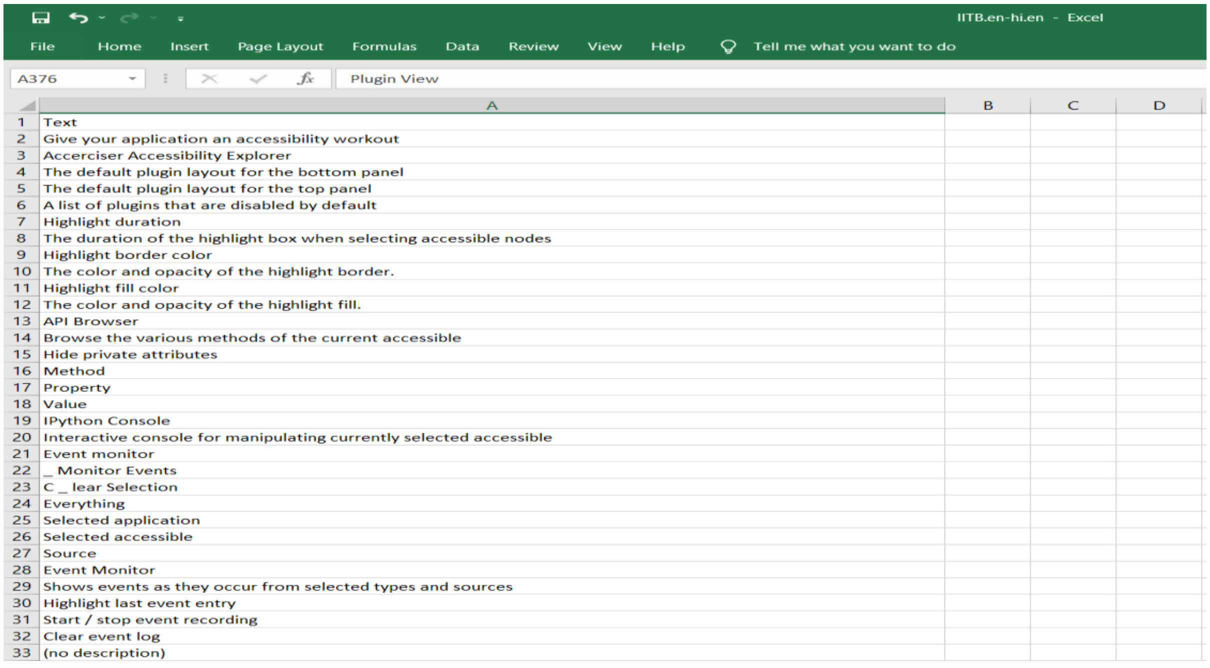
Size of dev set-20329

Different classes of labels and their counts:

Text

Language: Hindi (HIN), English (ENG)

Sample Data:

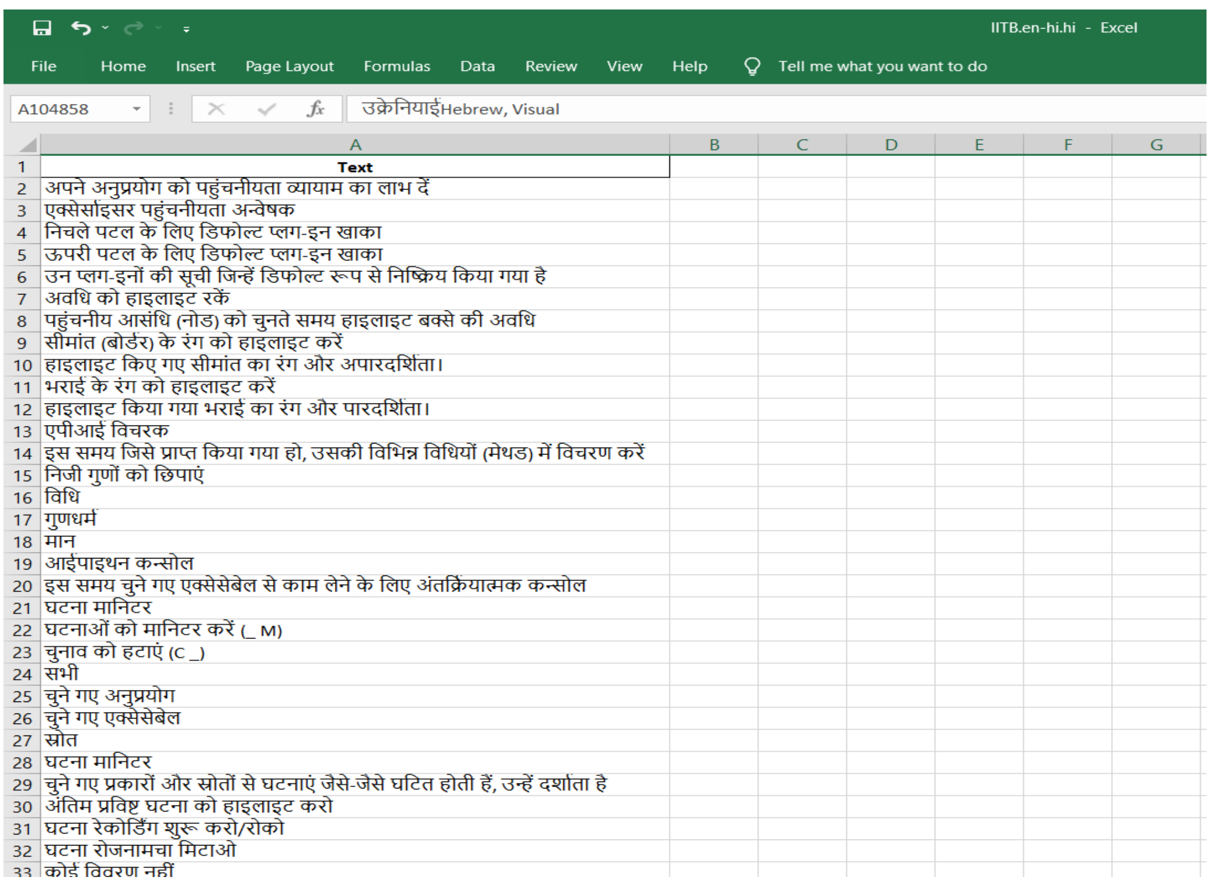


IITB.en-hi.en - Excel

File | Home | Insert | Page Layout | Formulas | Data | Review | View | Help | Tell me what you want to do

A376 | Plugin View

	A	B	C	D
1	Text			
2	Give your application an accessibility workout			
3	Accerciser Accessibility Explorer			
4	The default plugin layout for the bottom panel			
5	The default plugin layout for the top panel			
6	A list of plugins that are disabled by default			
7	Highlight duration			
8	The duration of the highlight box when selecting accessible nodes			
9	Highlight border color			
10	The color and opacity of the highlight border.			
11	Highlight fill color			
12	The color and opacity of the highlight fill.			
13	API Browser			
14	Browse the various methods of the current accessible			
15	Hide private attributes			
16	Method			
17	Property			
18	Value			
19	IPython Console			
20	Interactive console for manipulating currently selected accessible			
21	Event monitor			
22	_ Monitor Events			
23	C _ lear Selection			
24	Everything			
25	Selected application			
26	Selected accessible			
27	Source			
28	Event Monitor			
29	Shows events as they occur from selected types and sources			
30	Highlight last event entry			
31	Start / stop event recording			
32	Clear event log			
33	(no description)			



IITB.en-hi.hi - Excel

File | Home | Insert | Page Layout | Formulas | Data | Review | View | Help | Tell me what you want to do

A104858 | उक्रेनियाईHebrew, Visual

	A	B	C	D	E	F	G
1	Text						
2	अपने अनुप्रयोग को पहुंचनीयता व्यायाम का लाभ दें						
3	एक्सेसाइसर पहुंचनीयता अन्वेषक						
4	निचले पटल के लिए डिफोल्ट प्लग-इन खाका						
5	ऊपरी पटल के लिए डिफोल्ट प्लग-इन खाका						
6	उन प्लग-इनों की सूची जिन्हें डिफोल्ट रूप से निष्क्रिय किया गया है						
7	अवधि को हाइलाइट रकें						
8	पहुंचनीय आसंधि (नोड) को चुनते समय हाइलाइट बक्से की अवधि						
9	सीमांत (बोर्डर) के रंग को हाइलाइट करें						
10	हाइलाइट किए गए सीमांत का रंग और अपारदर्शिता।						
11	भराई के रंग को हाइलाइट करें						
12	हाइलाइट किया गया भराई का रंग और पारदर्शिता।						
13	एपीआई विचरक						
14	इस समय जिसे प्राप्त किया गया हो, उसकी विभिन्न विधियों (मेथड) में विचरण करें						
15	निजी गुणों को छिपाएं						
16	विधि						
17	गुणधर्म						
18	मान						
19	आईपाइथन कन्सोल						
20	इस समय चुने गए एक्सेसेबेल से काम लेने के लिए अंतर्क्रियात्मक कन्सोल						
21	घटना मानिटर						
22	घटनाओं को मानिटर करें (_ M)						
23	चुनाव को हटाएं (C _)						
24	सभी						
25	चुने गए अनुप्रयोग						
26	चुने गए एक्सेसेबेल						
27	स्रोत						
28	घटना मानिटर						
29	चुने गए प्रकारों और स्रोतों से घटनाएं जैसे-जैसे घटित होती हैं, उन्हें दर्शाता है						
30	अंतिम प्रविष्ट घटना को हाइलाइट करो						
31	घटना रेकोर्डिंग शुरू करो/रोको						
32	घटना रोजनामचा मिटाओ						
33	कोई विवरण नहीं						

7.7 HindEnCorp 0.5

Brief overview of Datasets -

HindEnCorp parallel texts (sentence-aligned) come from the following sources: Tides, which contains 50K sentence pairs taken mainly from news articles. This dataset was originally collected for the DARPA-TIDES surprise-language con- test in 2002, later refined at IIIT Hyderabad and provided for the NLP Tools Contest at ICON 2008 (Venkatapathy, 2008).

Commentaries by Daniel Pipes contain 322 articles in English written by a journalist Daniel Pipes and translated into Hindi.

EMILLE. This corpus (Baker et al., 2002) consists of three components: monolingual, parallel and annotated corpora. There are fourteen monolingual sub- corpora, including both written and (for some languages) spoken data for fourteen South Asian languages. The EMILLE monolingual corpora contain in total 92,799,000 words (including 2,627,000 words of transcribed spoken data for Bengali, Gujarati, Hindi, Punjabi and Urdu). The parallel corpus consists of 200,000 words of text in English and its accompanying translations into Hindi and other languages.

Smaller datasets as collected by Bojar et al. (2010) include the corpus used at ACL 2005 (a sub corpus of EMILLE), a corpus of named entities from Wikipedia (crawled in 2009), and Agriculture domain parallel corpus.

For the current release, we are extending the parallel corpus using these sources: Intercorp (Čermák and Rosen,2012) is a large multilingual parallel corpus of 32 languages including Hindi. The central language used for alignment is Czech. Intercorp's core texts amount to 202 million words. These core texts are most suitable for us because their sentence alignment is manually checked and therefore very reliable. They cover predominately short stories and novels. There are seven Hindi texts in Inter- corp. Unfortunately, only for three of them the English translation is available; the other four are aligned only with Czech texts. The Hindi sub corpus of Intercorp contains 118,000 words in Hindi.

Other smaller datasets. This time, we added Wikipedia entities as crawled in 2013 (including any morphological variants of the named entity that appears on the Hindi variant of the Wikipedia page) and words, word examples and quotes from the Shabdkosh online dictionary.

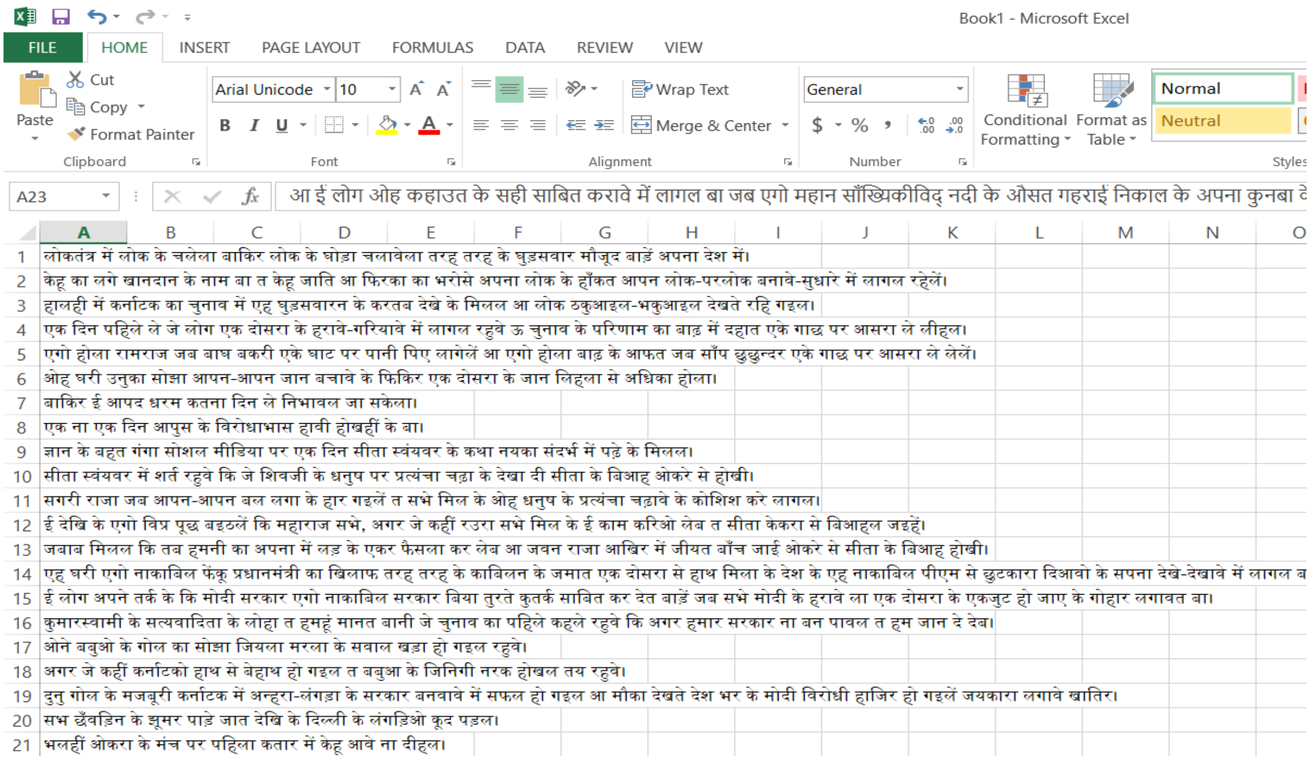
Datasets Size - Size of training, testing and dev sets:

Size of training set-118,000

Size of testing set-27692

Size of dev set-20569

Sample Data:



Bhojpuri

7.8 Word Similarity

Brief overview of Datasets –

Distributional semantics in the form of word embeddings are an essential ingredient to many modern natural language processing systems. The quantification of semantic similarity between words can be used to evaluate the ability of a system to perform semantic interpretation. To this end, a number of word similarity datasets have been created for the English language over the last decades. For Thai language few such resources are available. In this work, we create three Thai word similarity datasets by translating and re-rating the popular WordSim-353, SimLex-999 and SemEval-2017-Task-2 datasets. The three datasets contain 1852-word pairs in total and have different characteristics in terms of difficulty, domain coverage, and notion of similarity (relatedness vs. similarity). These features help to gain a broader picture of the properties of an evaluated word embedding model. We include baseline evaluations with existing Thai embedding models, and identify the high ratio of out-of-vocabulary words as one of the biggest challenges in the evaluation process. All datasets, evaluation results, and a tool for easy evaluation of new Thai embedding models are available to the NLP community online.

Distributional semantics in the form of word embeddings are an essential ingredient to many modern natural language processing systems. The quantification of semantic similarity between words can be used to evaluate the ability of a system to perform semantic interpretation. To this end, a number of word similarity datasets have been created for the English language over the last decades. For Thai language few such resources are available. In this work, we create three Thai word similarity datasets by translating and re-rating the popular WordSim-353, SimLex-999 and SemEval-2017-Task-2 datasets. The three datasets contain 1852-word pairs in total and have different characteristics in terms of difficulty, domain coverage, and notion of similarity (relatedness vs. similarity). These features help to gain a broader picture of the properties of an evaluated word embedding model. We include baseline evaluations with existing Thai embedding models, and identify the high ratio of out-of-vocabulary words as one of the biggest challenges in the evaluation process. All datasets, evaluation results, and a tool for easy evaluation of new Thai embedding models are available to the NLP community online.

Distributional semantics in the form of word embeddings are an essential ingredient to many modern natural language processing systems. The quantification of semantic similarity between words can be used to evaluate the ability of a system to perform semantic interpretation. To this end, a number of word similarity datasets have been created for the English language over the last

decades. For Thai language few such resources are available. In this work, we create three Thai word similarity datasets by translating and re-rating the popular WordSim-353, SimLex-999 and SemEval-2017-Task-2 datasets. The three datasets contain 1852-word pairs in total and have different characteristics in terms of difficulty, domain coverage, and notion of similarity (relatedness vs. similarity). These features help to gain a broader picture of the properties of an evaluated word embedding model. We include baseline evaluations with existing Thai embedding models, and identify the high ratio of out-of-vocabulary words as one of the biggest challenges in the evaluation process. All datasets, evaluation results, and a tool for easy evaluation of new Thai embedding models are available to the NLP community online.

Distributional semantics in the form of word embeddings are an essential ingredient to many modern natural language processing systems. The quantification of semantic similarity between words can be used to evaluate the ability of a system to perform semantic interpretation. To this end, a number of word similarity datasets have been created for the English language over the last decades. For Thai language few such resources are available. In this work, we create three Thai word similarity datasets by translating and re-rating the popular WordSim-353, SimLex-999 and SemEval-2017-Task-2 datasets. The three datasets contain 1852-word pairs in total and have different characteristics in terms of difficulty, domain coverage, and notion of similarity (relatedness vs. similarity). These features help to gain a broader picture of the properties of an evaluated word embedding model. We include baseline evaluations with existing Thai embedding models, and identify the high ratio of out-of-vocabulary words as one of the biggest challenges in the evaluation process. All datasets, evaluation results, and a tool for easy evaluation of new Thai embedding models are available to the NLP community online.

Distributional semantics in the form of word embeddings are an essential ingredient to many modern natural language processing systems. The quantification of semantic similarity between words can be used to evaluate the ability of a system to perform semantic interpretation. To this end, a number of word similarity datasets have been created for the English language over the last decades. The three datasets contain 1852-word pairs in total and have different characteristics in terms of difficulty, domain coverage, and notion of similarity (relatedness vs. similarity). These features help to gain a broader picture of the properties of an evaluated word embedding model. We include baseline evaluations with existing Thai embedding models, and identify the high ratio of out-of-vocabulary words as one of the biggest challenges in the evaluation process. All datasets, evaluation results, and a tool for easy evaluation of new Thai embedding models are available to the NLP community online.

Distributional semantics in the form of word embeddings are an essential ingredient to many modern natural language processing systems. The quantification of semantic similarity between words can be used to evaluate the ability of a system to perform semantic interpretation. To this end, a number of word similarity datasets have been created for the English language over the last decades. For Thai language few such resources are available. In this work, we create three Thai word similarity datasets by translating and re-rating the popular WordSim-353, SimLex-999 and SemEval-2017-Task-2 datasets. The three datasets contain 1852-word pairs in total and have different characteristics in terms of difficulty, domain coverage, and notion of similarity (relatedness vs. similarity). These features help to gain a broader picture of the properties of an evaluated word embedding model. We include baseline evaluations with existing Thai embedding models, and identify the high ratio of out-of-vocabulary words as one of the biggest challenges in the evaluation process. All datasets, evaluation results, and a tool for easy evaluation of new Thai embedding models are available to the NLP community online.

The capacity to quantify the degree of semantic similarity between terms is an archetypal way to evaluate the ability of a system to perform semantic interpretation. This operation of lightweight semantic interpretation is applicable in many scenes. The capacity to quantify the degree of semantic similarity between terms is an archetypal way to evaluate the ability of a system to perform semantic interpretation. This operation of lightweight semantic interpretation is applicable in many scenes. The capacity to quantify the degree of semantic similarity between terms is an archetypal way to evaluate the ability of a system to perform semantic interpretation. This operation of lightweight semantic interpretation is applicable in many scenes. The capacity to quantify the degree of semantic similarity between terms is an archetypal way to evaluate the ability of a system to perform semantic interpretation. This operation of lightweight semantic interpretation is applicable in many scenes. The capacity to quantify the degree of semantic similarity between terms is an archetypal way to evaluate the ability of a system to perform semantic interpretation. This operation of lightweight semantic interpretation is applicable in many scenes used to evaluate the ability of a system to perform semantic interpretation. To this end, a number of words used to evaluate the ability of a system to perform semantic interpretation. To this end, a number of words used to evaluate the ability of a system to perform semantic interpretation. To this end, a number of word Distributional semantics in the form of word embeddings are an essential ingredient to many modern natural language processing systems. The quantification of semantic similarity between words can be used to evaluate the ability of a system to perform semantic interpretation. To this end, a number of word similarity datasets have been created for the English language over the last decades. For Thai language few such resources are available. In this work, we create three Thai word similarity datasets by translating and re-rating the popular WordSim-353, SimLex-999 and SemEval-2017-Task-2 datasets. The three datasets contain 1852-word pairs in total and have different characteristics in terms of difficulty, domain coverage, and notion of similarity (relatedness vs. similarity). These features help to gain a broader picture of the properties of an evaluated word embedding model. We include baseline evaluations with existing Thai embedding models, and identify the high ratio of out-of-vocabulary words as one of the biggest challenges in the evaluation process. All datasets, evaluation results, and a tool for easy evaluation of new Thai embedding models are available to the NLP community online.

Some of the well-known existing word similarity datasets include RG-65 (containing 65-word pairs), WordSim353 (353 pairs), and SimLex-999 (with 999-word pairs). Typically, word similarity gold-standards were initially created for the English language only, although in recent years there have been increased efforts to translate some of the datasets into various other European languages Chinese, Indian languages, etc. However, for Thai language, to the best of our knowledge, there only exists a very small dataset (65-word pairs) by Osathanunkul et al based on the Rubenstein & Goodenough's RG- 65 dataset – which is too limited in size and other aspects like domain coverage to allow a comprehensive evaluation of Thai word embedding models.

Some of the well-known existing word similarity datasets include RG-65 (containing 65-word pairs), WordSim-353 (353 pairs), and SimLex-999 (with 999-word pairs). Typically, word similarity gold-standards were initially created for the English language only, although in re-cent years there have been increased efforts to translate some of the datasets into various other European languages Chinese, Indian languages etc. However, for Thai language, to the best of our knowledge, there only exists a very small dataset (65-word pairs) by Osathanunkuletal based on the Rubenstein & Goodenough's RG-65 dataset – which is too limited in size and other aspects like domain coverage to allow a comprehensive evaluation of Thai word embedding models.

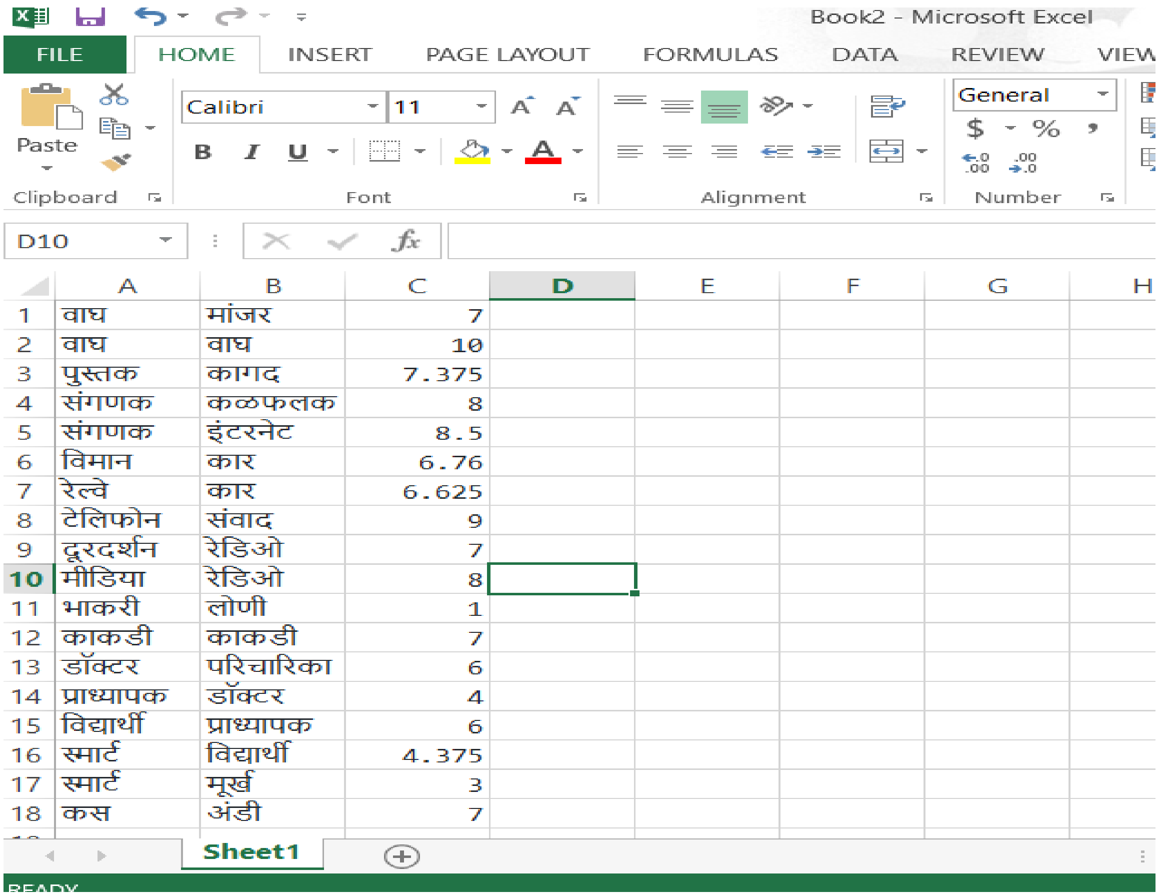
Datasets Size - Size of training, testing and dev sets:

Size of training set-945(each language)

Size of testing set-270(each language)

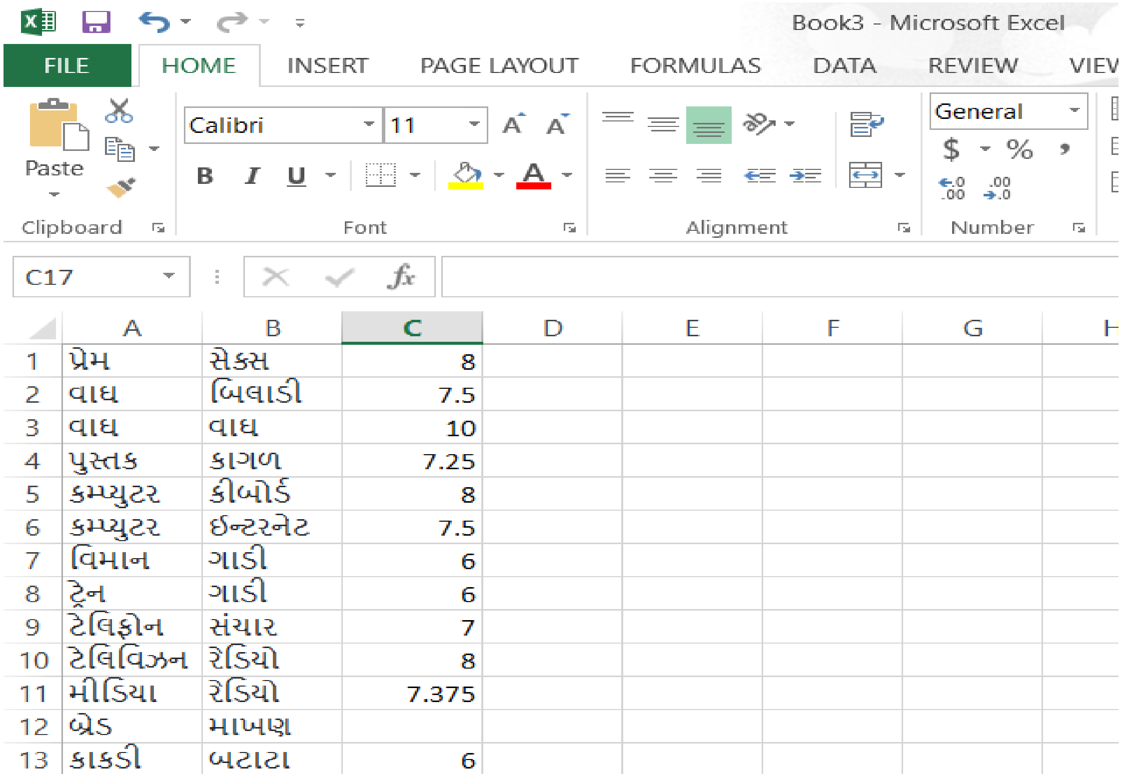
Size of dev set-205(each language)

Sample Dataset:



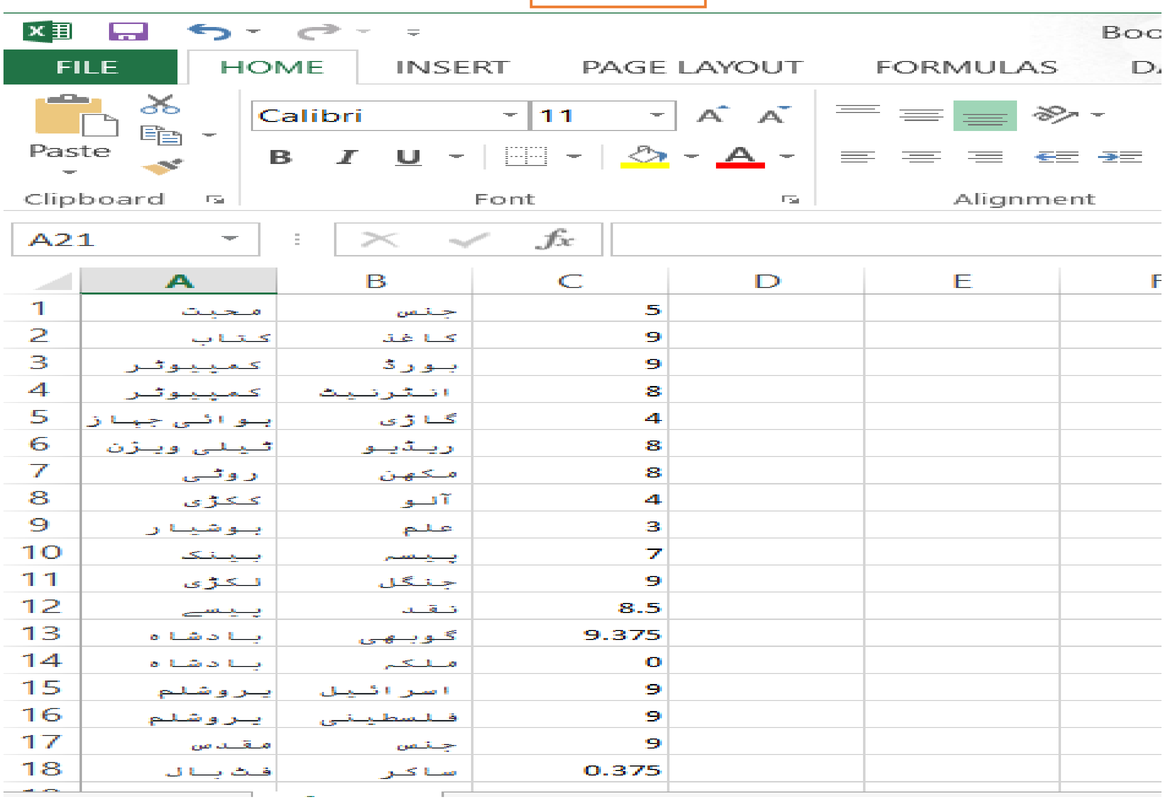
	A	B	C
1	वाघ	मांजर	7
2	वाघ	वाघ	10
3	पुस्तक	कागद	7.375
4	संगणक	कळफलक	8
5	संगणक	इंटरनेट	8.5
6	विमान	कार	6.76
7	रेल्वे	कार	6.625
8	टेलिफोन	संवाद	9
9	दूरदर्शन	रेडिओ	7
10	मीडिया	रेडिओ	8
11	भाकरी	लोणी	1
12	काकडी	काकडी	7
13	डॉक्टर	परिचारिका	6
14	प्राध्यापक	डॉक्टर	4
15	विद्यार्थी	प्राध्यापक	6
16	स्मार्ट	विद्यार्थी	4.375
17	स्मार्ट	मूर्ख	3
18	कस	अंडी	7

MARATHI



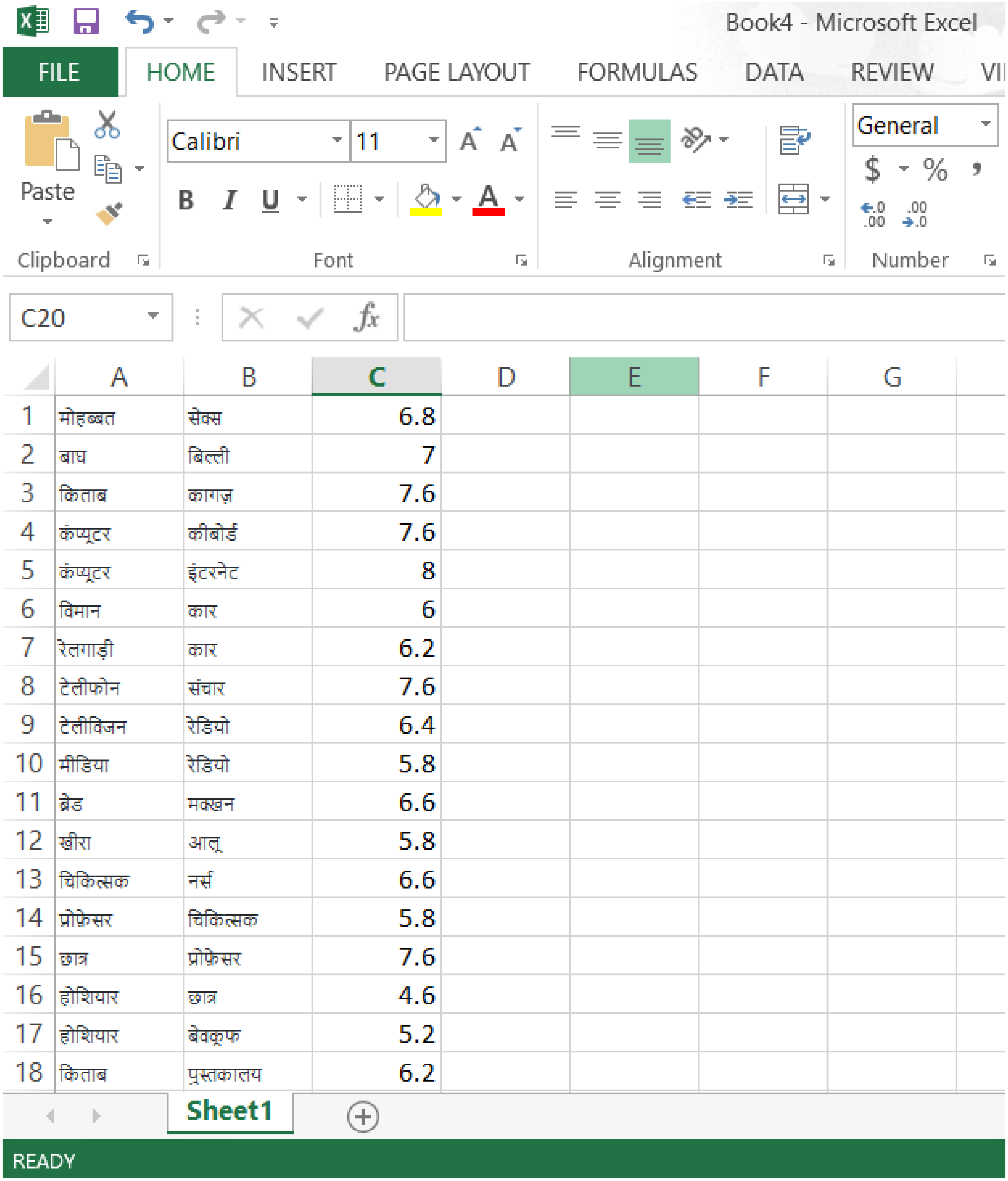
	A	B	C
1	પ્રેમ	સેક્સ	8
2	વાઘ	બિલાડી	7.5
3	વાઘ	વાઘ	10
4	પુસ્તક	કાગળ	7.25
5	કમ્પ્યુટર	કીબોર્ડ	8
6	કમ્પ્યુટર	ઇન્ટરનેટ	7.5
7	વિમાન	ગાડી	6
8	ટ્રેન	ગાડી	6
9	ટેલિફોન	સંચાર	7
10	ટેલિવિઝન	રેડિયો	8
11	મીડિયા	રેડિયો	7.375
12	બ્રેડ	માખણ	
13	કાકડી	બટાટા	6

GUJARATI



	A	B	C
1	محبت	جنس	5
2	کتاب	کاغذ	9
3	کمپیوٹر	بورڈ	9
4	کمپیوٹر	انٹرنیٹ	8
5	ہوائی جہاز	گاڑی	4
6	ٹیلی ویژن	ریڈیو	8
7	روٹی	مکھن	8
8	ککڑی	آلو	4
9	ہوشیار	علم	3
10	بینک	پیسہ	7
11	لکڑی	جنگل	9
12	پیسے	نقد	8.5
13	بادشاہ	گوبھی	9.375
14	بادشاہ	ملکہ	0
15	یروشلم	اسرائیل	9
16	یروشلم	فلسطینی	9
17	مقدس	جنس	9
18	فٹ بال	ساکر	0.375

URDU



	A	B	C
1	मोहब्बत	सेक्स	6.8
2	बाघ	बिल्ली	7
3	किताब	कागज़	7.6
4	कंप्यूटर	कीबोर्ड	7.6
5	कंप्यूटर	इंटरनेट	8
6	विमान	कार	6
7	रेलगाड़ी	कार	6.2
8	टेलीफोन	संचार	7.6
9	टेलीविजन	रेडियो	6.4
10	मीडिया	रेडियो	5.8
11	ब्रेड	मक्खन	6.6
12	खीरा	आलू	5.8
13	चिकित्सक	नर्स	6.6
14	प्रोफ़ेसर	चिकित्सक	5.8
15	छात्र	प्रोफ़ेसर	7.6
16	होशियार	छात्र	4.6
17	होशियार	बेवकूफ	5.2
18	किताब	पुस्तकालय	6.2

HINDI

7.9 NER Hindi

Brief overview of Datasets –

Named Entity Recognition (NER) Refers to automatic identification of named entities in a given text document. Given a text document, named entities such as Person names, Organization names, Location names, Product names are identified and tagged. Identification of named entities is important in several higher language technology systems such as information extraction systems, machine translation systems, and cross-lingual information access systems.

Over the past decade Indian language content on various media types such as websites, blogs, email, chats have increased significantly. Content growth is driven by people from non-metros and small cities. Need to process this huge data automatically especially companies are interested to ascertain public view on their products and processes. This requires natural language processing software systems which identify entities, identification of associations or relation between entities. Hence an automatic Named Entity recognizer is required.

The objectives of this evaluation exercise are:

Creation of benchmark data for Evaluation of Named Entity Recognition for Indian Languages

Encourage researchers to develop Named Entity Recognition (NER) systems for Indian languages.

Challenges in Indian Language NER

Indian languages belong to several language families, the major ones being the Indo-European languages, Indo-Aryan and the Dravidian languages.

The challenges in NER arise due to several factors. Some of the main factors are listed below

Morphologically rich - identification of root is difficult, require use of morphological analyzers

No Capitalization feature - In English, capitalization is one of the main features, whereas that is not there in Indian languages

Ambiguity - ambiguity between common and proper nouns. E.g.: common words such as "Roja" meaning Rose flower is a name of a person

Spell variations - In the web data is that we find different people spell the same entity differently - for example: In Tamil person name -Roja is spelt as "rosa", "roja".

Datasets Size - Size of training, testing and dev sets:

Size of training set-34586

Size of testing set-22066

Size of dev set-19571

Results

Language	Team SystemID	Precision	Recall	F-Measure
Bengali	ISI Kolkata Sys 1	23.69	28.02	25.68
	ISI Kolkata Sys 2	28.61	16.09	20.59
English	TRDDC Sys 1	64.79	67.23	65.99
	TRDCC Sys 2	64.92	68.63	66.73
	ISM Sys 1	14.89	32.02	20.33
	ISM Sys 2	39.33	34.46	36.74
Hindi	TRDCC	47.51	68.35	56.06
	IITB	83.68	74.14	78.62
	MNIT	01.72	04.82	02.53

7.10 Text Classification

Brief overview of Datasets –

The AI4Bharat-IndicNLP dataset is an ongoing effort to create a collection of large-scale, general-domain corpora for Indian languages. Currently, it contains 2.7 billion words for 10 Indian languages from two language families. We share pre-trained word embeddings trained on these corpora. We create news article category classification datasets for 9 languages to evaluate the embeddings. We evaluate the IndicNLP embeddings on multiple evaluation tasks.

We can read details regarding the corpus and other resources. We showcased the AI4Bharat-IndicNLP dataset at REPL4NLP 2020 (collocated with ACL 2020) *(non-archival submission as extended abstract)*.

We can use the IndicNLP corpus and embeddings for multiple Indian language tasks. A comprehensive list of Indian language NLP resources can be found in the IndicNLP Catalog. For processing the Indian language text, you can use the Indic NLP Library.

Note:

The vocabulary frequency files contain the frequency of all unique tokens in the corpus. Each line contains one word along with frequency delimited by tab.

For convenience, the corpus is already tokenized using the IndicNLP tokenizer. You can use the IndicNLP detokenizer in case you want a detokenized version.

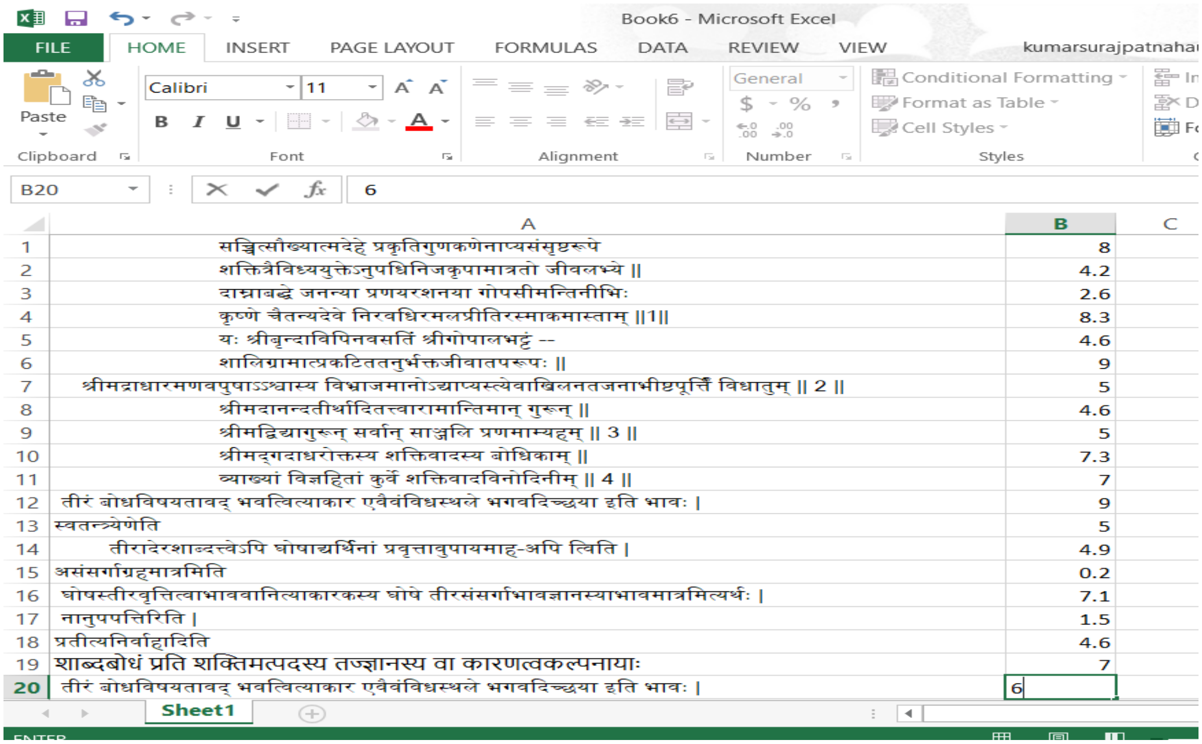
Datasets Size - Size of training, testing and dev sets:

Size of training set-8549

Size of testing set-5267

Size of dev set-2982

Sample Dataset:



	A	B
1	सच्चित्सौख्यात्मदेहे प्रकृतिगुणकणेनाप्यसंसृष्टरूपे	8
2	शक्तित्रैविध्ययुक्तेऽनुपधिनिजकृपामात्रतो जीवलभ्ये ‖	4.2
3	दाम्नाबद्धे जनन्या प्रणयरशनया गोपसीमन्तिनीभिः	2.6
4	कृष्णे चैतन्यदेवे निरवधिरमलप्रीतिरस्माकमास्ताम् ‖1‖	8.3
5	यः श्रीबृन्दाविपिनवसतिं श्रीगोपालभट्टं --	4.6
6	शालिग्रामात्प्रकटिततनुर्भक्तजीवातपरूपः ‖	9
7	श्रीमद्राधारमणवपुषाऽऽश्वास्य विभ्राजमानोऽद्याप्यस्त्येवाखिलनतजनाभीष्टपूर्त्तिं विधातुम् ‖ 2 ‖	5
8	श्रीमदानन्दतीर्थादितत्त्वारामान्तिमान् गुरून् ‖	4.6
9	श्रीमद्विद्यागुरून् सर्वान् साञ्जलि प्रणमाम्यहम् ‖ 3 ‖	5
10	श्रीमद्गदाधरोक्तस्य शक्तिवादस्य बोधिकाम् ‖	7.3
11	व्याख्यां विज्ञहितां कुर्वे शक्तिवादविनोदिनीम् ‖ 4 ‖	7
12	तीरं बोधविषयतावद् भवत्वित्याकार एवैवंविधस्थले भगवदिच्छया इति भावः ।	9
13	स्वतन्त्र्येणेति	5
14	तीरादेरशाब्दत्वेऽपि घोषाद्यर्थिनां प्रवृत्तावुपायमाह-अपि त्विति ।	4.9
15	असंसर्गाग्रहमात्रमिति	0.2
16	घोषस्तीरवृत्तित्वाभाववानित्याकारकस्य घोषे तीरसंसर्गाभावज्ञानस्याभावमात्रमित्यर्थः ।	7.1
17	नानुपपत्तिरिति ।	1.5
18	प्रतीत्यनिर्वाहादिति	4.6
19	शाब्दबोधं प्रति शक्तिमत्पदस्य तज्ज्ञानस्य वा कारणत्वकल्पनायाः	7
20	तीरं बोधविषयतावद् भवत्वित्याकार एवैवंविधस्थले भगवदिच्छया इति भावः ।	6

SANSKRIT

8 Conclusion

In a nutshell, this internship has been an excellent and rewarding experience. We can conclude that there have been a lot We have learnt from our work at YScholar. We have completed our tasks in stipulated time interval.

Two main things that we have learned the importance of time-management skills and self-motivation.

During the internship, we worked with YScholar Technology LLP on the project to develop a Natural Language Processing (NLP) Data Analysis for low resource Indo-European languages using Python for the organization.

During the internship, we had a 14 days sprint cycle which included the developmental and testing phases of the tasks assigned to us. We had bi-weekly standup meetings discussing the progress of the tasks and doubts. This internship also helped us to get familiar with the GitHub Projects' Kanban-style board for managing our tasks.

This has been a great learning for us during the 8th Semester Internship project opportunity in our B.Tech curriculum.

I would like to thank the Indian Institute of Information Technology Bhagalpur and YScholar Technology LLP for giving me this opportunity for understanding the recent industry styles and protocols that are followed in the current timeline.

9 Bibliography

1. https://www.kaggle.com/agrawaladitya/step-by-step-data-preprocessing-eda
2. https://www.nltk.org/
3. https://spacy.io/usage/models
4. https://www.coquery.org/
5. https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/
6. https://www.geeksforgeeks.org/python-efficient-text-data-cleaning/
7. https://www.kdnuggets.com/2018/03/text-data-preprocessing-walkthrough-python.html
8. https://medium.com/towards-artificial-intelligence/natural-language-processing-nlp-with-python-tutorial-for-beginners-1f54e610a1a0#3be2
9. https://realpython.com/natural-language-processing-spacy-python/
10. https://github.com/syedsarfarazakhtar/Word-Similarity-Datasets-for-Indian-Languages
11. https://github.com/kmi-linguistics/vardial2018
12. https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-625F-0
13. http://www.iitp.ac.in/~ai-nlp-ml/resources.html
14. http://amitavadas.com/sentiwordnet.php
15. https://www.kdnuggets.com/2018/03/text-data-preprocessing-walkthrough-python.html