Building a News Recommender Using NER and Sentence Transformer

Ming Cen, Matthew Gnanadass, Qihang Wang, Zhibao Li

Introduction

- How do news sites recommend articles?

Macron arrives at the White House for first state visit of the Biden administration

By Maegan Vazquez, CNN

Updated 10:31 AM EST, Thu December 1, 2022

RELATED ARTICLE

The Bidens' first state dinner features butter-poached lobster with a side of hospitality

RELATED ARTICLE

First on CNN: Harris and Macron to strengthen working relationship with NASA headquarters visit

Introduction

- Our idea was to used **Named Entities**!

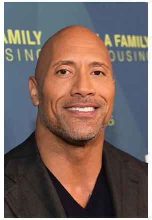
Organizations

amazon

girl scouts

People

Lionel Messi



The Rock

Locations

Washington Dc

Stonehenge

+ Miscellaneous

Idea

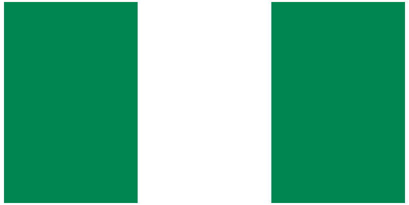
Goat accused of robbery

POLICE in Nigeria are holding a goat on suspicion of attempted armed robbery.

Vigilantes seized the black and white goat, saying it was an armed robber who had used black magic to transform himself into an animal to escape after trying to steal a Mazda 323.

A spokesman for police in the eastern state of Kwara said: 'The goat is in our custody.

'Vigilantes saw some hoodlums attempting to rob a car. One escaped while the other turned into a goat.'



METRO NEWS.

NEWS WORLD CUP SPORT ENTERTAINMENT SOAPS LIFESTYLE VIDI

LONDON UK US WORLD COST OF LIVING ROYAL FAMILY RUSS

Goat arrested for Mazda robbery

Comment

Metrowebukmetro
Friday 23 Jan 2009 4:06 pm

Share

Police in Nigeria are holding a goat on suspicion of the attempted armed robbery of a Mazda 323.

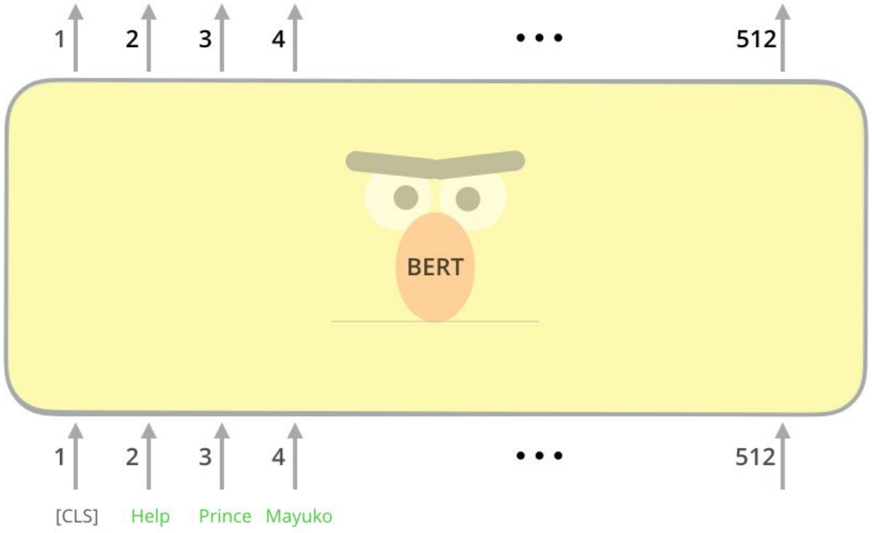
Vigilantes took the animal to the police, claiming it was a criminal who had used black magic to transform himself into a goat to escape arrest after trying to steal the car.

Technical Explanations

Overview:

- Training a BERT model on a named entity recognition dataset (CoNLL-2003 Dataset)
- Testing the model using CNN news articles
- Using the Sentence Transformer algorithm for embedding.
- Evaluating the model by computing the most similar texts

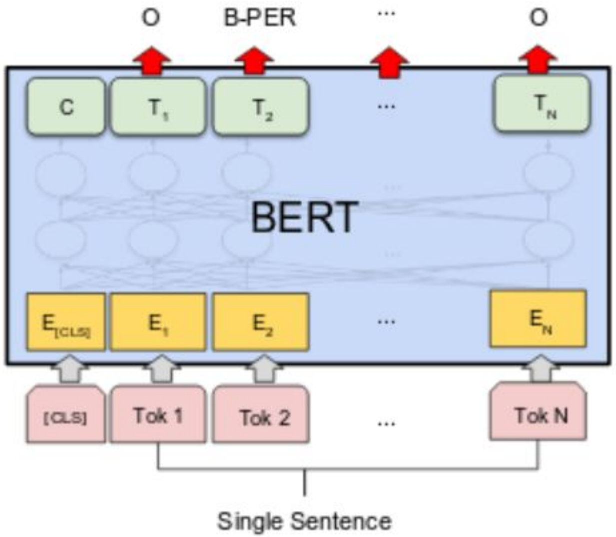
BERT model



Note, that Bert supports sequences of up to 512 tokens.

Model: “bert-base-cased”, “BertForTokenClassification”

	Word	POS	IOB tags	Tag	Sentence
0	EU	NNP	I-NP	I-ORG	0
1	rejects	VBZ	I-VP	O	0
2	German	JJ	I-NP	I-MISC	0
3	call	NN	I-NP	O	0
4	to	TO	I-VP	O	0
5	boycott	VB	I-VP	O	0
6	British	JJ	I-NP	I-MISC	0
7	lamb	NN	I-NP	O	0
8	.	.	O	O	0
10	Peter	NNP	I-NP	I-PER	1



Ideal output of the testing data

“It has been an interminable month since Elon Musk assumed control of Twitter and showed up in its headquarters while carrying a bathroom sink. “

Twitter ➜ I-ORG

Elon ➜ I-PER

Musk ➜ I-PER

[“Twitter”, “Elon”, “Musk”] ➜ [“I-ORG”, “I-PER”, “I-PER”]

Sentence Transformers for embedding:

A framework for state-of-the-art sentence, text and image embeddings.

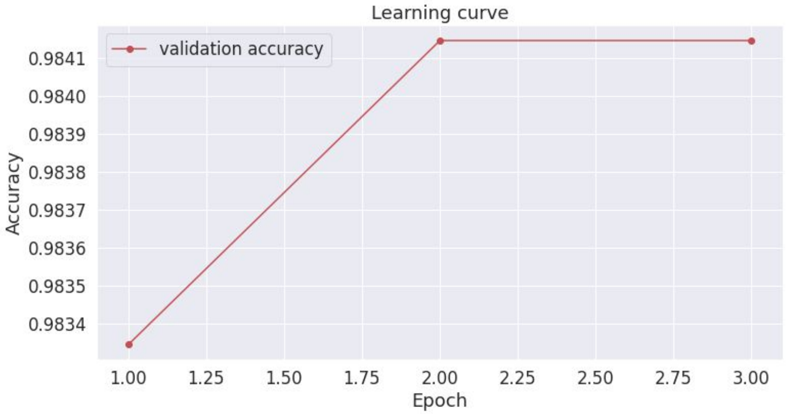
["Twitter", "Elon", "Musk"] ➜ "Twitter Elon Musk"

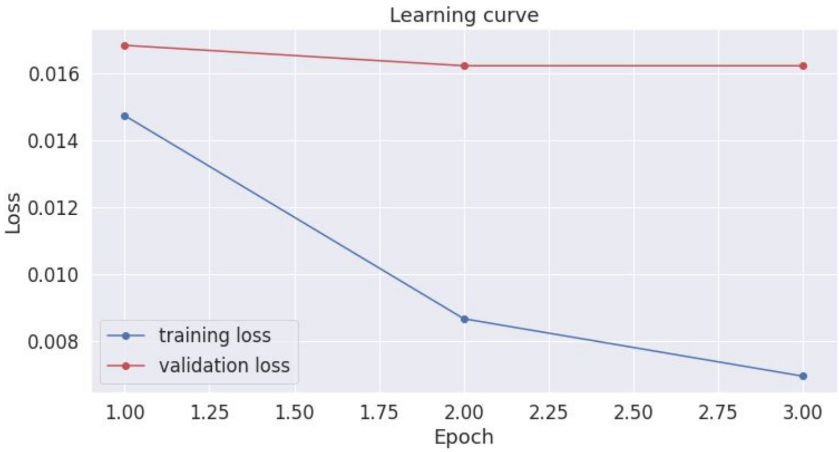
list ➜ string

"Twitter Elon Musk" ➜ [0.034875, .096636, ...]

string ➜ vector

Named-Entity Recognition Result



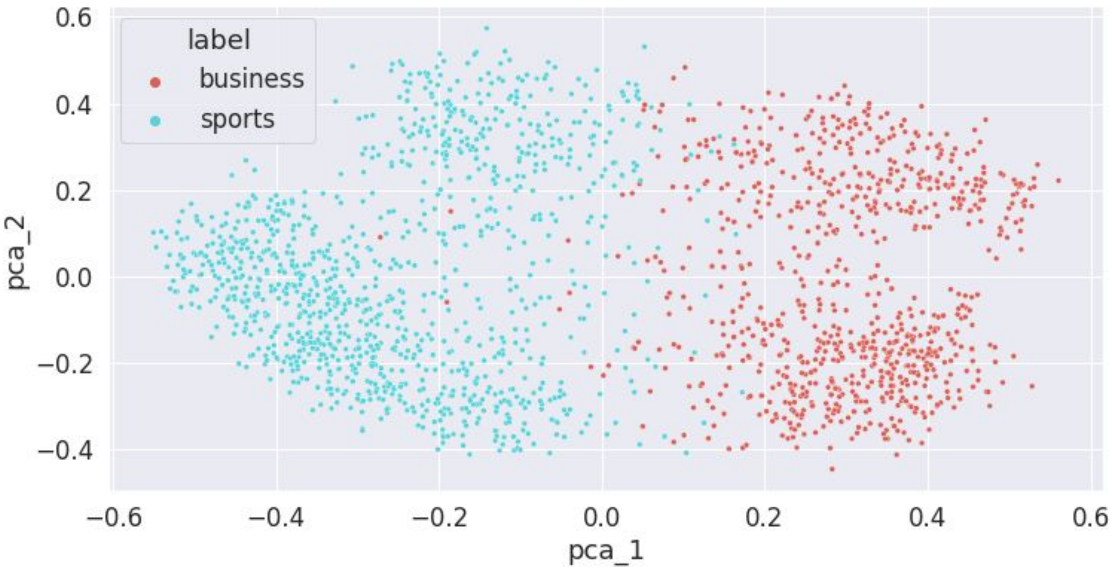


Named-Entity Recognition Result

ner	ner_labels
['KARACHI', 'Sindh', 'Geo', 'News', 'Karachi',...	['I-ORG', 'I-LOC', 'I-ORG', 'I-ORG', 'I-ORG', ...
['Hong', 'Kong', 'Wall', 'Street', 'Hang', 'Se...	['I-LOC', 'I-LOC', 'I-LOC', 'I-LOC', 'I-MISC',...
['New', 'York', 'Saudi', 'Arabia', 'US', 'West...	['I-LOC', 'I-LOC', 'I-LOC', 'I-LOC', 'I-LOC', ...
['KARACHI', 'KSE', '-', '100', 'Index', 'Karac...	['I-ORG', 'I-MISC', 'I-MISC', 'I-MISC', 'I-MIS...
['Singapore', 'Asia', 'US', 'West', 'Texas', '...	['I-LOC', 'I-LOC', 'I-LOC', 'I-ORG', 'I-ORG', ...
['KARACHI', 'Sindh', 'Karachi', 'Sindh']	['I-ORG', 'I-LOC', 'I-LOC', 'I-LOC']
['TOKYO', 'Tokyo', 'Wall', 'Street', 'Nikkei',...	['I-ORG', 'I-LOC', 'I-LOC', 'I-LOC', 'I-MISC',...
['Hong', 'Kong', 'Wall', 'Street', 'Hang', 'Se...	['I-LOC', 'I-LOC', 'I-LOC', 'I-LOC', 'I-MISC',...
['Federal', 'Petroleum', 'Shahid', 'Khaqan', '...	['I-ORG', 'I-ORG', 'I-PER', 'I-PER', 'I-PER', ...
['ISLAMABAD', 'OGRA', 'Oil', 'and', 'Gas', 'Re...	['I-ORG', 'I-ORG', 'I-ORG', 'I-ORG', 'I-ORG', ...

Pre-Trained-Embedding Model Result

Dimension of the embedding space: 384



Recommendation results

```
recommendation(1000)
```

```
# Input Healine:  India opt to bowl against Bangladesh in rain reduced Asia Cup fi
# Input Label:  sports
# Top 10 similar articles
```

	headline	text	label	cosine_similarity
941	Bangladesh put India into bat in Asia Cup opener	DHAKA: Bangladesh captain Mashrafe Mortaza won...	sports	0.912977
958	India win toss bowl against Paki	DHAKA: India captain Mahindra Singh Dhoni won ...	sports	0.883635
939	Rohits 83 lifts India to 166 6 against Banglad	DHAKA: Rohit Sharma s 55-ball 83 and late surg...	sports	0.874810
989	Pakistan win toss bowl against Sri L	DHAKA: Pakistan captain Shahid Khan Afridi won...	sports	0.873222
982	Pakistan face Bangladesh in must win game today	strong>DHAKA: Pakistan will bank a lot on thei...	sports	0.871331
980	Pakistan win toss bat against Bangladesh in do...	DHAKA: Pakistan captain Shahid Khan Afridi won...	sports	0.867156
985	Asia Cup UAE win toss decides to bat fir	DHAKA: United Arab Emirate (UAE) won the toss ...	sports	0.849027
972	Asia Cup India opt to bowl against Sri L	strong>DHAKA: India won the toss and elected t...	sports	0.842350
998	India overpower Tigers to become Asian	DHAKA: India defeated Bangladesh by eight wick...	sports	0.830616
984	India defeat UAE by 9 wi	strong>DHAKA: India enjoyed an easy victory ah...	sports	0.830314

Discussion and Conclusion

Major Concerns:

- Long training time
- The CNN articles used for testing might not align with the initial BERT pre-trained model

Discussion and Conclusion

Major Concerns:

- Evaluating the model with labels (Business, Sports, Tech...) is controversial
- For example Elon Musk might appear on a tech article as well as a business article, the articles are similar to users, but the labels are different

Discussion and Conclusion

Real-life application:

- News article recommendations
- Advertisement and Marketing
- User preference analysis
- Police Report Identification
- Stackoverflow Code Classification

Discussion and Conclusion

Final Comments:

- The model perform well on classifying the name entity and then grouping similar articles
- Will benefit many real life applications such as publications and social networks.

Hope you like it, thanks!