Natural Language Processing with Disaster Tweets

R26091032 戴庭筠

OUTLINE

- Motivation
- Problem statement
- Data Description
- Data Analysis (balanced data/ imbalanced data)
- Conclusion

Motivations

- ✗ People use smartphone to announce an emergency they're observing in time, so twitter is an important communication channel under emergencies.
- ✗ It's important for disaster relief organizations and news agencies to monitor and know the real situation with tweets.

Problem Statement

- ✗ Input : text in twitter
- ✗ Output：1, tweet is real disaster；0, tweet isn’t real disaster

	text
	Our Deeds are the Reason of this #earthquake M...
	Forest fire near La Ronge Sask. Canada
	All residents asked to 'shelter in place' are ...
	13,000 people receive #wildfires evacuation or...
	Just got sent this photo from Ruby #Alaska as ...

Data Description

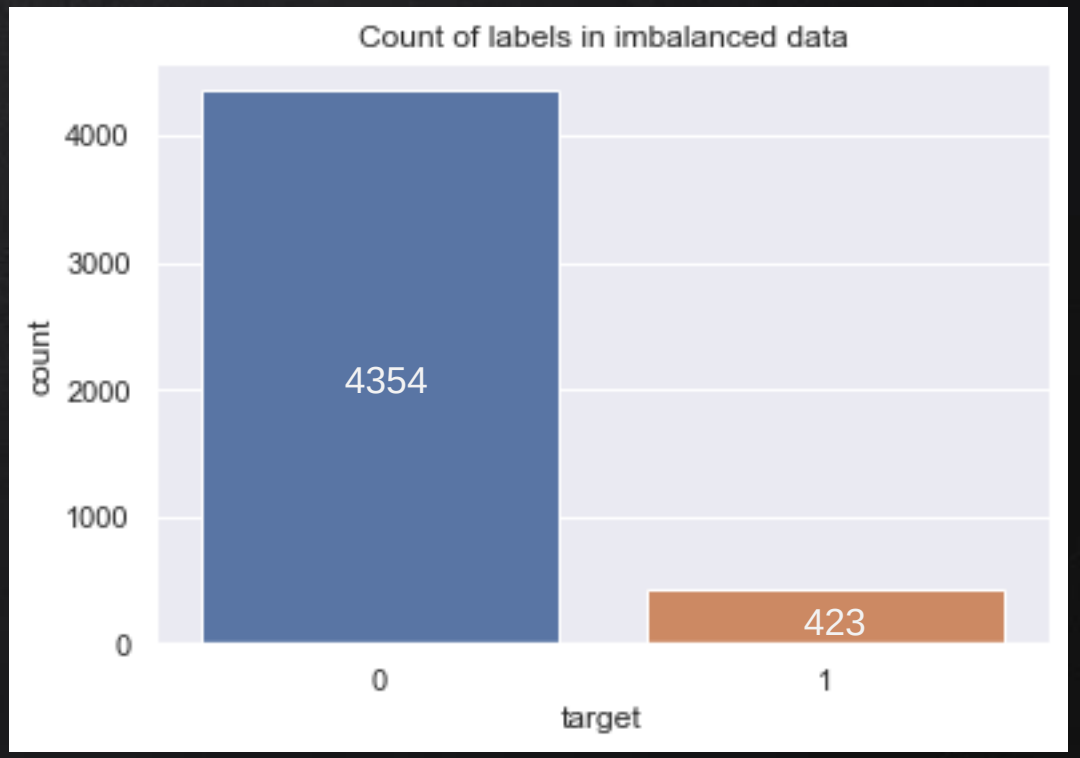
- ✗ Dataset from Kaggle ‘Natural Language Processing with Disaster Tweets’
- ✗ There are 7613 samples in training set, and 3263 samples in testing set.

	id	keyword	location	text	target
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	1
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	1
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1

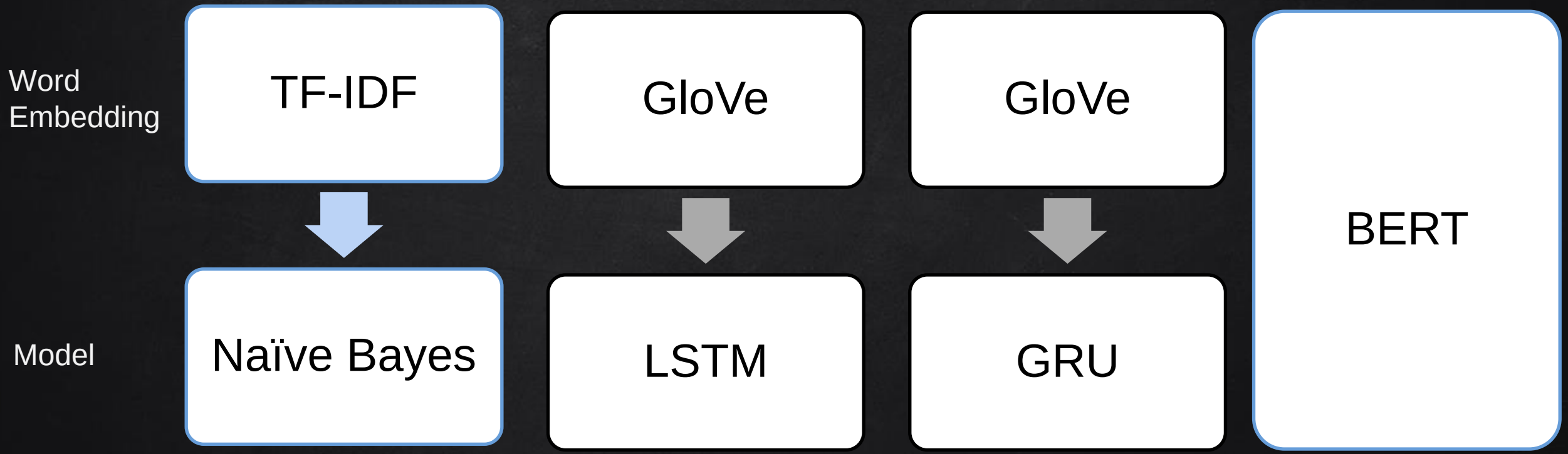
Data Description

- Balanced Data (4:3)

- Imbalanced Data (10:1)



Data Analysis – balanced data



Data Analysis – balanced data

Method	Naïve Bayes	LSTM	GRU	BERT
F-score	0.7992	0.8023	0.8063	0.8326

Choose BERT as main model in imbalanced data.

Data Analysis – imbalanced data

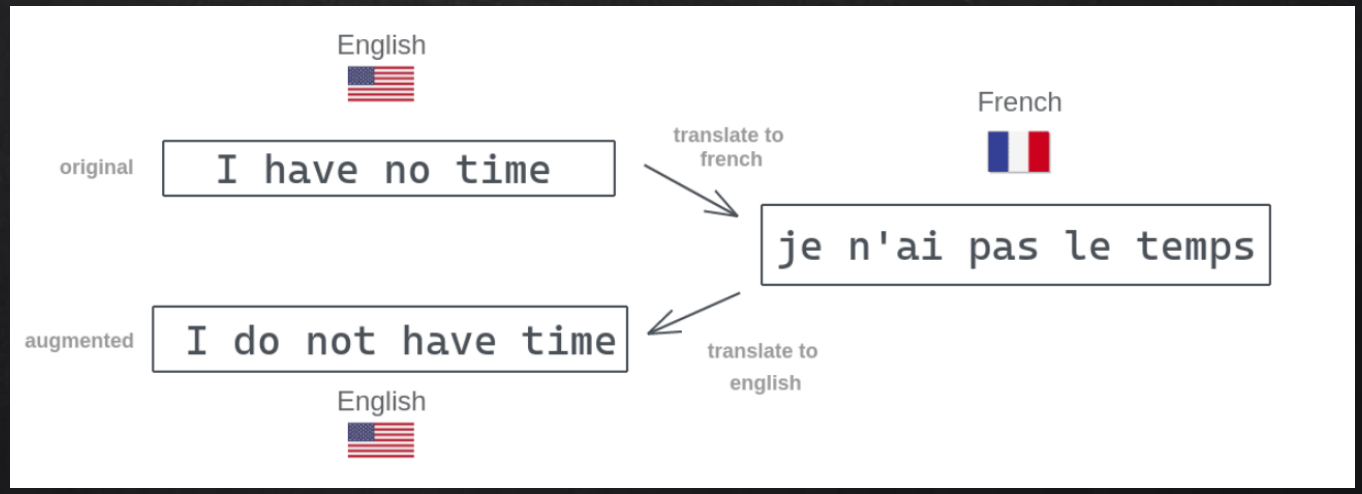
- ✗ Adjust class weight
- ✗ Data augmentation (increase sample size)
  - Back translation
  - Text generation
  - Random insertion

The number of sample size?

Combination of data augmentation?

Data Analysis – data augmentation

- ✗ Back translation
  - English ⟹ Spanish / German ⟹ English



Data Analysis – data augmentation

- ✗ Text generation

I went to see a movie in the theater.

“I went to see a new movie in the theater.” says Bobo.

- ✗ Random insertion

I went to see a movie in the theater.

I went to see a new movie in the theater.

Data Analysis – imbalanced data

Methods	Label Ratio	F-Score
Baseline	10 : 1	0.6745
Class weight	10 : 1	0.7873
Back translation * 1 + class weight	5 : 1	0.7928
Text generation * 1 + class weight	5 : 1	0.7689
Back translation * 1 + Text generation	3.4 : 1	0.7986
Back translation * 1 + Text generation + Random insertion	2.6 : 1	0.8044
Back translation * 2 + Text generation + Random insertion	2 : 1	0.7949

Conclusion

- BERT model has best performance.
- Optimal number of sample size we create is about 3 times.
- Back translation is best method in data augmentation.
- Data augmentation is better than adjust class weight.
- Balanced data F1-score = 0.8326 and imbalanced data F1-score = 0.8044

Thank you for your attention!