Bangla Article Classification With TensorFlow

Group TSIA:
Zhang Xinyue
Fu Chennan
Wu Di

2020/12/5

Content

Introduction

Set up & FastText

- Data preprocessing
- Data set
- FasText

Model Training & Evaluation

- Data set converting
- Model training
- Model evaluation

Prediction & Conclusion

- F1-score, Recall, Precision

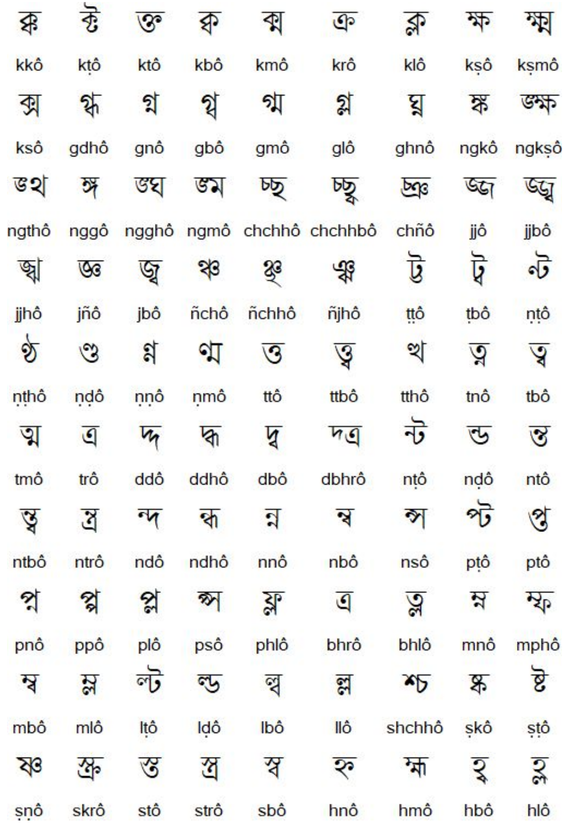
Introduction

01 | Bengali data set is small(in many works)

02 | For SL, hard to train better

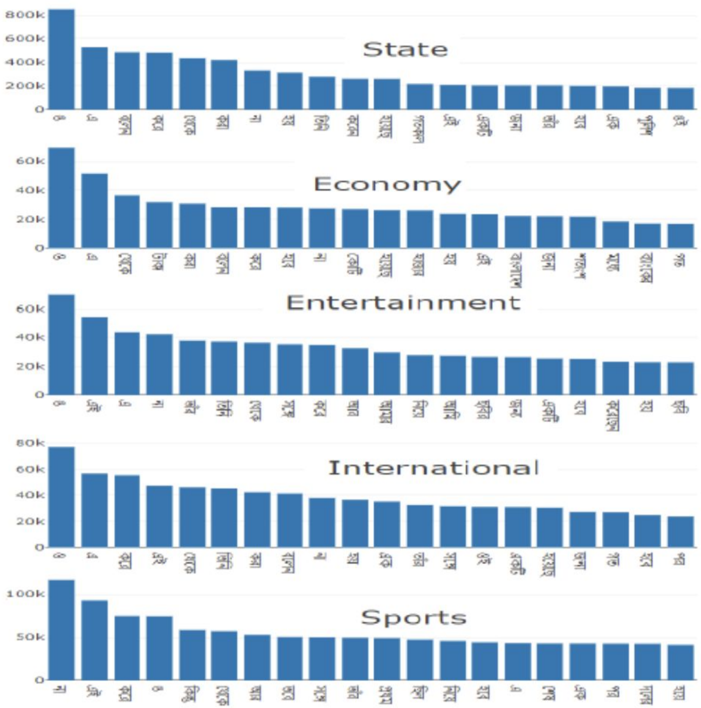
03 | Kaggle’s outcome is 70% using RNN

04 | Did not apply word embedding such FASTtext



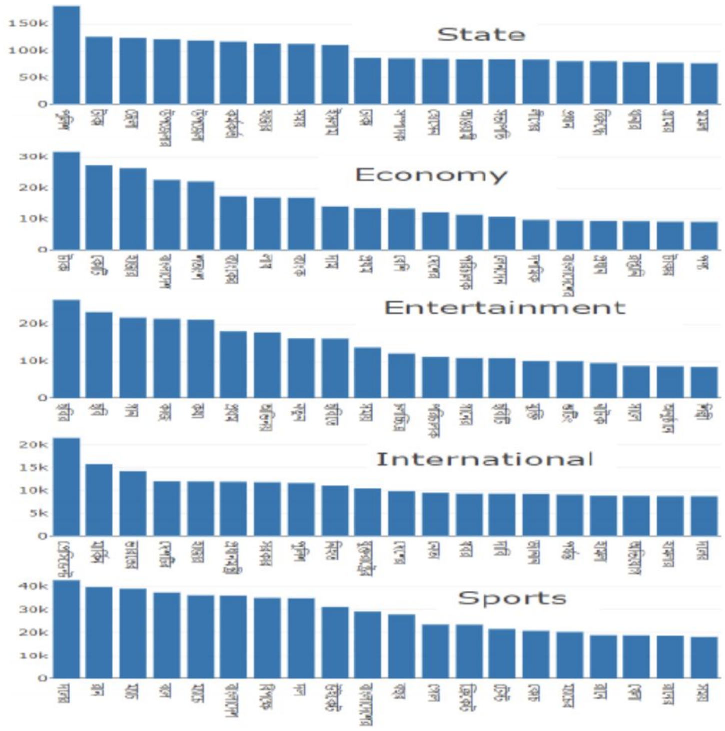
ক্ক kkô	ক্ট kṭô	ক্ত ktô	ক্ব kbô	ক্ম kmô	ক্র krô	ক্ল klô	ক্ষ kṣô	ক্ষ্ম kṣmô
ক্স ksô	গ্ধ gdhô	গ্ন gnô	গ্ব gbô	গ্ম gmô	গ্ল glô	ঘ্ন ghnô	ঙ্ক ngkô	ঙ্ক্ষ ngkṣô
ঙ্থ ngthô	ঙ্গ nggô	ঙ্ঘ ngghô	ঙ্ম ngmô	চ্ছ chchhô	চ্ছ্ব chchhbô	চ্ঞ chñô	জ্জ jjô	জ্জ্ব jjbô
জ্ঝ jjhô	জ্ঞ jñô	জ্ব jbô	ঞ্চ ñchô	ঞ্ছ ñchhô	ঞ্ঝ ñjhô	ট্ট ṭṭô	ট্ব ṭbô	ণ্ট ṇṭô
ণ্ঠ ṇṭhô	ণ্ড ṇḍô	ণ্ণ ṇṇô	ণ্ম ṇmô	ত্ত ttô	ত্ত্ব ttbô	ত্থ tthô	ত্ন tnô	ত্ব tbô
ত্ম tmô	ত্র trô	দ্দ ddô	দ্ধ ddhô	দ্ব dbô	দ্ভ্র dbhrô	ন্ট nṭô	ন্ড nḍô	ন্ত ntô
ন্ত্ব ntbô	ন্ত্র ntrô	ন্দ ndô	ন্ধ ndhô	ন্ন nnô	ন্ব nbô	ন্স nsô	প্ট pṭô	প্ত ptô
প্ন pnô	প্প ppô	প্ল plô	প্স psô	ফ্ল phlô	ভ্র bhrô	ভ্ল bhlô	ম্ন mnô	ম্ফ mphô
ম্ব mbô	ম্ল mlô	ল্ট lṭô	ল্ড lḍô	ল্ব lbô	ল্ল llô	শ্চ shchhô	ষ্ক ṣkô	ষ্ট ṣṭô
ষ্ণ ṣṇô	স্ক্র skrô	স্ত stô	স্ত্র strô	স্ব sbô	হ্ন hnô	হ্ম hmô	হ্ব hbô	হ্ল hlô

Data preprocessing



Remove

stop words



Data preprocessing

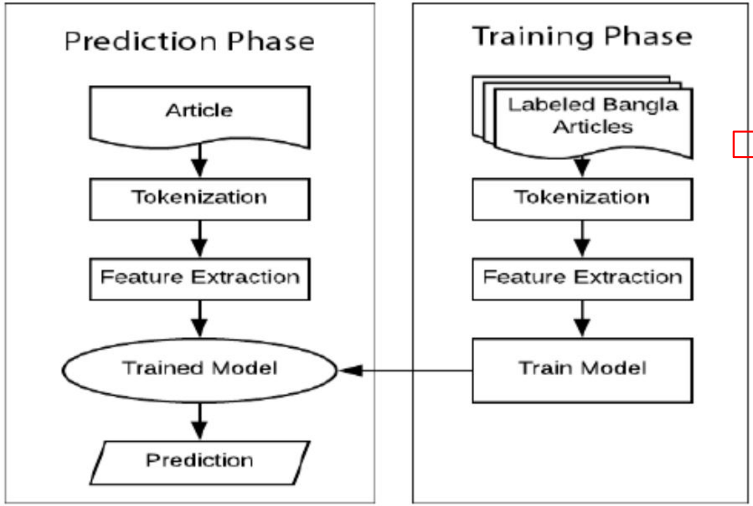


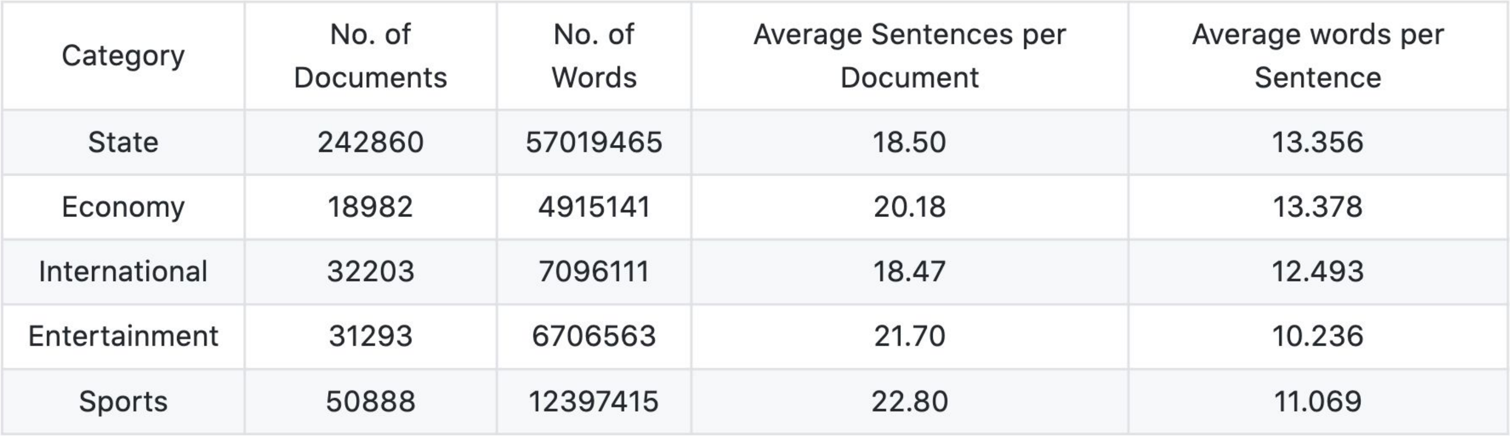
Fig. Structure of Bangala classification model

Removing stop words

```
# Returns words from a text
def get_vector(text):
    ret = ""
    stp=["!", "@",'–', "#", "|", "%", "(", ")", "।", "—", ".",
"-", "", ",", "’", "•", "‘", ":", "*", "?",
            "০", "১", "২", "৩", "৪", "৫", "৬", "৭", "৮", "৯"]
    for x in text:
        if x in stp:
            ret = ret + " "
        else:
            ret = ret + x
    ret = ret.replace("  ", " ")
    ret = ret.replace("  ", " ")
    ret = ret.split()
    return ret
```

Code. Tokenizing

Data set



Category	No. of Documents	No. of Words	Average Sentences per Document	Average words per Sentence
State	242860	57019465	18.50	13.356
Economy	18982	4915141	20.18	13.378
International	32203	7096111	18.47	12.493
Entertainment	31293	6706563	21.70	10.236
Sports	50888	12397415	22.80	11.069

Table. Data set details

376K

articles

5

categories

FastText

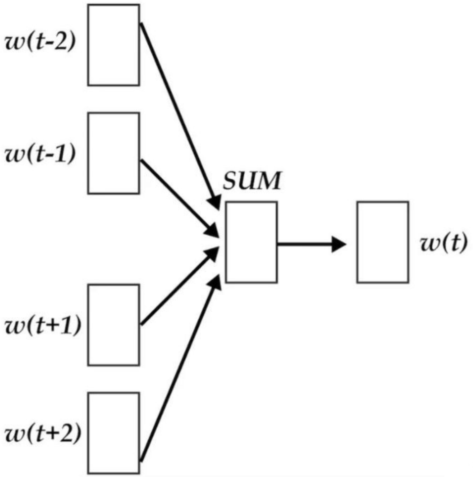


Fig. Key idea of fastText

Example

love

{ <l , lo ,ov ,ve ,e>, love }

- ❑ For low-frequency words: their n-grams can be shared with other words.
- ❑ For words outside the trained vocabulary: their n-gram vectors are superimposed.

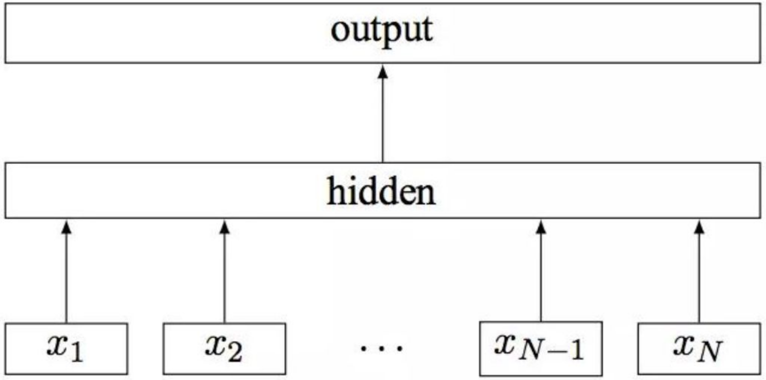
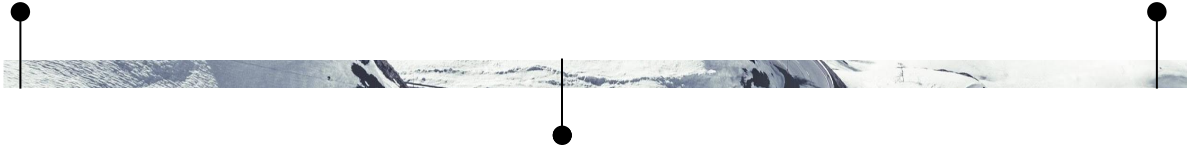


Fig. Structure of fastText

Data set converting



Unbalanced data set

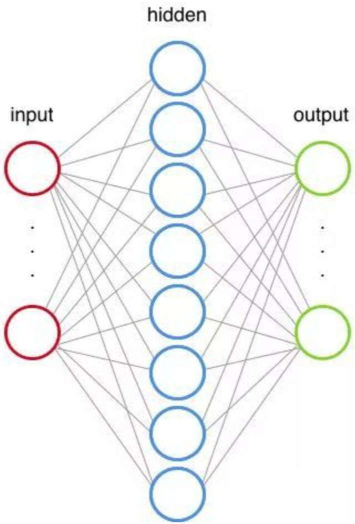
Generate the sample data into batches

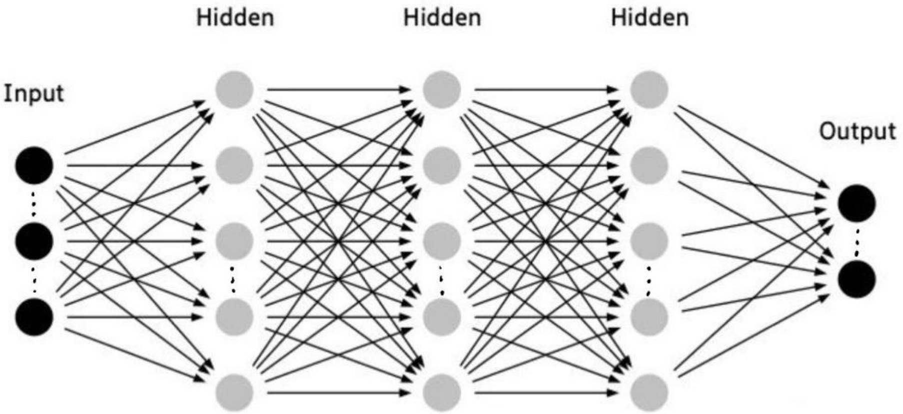
Disturb the order

```
def load_file(path, label):
   return tf.io.read_file(path), label
def make_datasets(train_size):
 ...
 train_ds = tf.data.Dataset.from_tensor_slices((train_files, train_labels))
 train_ds = train_ds.map(load_file).shuffle(5000)
 train_ds = train_ds.batch(batch_size).prefetch(tf.data.experimental.AUTOTUNE)
 ...
 test_ds = tf.data.Dataset.from_tensor_slices((test_files, test_labels))
 test_ds = test_ds.map(load_file)
 test_ds = test_ds.batch(batch_size).prefetch(tf.data.experimental.AUTOTUNE)
 ...
```

Code for input data processing

Model Training





The main function of the hidden layer is to convert the data of the input layer into a more convenient form of the output layer.

01 | Create a sequential model through Keras.

02 | Add a layer instance at the top of the layer stack.

In this model, one input layer, one embedding layer,two fully-connected layers and one output layer are included.

Model Evaluation

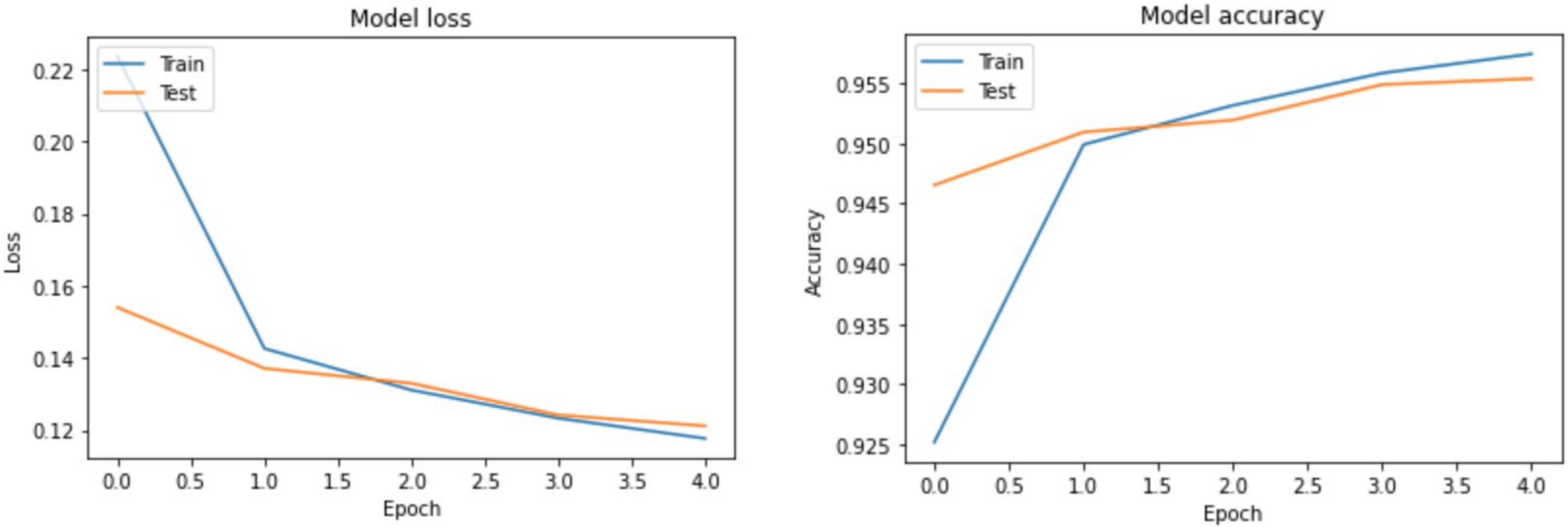
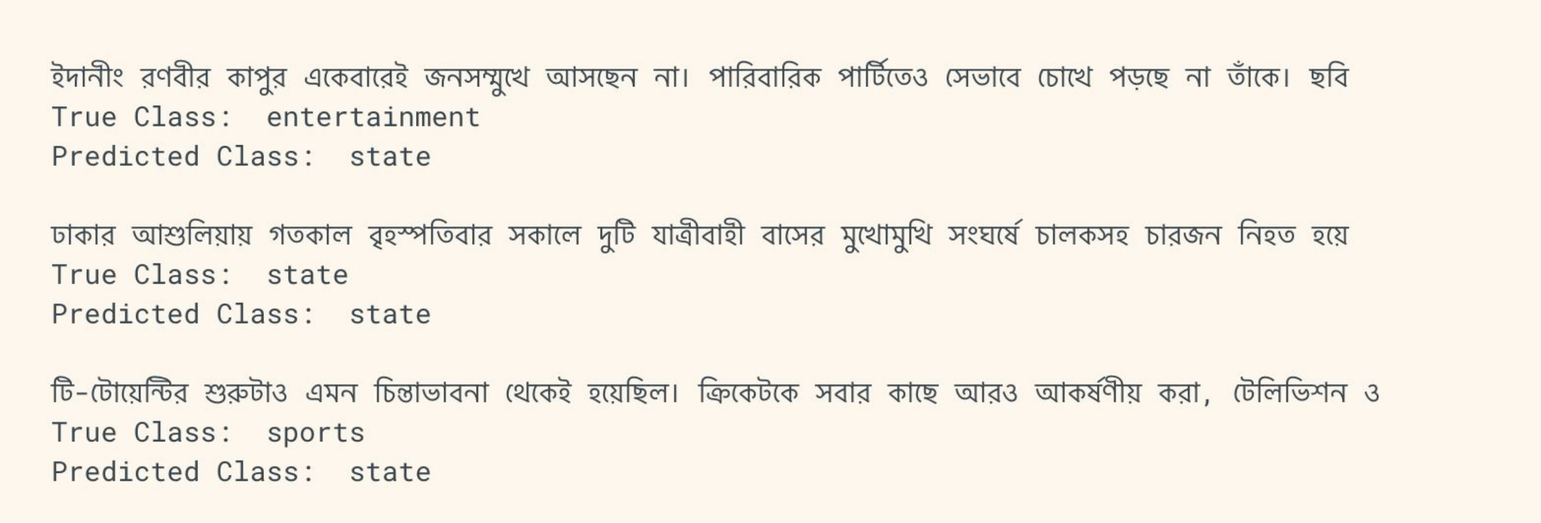


Fig. Visualize loss curve and accuracy of training and validation data.

Data Prediction

The first line: randomly print the first hundred characters of the file
The second line: the ground truth category
The third line: the predicted category



```
ইদানীং রণবীর কাপুর একেবারেই জনসম্মুখে আসছেন না। পারিবারিক পার্টিতেও সেভাবে চোখে পড়ছে না তাঁকে। ছবি
True Class:  entertainment
Predicted Class:  state

ঢাকার আশুলিয়ায় গতকাল বৃহস্পতিবার সকালে দুটি যাত্রীবাহী বাসের মুখোমুখি সংঘর্ষে চালকসহ চারজন নিহত হয়ে
True Class:  state
Predicted Class:  state

টি-টোয়েন্টির শুরুটাও এমন চিন্তাভাবনা থেকেই হয়েছিল। ক্রিকেটকে সবার কাছে আরও আকর্ষণীয় করা, টেলিভিশন ও
True Class:  sports
Predicted Class:  state
```

Data Comparison

Evaluation Indicators

Precision

$$Precision = \frac{TP}{TP + FP}$$

Recall

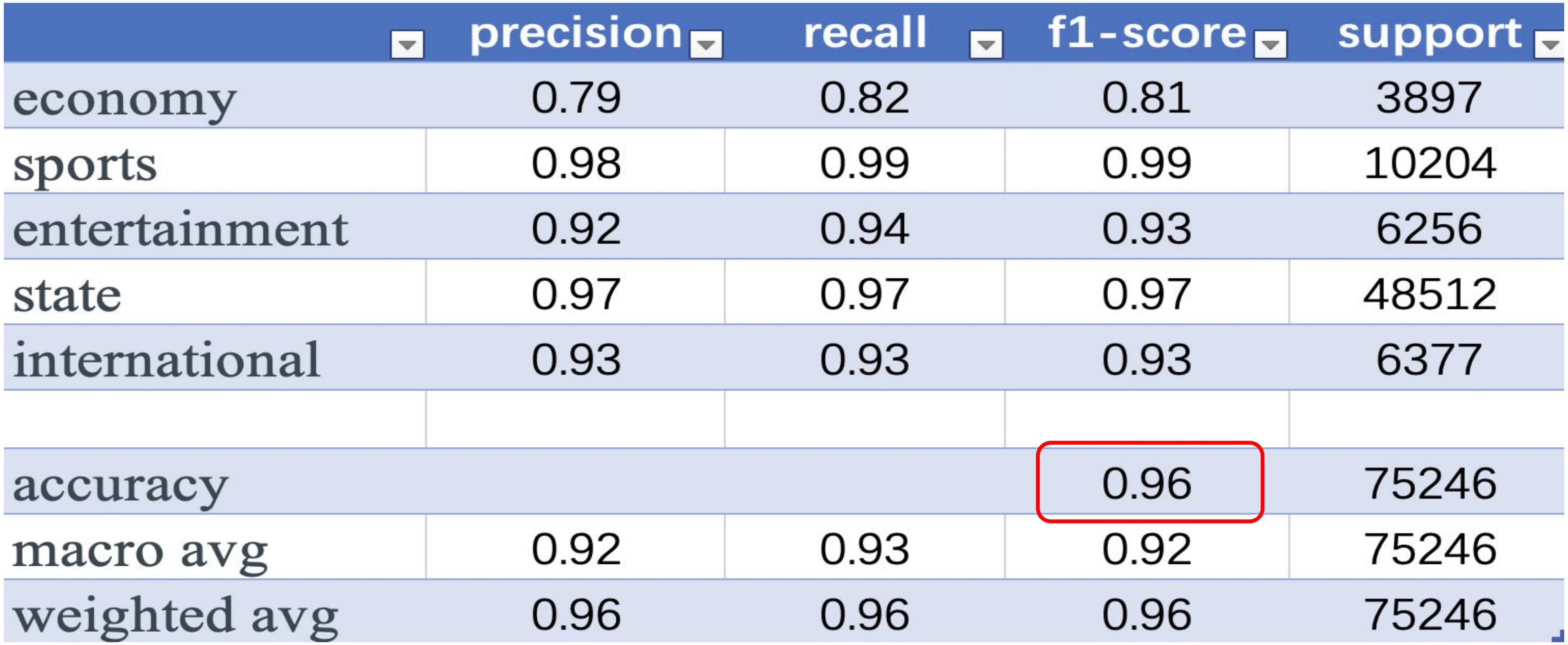
$$Recall = \frac{TP}{TP + EN}$$

F1-score

$$F1 - \text{score} = \frac{2 * (\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}}$$

◇ ***Here, TP = True Positives , FP = False Positives , FN = False Negatives.***

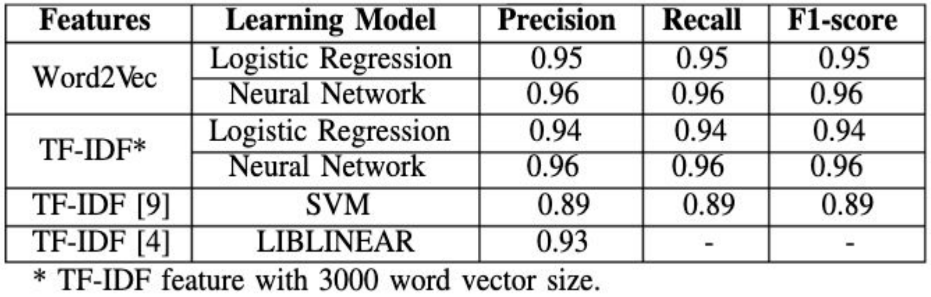
Data Comparison



	precision	recall	f1-score	support
economy	0.79	0.82	0.81	3897
sports	0.98	0.99	0.99	10204
entertainment	0.92	0.94	0.93	6256
state	0.97	0.97	0.97	48512
international	0.93	0.93	0.93	6377
accuracy			0.96	75246
macro avg	0.92	0.93	0.92	75246
weighted avg	0.96	0.96	0.96	75246

Here is Result!

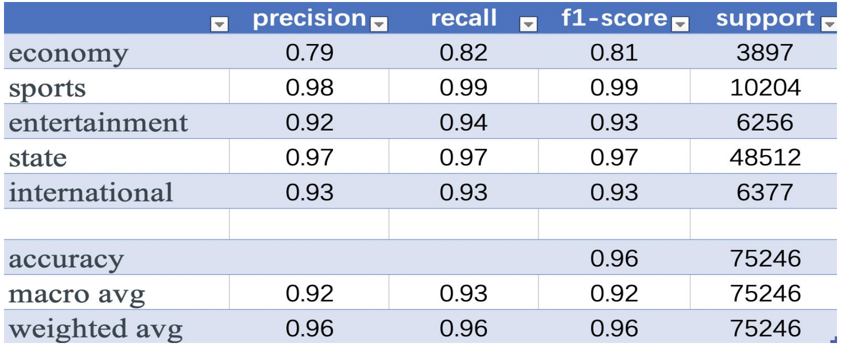
Summary & Future Work



Features	Learning Model	Precision	Recall	F1-score
Word2Vec	Logistic Regression	0.95	0.95	0.95
	Neural Network	0.96	0.96	0.96
TF-IDF*	Logistic Regression	0.94	0.94	0.94
	Neural Network	0.96	0.96	0.96
TF-IDF [9]	SVM	0.89	0.89	0.89
TF-IDF [4]	LIBLINEAR	0.93	-	-

* TF-IDF feature with 3000 word vector size.

Table1: other state-of-the-art works

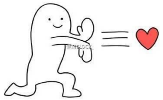


	precision	recall	f1-score	support
economy	0.79	0.82	0.81	3897
sports	0.98	0.99	0.99	10204
entertainment	0.92	0.94	0.93	6256
state	0.97	0.97	0.97	48512
international	0.93	0.93	0.93	6377
accuracy			0.96	75246
macro avg	0.92	0.93	0.92	75246
weighted avg	0.96	0.96	0.96	75246

Table2: our model performance

Pretrained Models (Neural Network)

Word Embedding Methods (Word2Vec and TF-IDF)



Simple Model

FastText

Five Epochs

👍

Reference

[1] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching wordvectors with subword information,” CoRR, vol. abs/1607.04606, 2016.

[2] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Inf. Process. Manage., vol. 24, no. 5, pp. 513–523, Aug.1988.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean,“Distributed representations of words and phrases and their composi-tionality,” in Advances in Neural Information Processing Systems 26,C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q.

[4] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” NeuralComput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[5] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Inf. Process. Manage., vol. 24, no. 5, pp. 513–523, Aug.1988.

交付

在此处插入文本在此 处插入文本在此 处插入文本在此 处插入文本在此 处插入文本在此 处插入文本在此 处插入文本在此 处插入文本在此 处插入文本在此处插入文本在此 处插入文本在此 处插入文本在此 处插入文本在此 处插入文本在此 处插入文本在此 处插入文本在此 处插入文本在此 处插入文本在此处插入文本在此 处插入文本在此 处插入文本在此 处插入文本。

在此处插入文本

45K

在此处插入文本在此 处插入文本在此处插入文本

在此处插入文本

690K

在此处插入文本在此 处插入文本在此处插入文本

在此处插入文本

100K

在此处插入文本在此 处插入文本在此处插入文本在此 处插入文本在此处插入文本