CS579
Fake News Classification

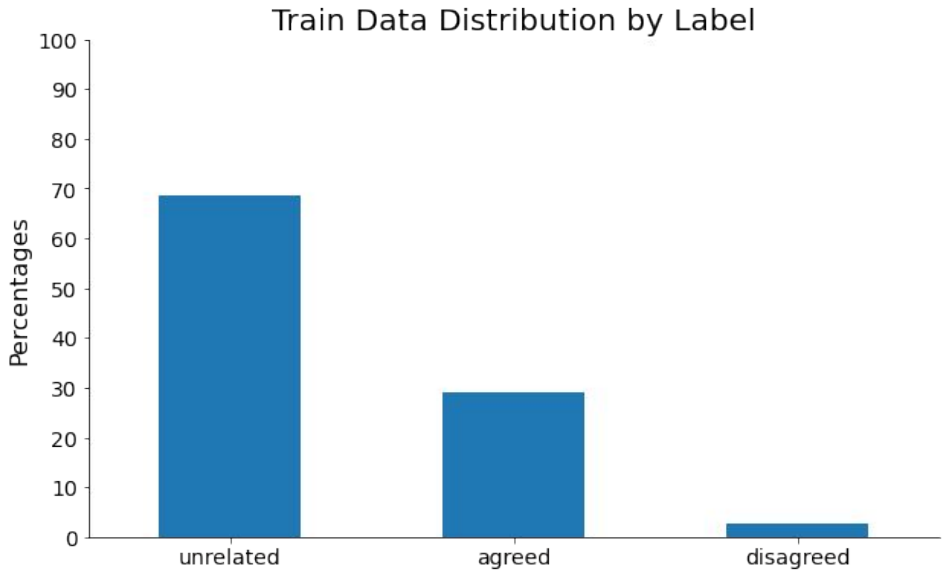
Instructor: Dr. Kai Shu
Student: Christopher Hong

Project Scope

- Title 1
  - “A few characteristics, let a pregnant mother know the difference between having a boy and having a baby girl!”
- Title 2
  - “To share the characteristics of a girl, to have a baby, to have a baby, to have a successful pregnancy, and to follow.”
- Title 1 && Title 2 => unrelated, agreed or disagreed.

Project Objective

- Build a **multinomial classification model** that helps solve the project scope
- Minimum acceptable level of performance, f1 score, should be **higher** than underlying label distribution



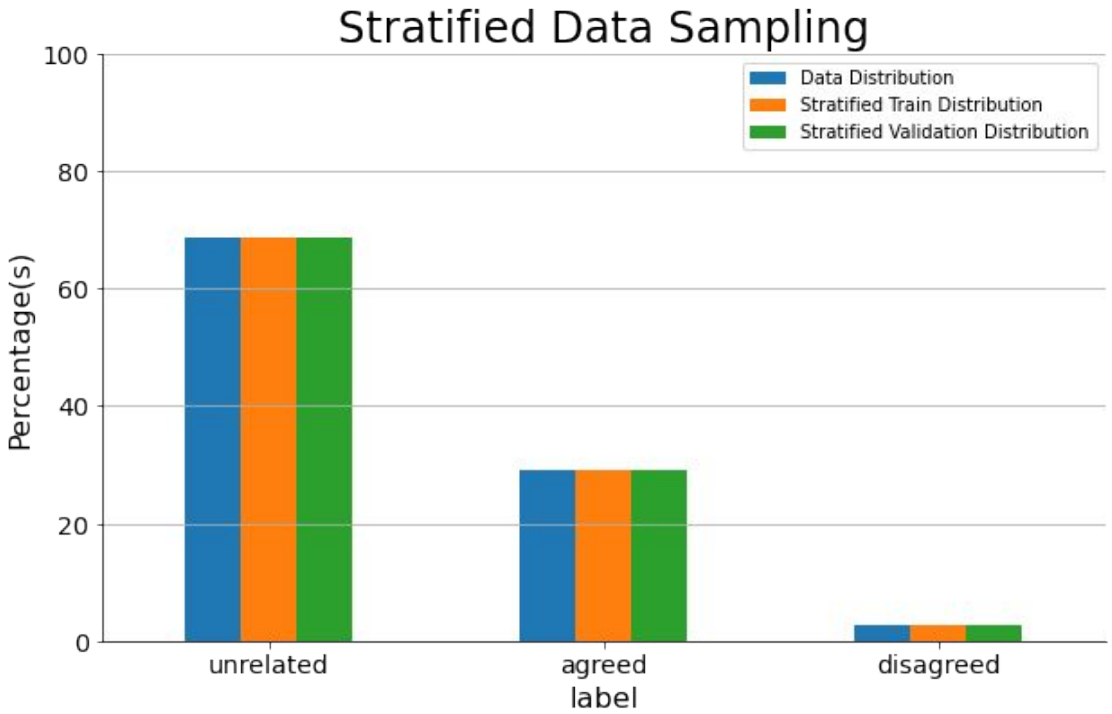
Data Collection

- Train data
  - 256,442 entries
  - 6 attributes

	id	tid1	tid2	title1_en	title2_en	label
0	195611	0	1	There are two new old-age insurance benefits f...	Police disprove "bird's nest congress each per...	unrelated
1	191474	2	3	"If you do not come to Shenzhen, sooner or lat...	Shenzhen's GDP outstrips Hong Kong? Shenzhen S...	unrelated
2	25300	2	4	"If you do not come to Shenzhen, sooner or lat...	The GDP overtopped Hong Kong? Shenzhen clarifi...	unrelated
3	123757	2	8	"If you do not come to Shenzhen, sooner or lat...	Shenzhen's GDP overtakes Hong Kong? Bureau of ...	unrelated
4	141761	2	11	"If you do not come to Shenzhen, sooner or lat...	Shenzhen's GDP outpaces Hong Kong? Defending R...	unrelated

- Test data
  - 64,100 entries
  - 5 attributes (excluded label)

Data Partitioning



Data Wrangling - **Diagnosis**

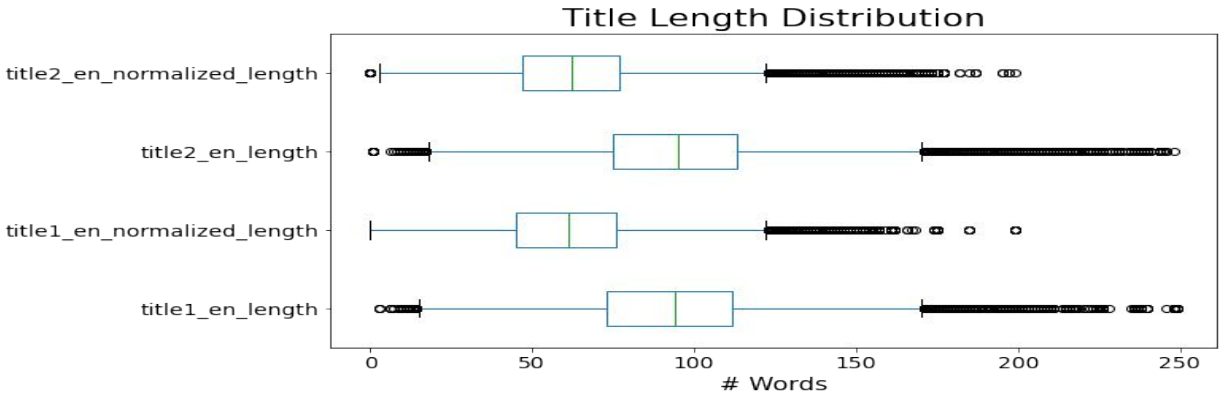
- Content diagnosis
  - Complete (no missing records)
  - Valid (no constraint violations)
  - Accurate (no incorrect records)
  - Consistent (no inconsistent formats except “title1_en” & “title2_en”)
- Structural diagnosis
  - Each variable forms a column (except “title1_en” & “title2_en”)
  - Each observation forms a row
  - Each type of observational unit forms a table

Data wrangling - **Cleaning**

- Text normalization
  - Lowercasing
  - Lemmatization
  - Stop word, punctuation, unwanted tokens (length less than 3) removal

“A few characteristics, let a pregnant mother know the difference between having a boy and having a baby girl!”

“characteristic let pregnant mother know difference boy baby girl”



Feature engineering - Count bag-of-one-gram

mechat	medical	medicine	medium	medlar	meet	meeting	mellitus	melon	member	memory	men
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0
0	0	4	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0

Hypothesis: word counts => label association

Feature engineering - Binary bag-of-one-gram

zero	zhang	zhangbaiji	zhao	zhaowei	zhejiang	zheng	zhengshuang	zhenning	zhi	zhong	zhou	zhu	zhuo	ziyi
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Hypothesis: word presence => label association

Feature engineering - TF-IDF-of-one-gram

characteristic	charge	charged	charging	chat	cheap	cheat	cheated	cheating	check	chen
0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0
0.0	0.0	0.0	0.0	0.245274	0.0	0.0	0.0	0.0	0.243246	0.0
0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0
0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0
0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0

Hypothesis: word importance => label association

Feature engineering - Order & one-hot encoding

- Shallow learning
  - Order-encoding labels
    - unrelated -> 0
    - agreed -> 1
    - disagreed -> 2

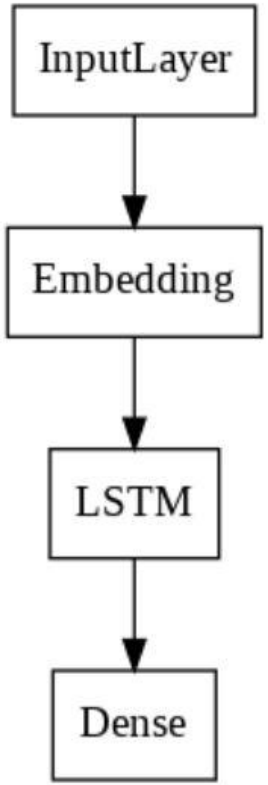
- Deep learning
  - One-hot encoding labels

```
0    [[1. 0. 0.]
1     [0. 1. 0.]
0     [1. 0. 0.]
0     [1. 0. 0.]
0     [1. 0. 0.]]
```

- Deep learning
  - Order-encoding features

characteristic	let	pregnant	mother	know	difference	boy	baby	girl	share
1209	104	47	110	30	1283	84	55	68	347

Feature engineering - Word embeddings



Feature engineering - Cosine similarity

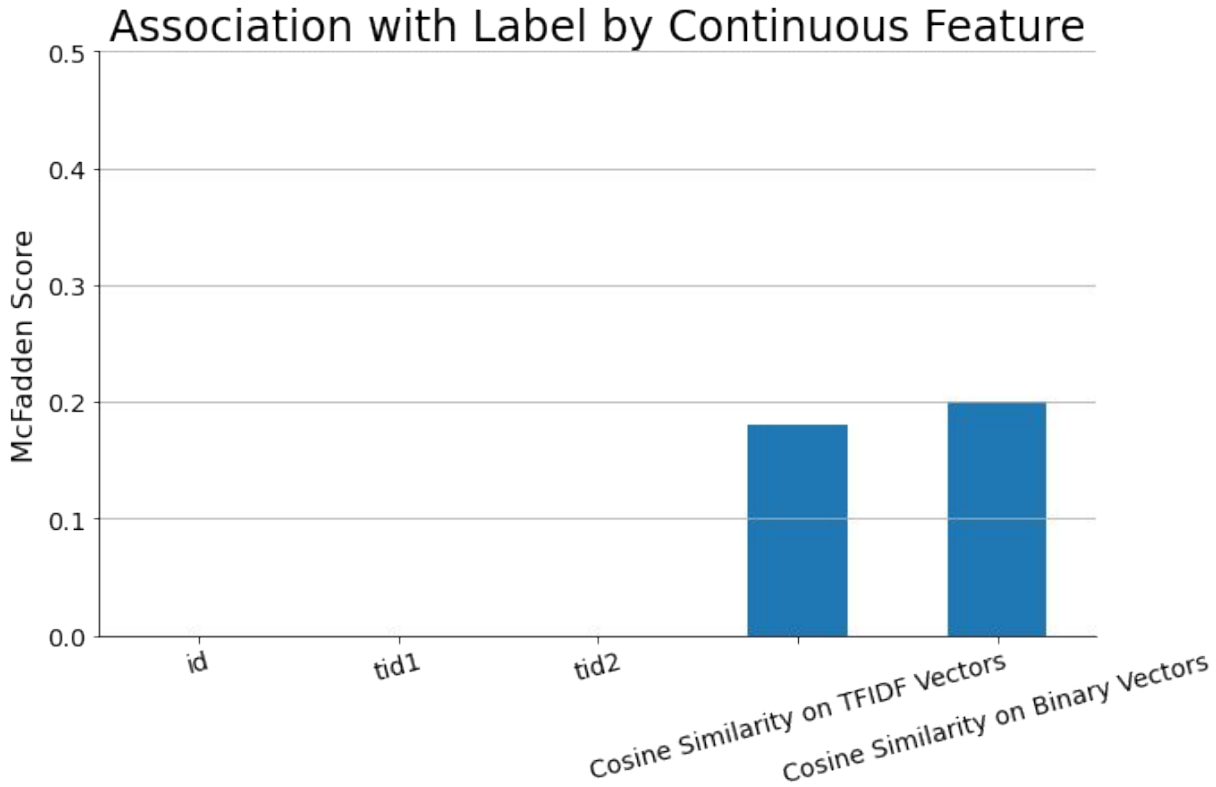
- Hypothesis:
  - Cosine similarity (normalized title 1 vector, normalized title 2 vector) => label association

Feature engineering - bucketing

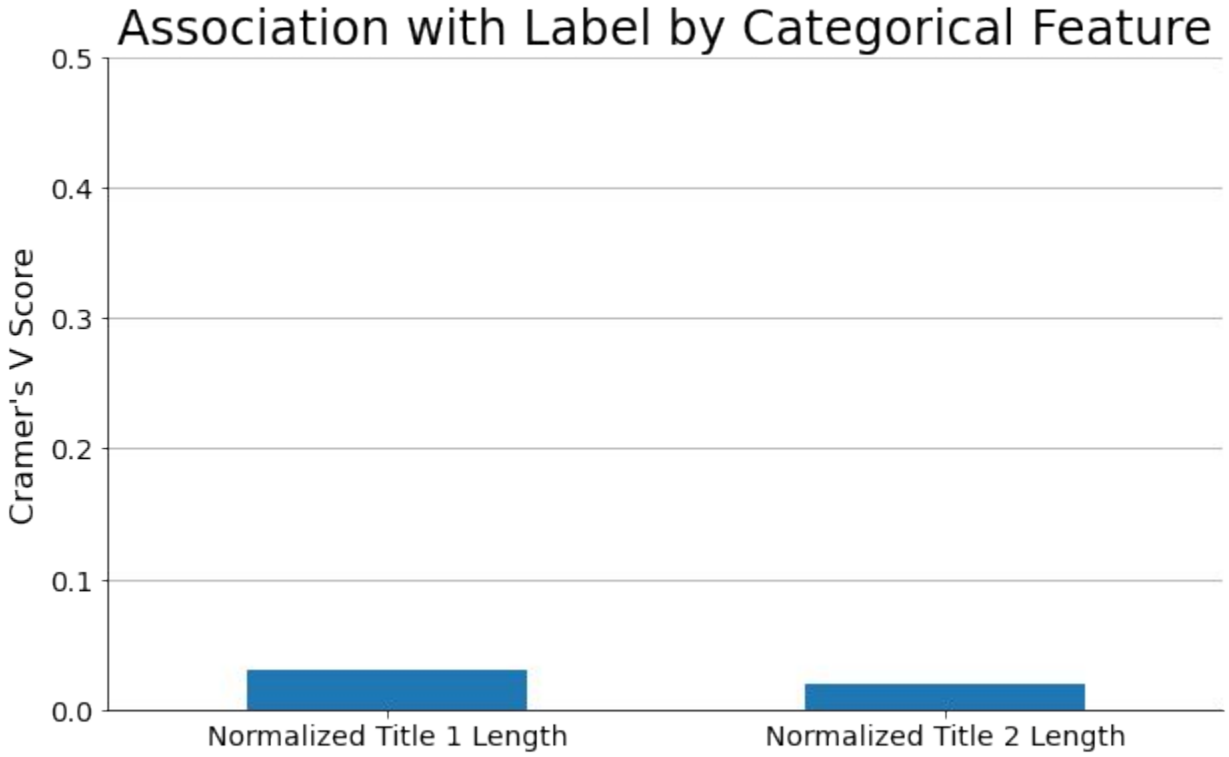
- normalized title length < 47 characters -> 1
- normalized title length < 62 characters -> 2
- normalized title length < 77 charters -> 3
- otherwise -> 4

Hypothesis: title length => label association

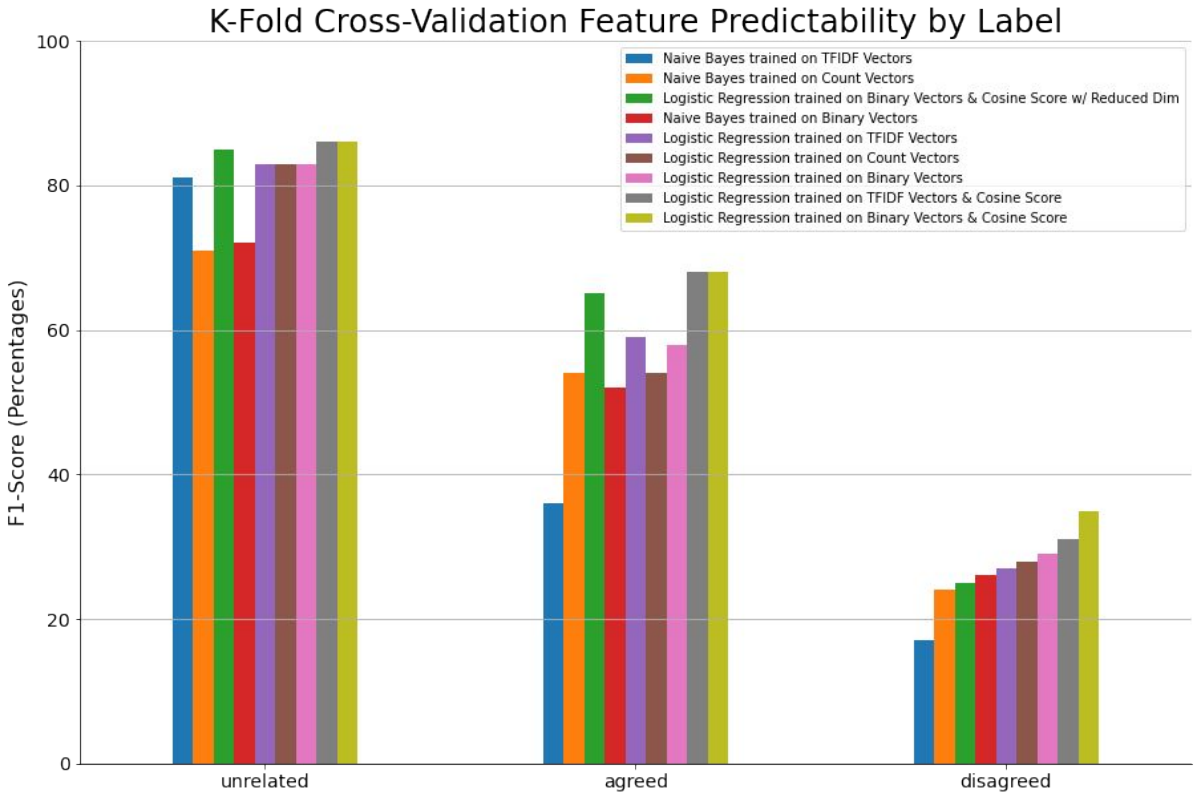
Feature Selection - Continuous features vs. categorical target



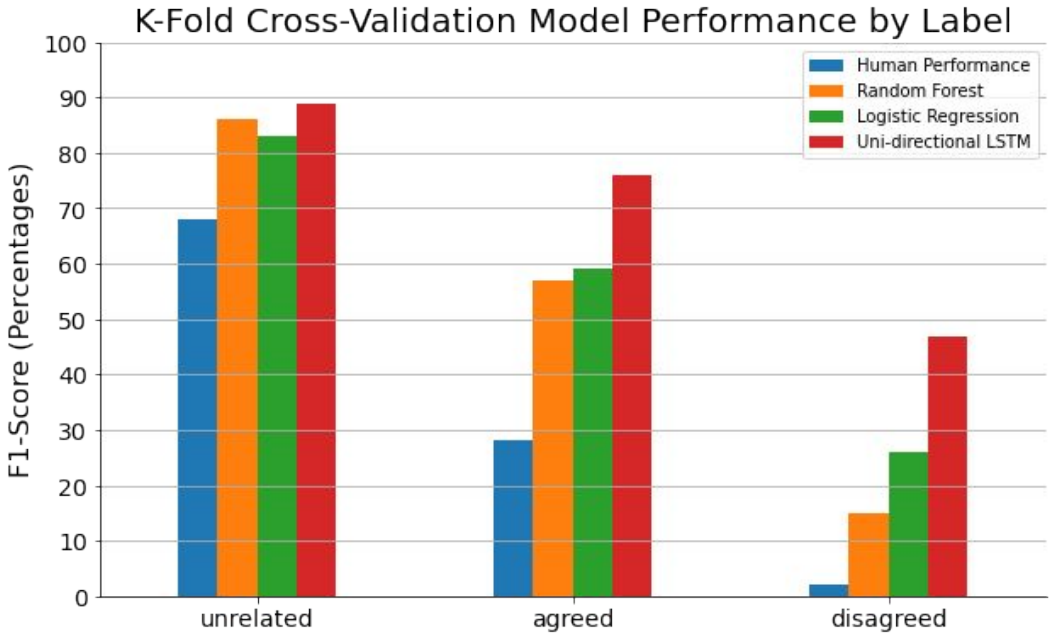
Feature Selection - Categorical (Nominal) features vs. categorical target



Feature Selection - Vectorized features



Model Selection



Binary BoW (word presence matters): A B C agreed with C B A
LSTM (word order && presence matter): A B C disagreed with C B A

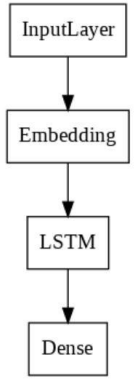
Hyperparameter Tuning - Random search & early stopping

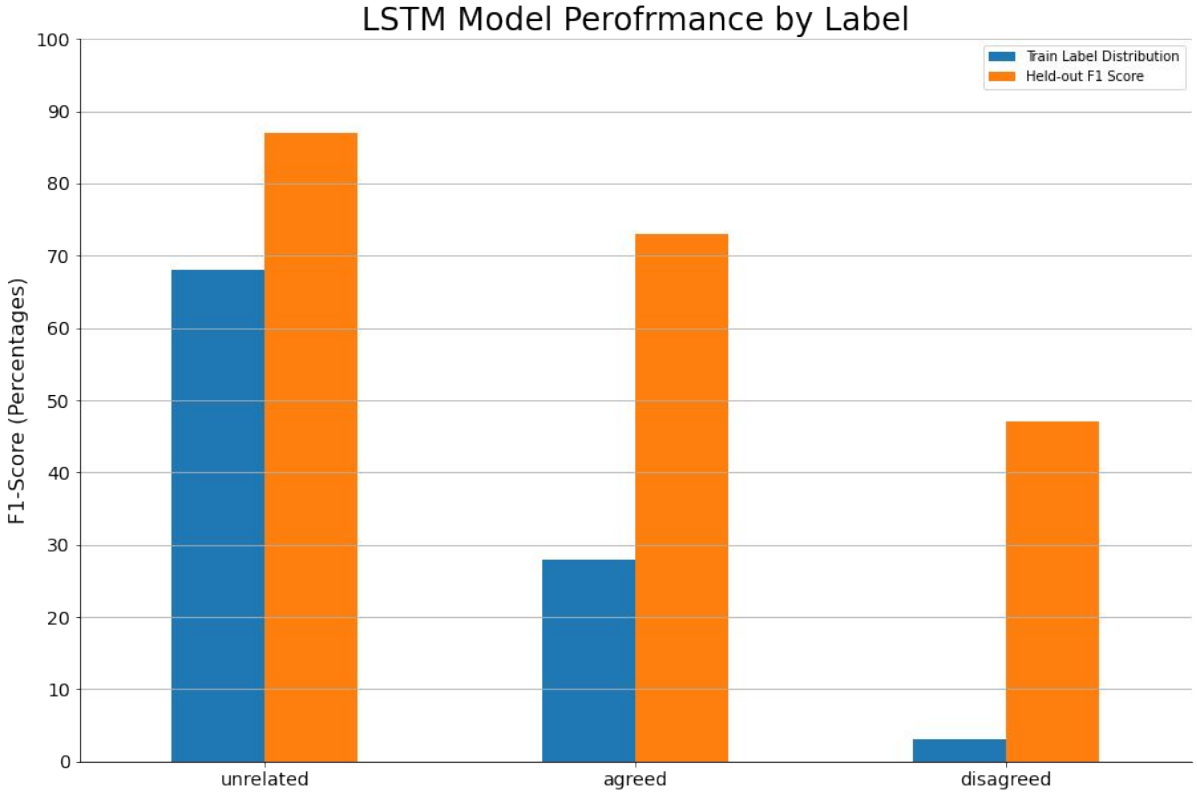
	vocabulary	sequence	embedding	units	layer	n_lyaer	unrelated	agreed	disagreed
0	2000	128	64	32	LSTM	1	0.87	0.70	0.41
1	2000	128	128	32	LSTM	1	0.87	0.70	0.43
2	2000	128	128	64	LSTM	1	0.87	0.71	0.43
3	2000	128	128	128	Bi-LSTM	1	0.87	0.71	0.46
4	5000	256	256	128	LSTM	1	0.87	0.73	0.50
5	5000	256	256	256	Bi-LSTM	1	0.88	0.73	0.41
6	5000	256	256	128	LSTM	2	0.88	0.74	0.48
7	5000	256	256	256	LSTM	1	0.88	0.73	0.48
8	8000	256	512	256	LSTM	1	0.88	0.74	0.49

vocabulary size || sequence length || embedding size || LSTM units || LSTM layers || LSTM directions

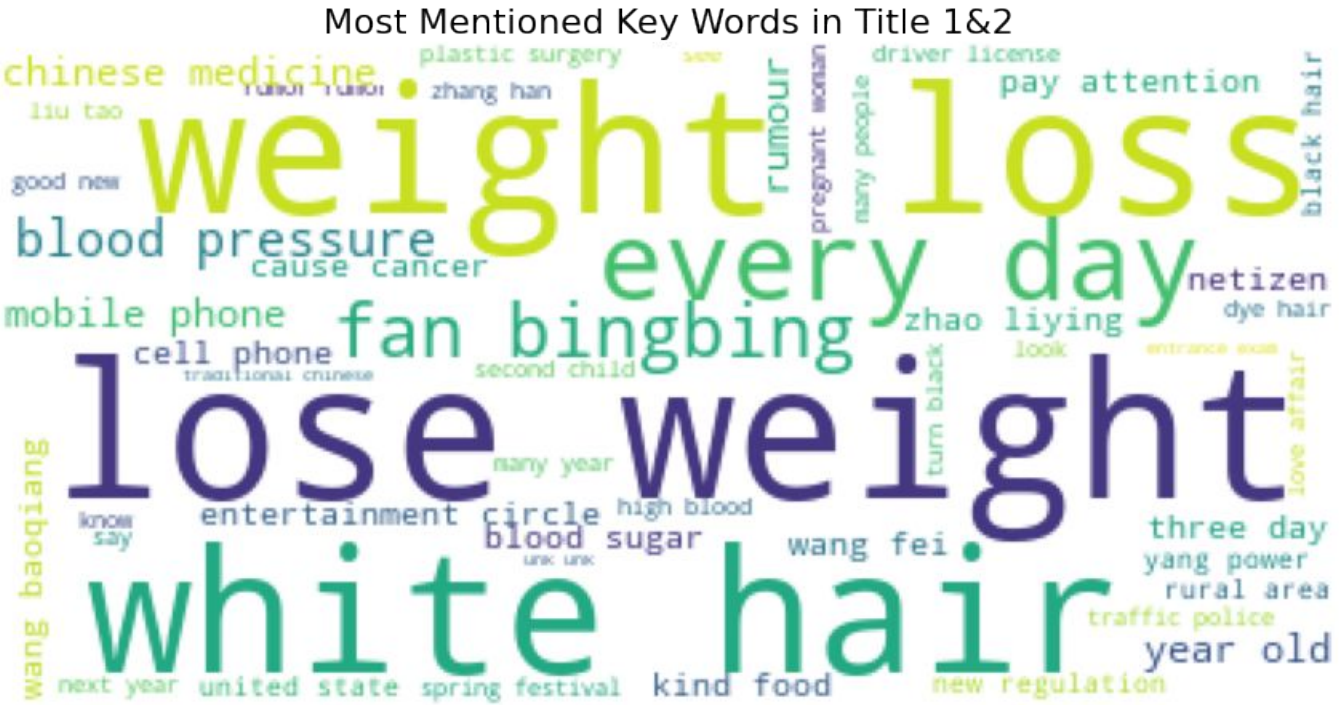
Final Deliverable - LSTM model

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 256, 256)          1280000
_________________________________________________________________
lstm (LSTM)                  (None, 128)               197120
_________________________________________________________________
dense (Dense)                (None, 3)                 387
=================================================================
Total params: 1,477,507
Trainable params: 1,477,507
Non-trainable params: 0
_________________________________________________________________
```





Bonus


Most Mentioned Key Words in Title 1&2
chinese medicine
plastic surgery
driver license
zhang han
pay attention
liu tao
weight loss
rumour
pregnant woman
many people
black hair
good new
blood pressure
every day
cause cancer
netizen
dye hair
mobile phone
fan bingbing
zhao liying
cell phone
look
second child
lose weight
turn black
love affair
many year
entertainment circle
high blood
three day
blood sugar
wang fei
yang power
wang baoqiang
rural area
white hair
traffic police
year old
next year
united state
spring festival
kind food
new regulation