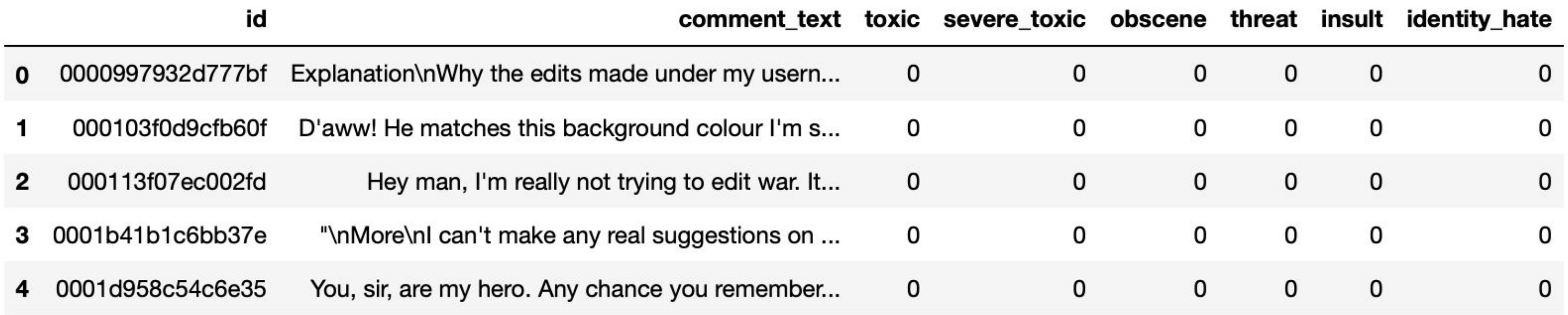
Identifying Toxic Comments by Internet Trolls

Group 5: Mengran Xia, Caihui Xiao, Xinyi Chen, Weixuan Wu, Qiaozhen Ma

Data

- **159,571** comments, **90.4%** are clean comments, and **9.6%** are toxic
- Toxic comments are classified into the following categories: **severe toxic**, **obscene**, **threat**, **insult**, and **identity hate**.
- One comment can have multiple labels.



	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0

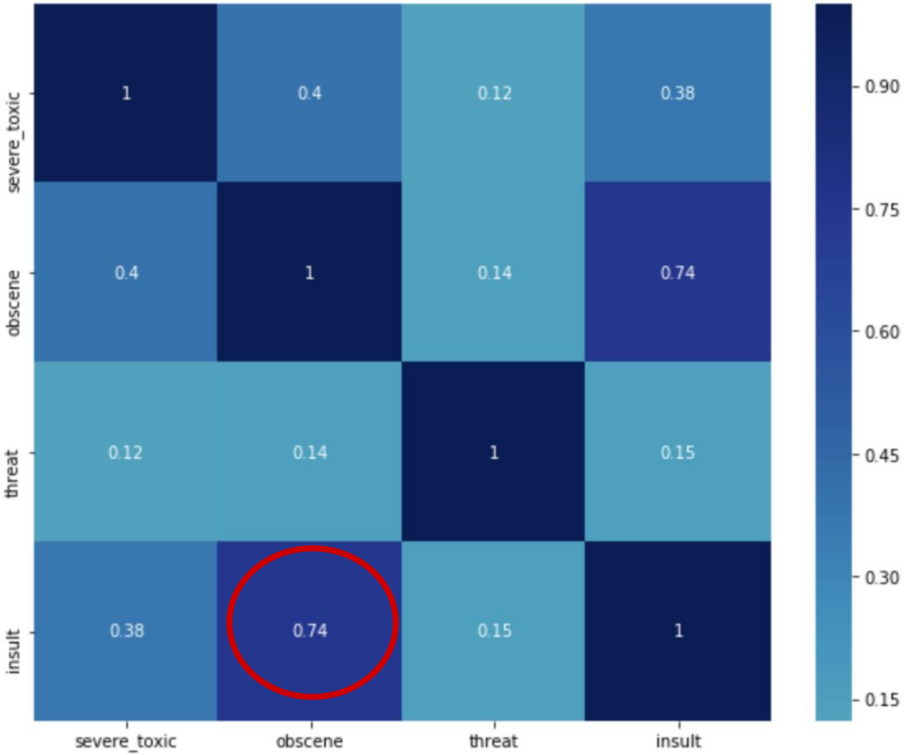
Some Examples

“COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK”

“Fuck you, block me, you faggot pussy!”

“FUCK YOUR FILTHY MOTHER IN THE ASS, DRY!”

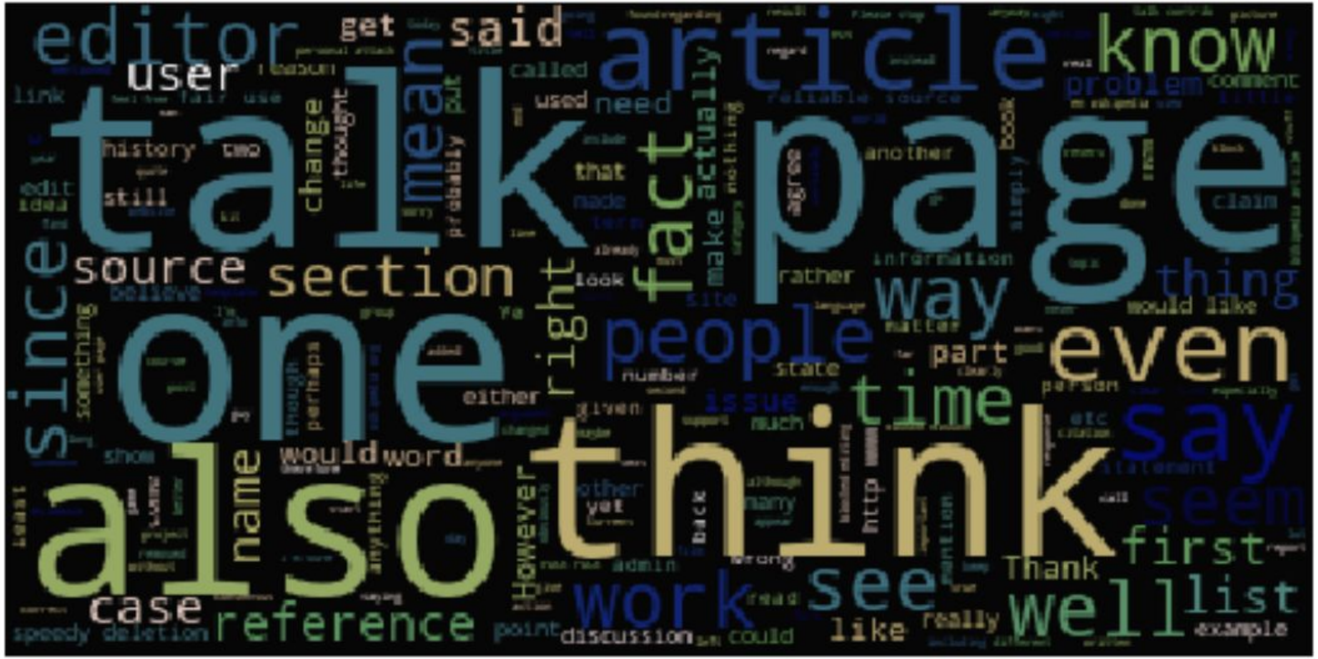
Exploratory Data Analysis



	severe_toxic	obscene	threat	insult
severe_toxic	1	0.4	0.12	0.38
obscene	0.4	1	0.14	0.74
threat	0.12	0.14	1	0.15
insult	0.38	0.74	0.15	1

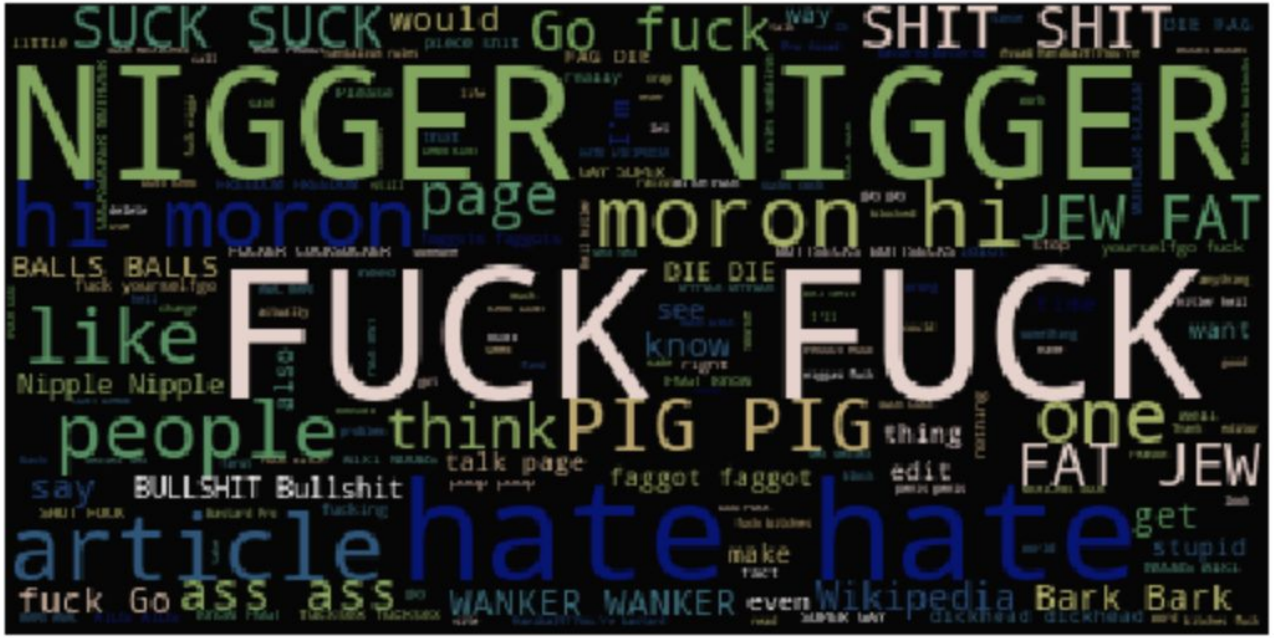
Word Cloud - All Comments

Words frequented in all Comments



Word Cloud - Toxic

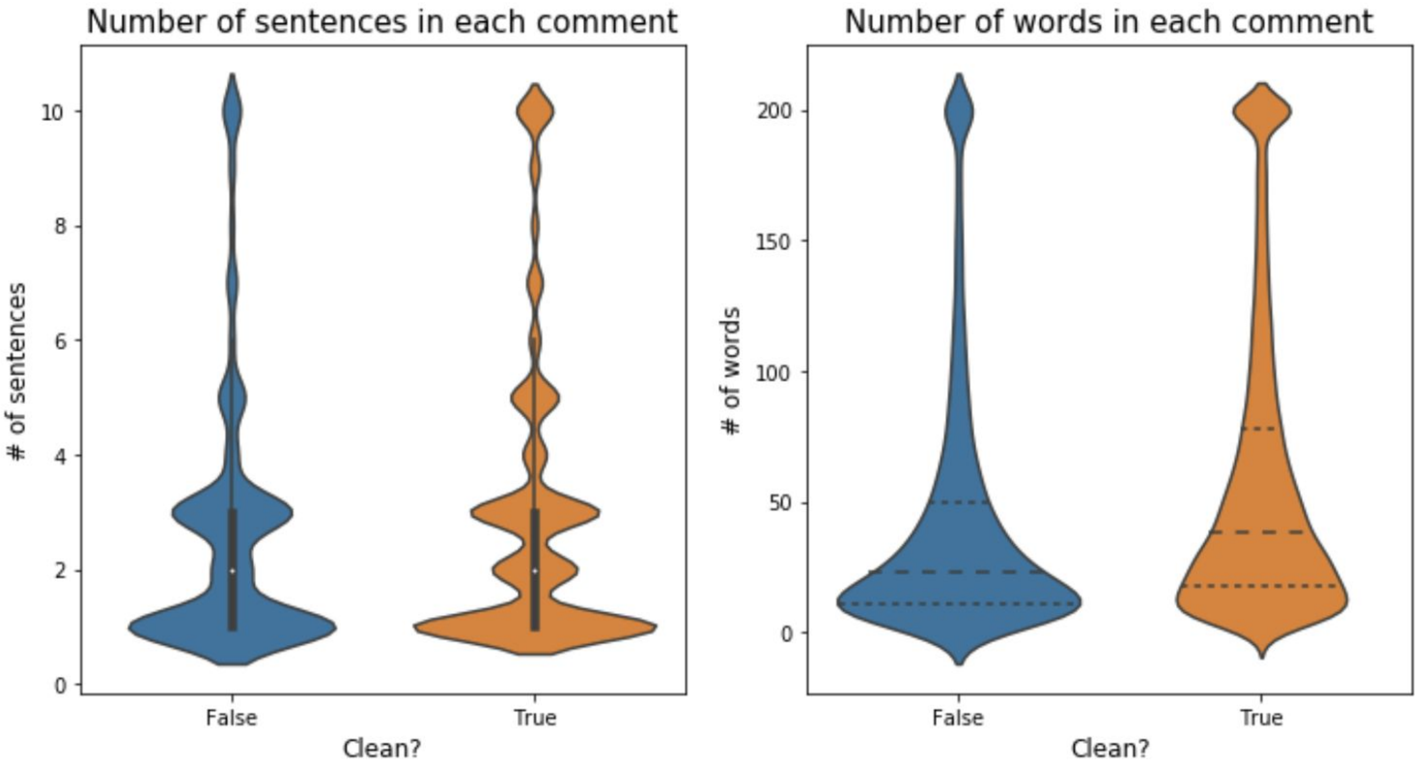
Words frequented in Toxic Comments



Word Cloud - Severe Toxic

FUCK SHUT
SHIT SHIT
SHUT FUCK
ass ass
NIGGER NIGGER
HUGE FAGGOT
Go fuck
FUCKER
NOTHJER
COCKSUCKER
fuck
Go
FUCK DOG
FAGGOT HUGE
MEXICANS SUCK
bastard
fucksex fucksex
Hanibal911You're bastard
faggot faggot
BITCH FUCK
fuck
yourselfgo
FUCK BITCH
DAMN CUNT
RAPE ANAL
SUCK SUCK
DIE DIE
SUCK COCK
FUCK FUCK
ANAL RAPE
FAT SHIT
SUCK DICK
MEXICANS
yourselfgo
NOTHJER
COCKSUCKER
fuck
Assad Hanibal911You're

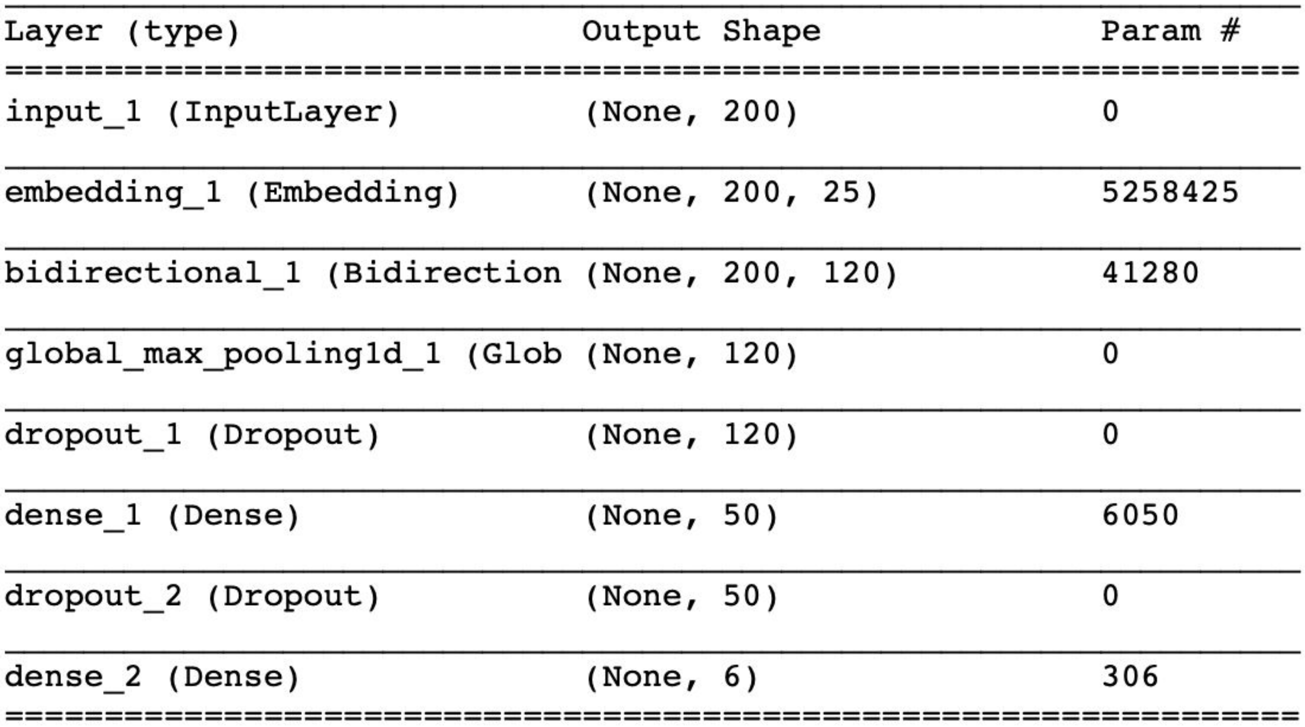
Are Toxic Comments Longer or Shorter?



Word Embedding

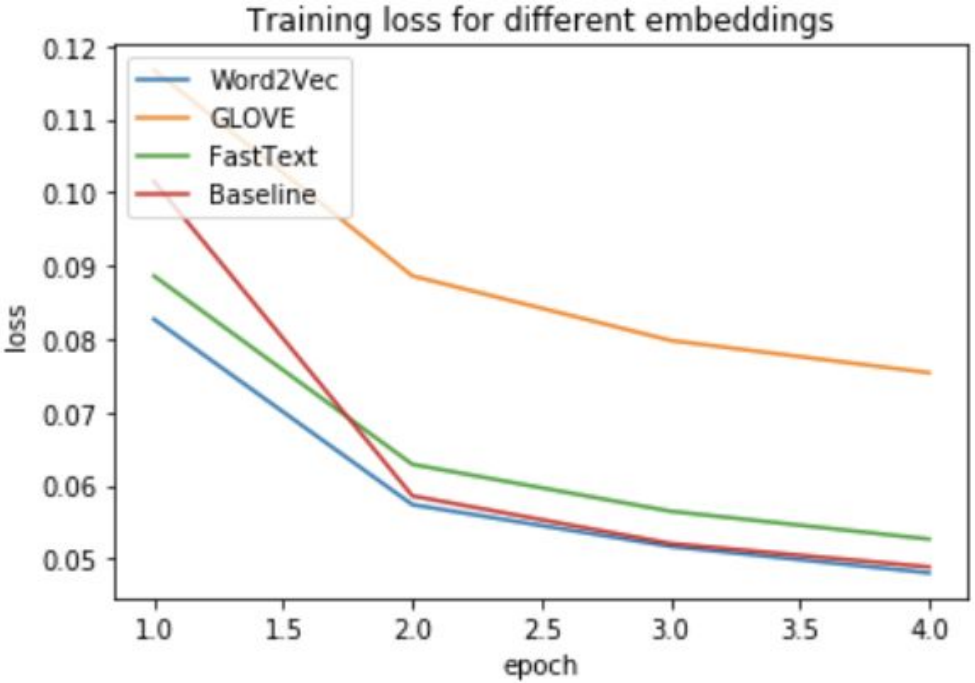
1. Baseline
   a. trained using the toxic comments text
2. Pre-trained Embedding
   a. **GloVe:**
      i. Twitter 25 dimension
   b. **word2vec**: Google News Negative 300
   c. **fastText**
      i. English Word Vectors: Pretrained on English Webcrawler and Wikipedia

Model Architecture

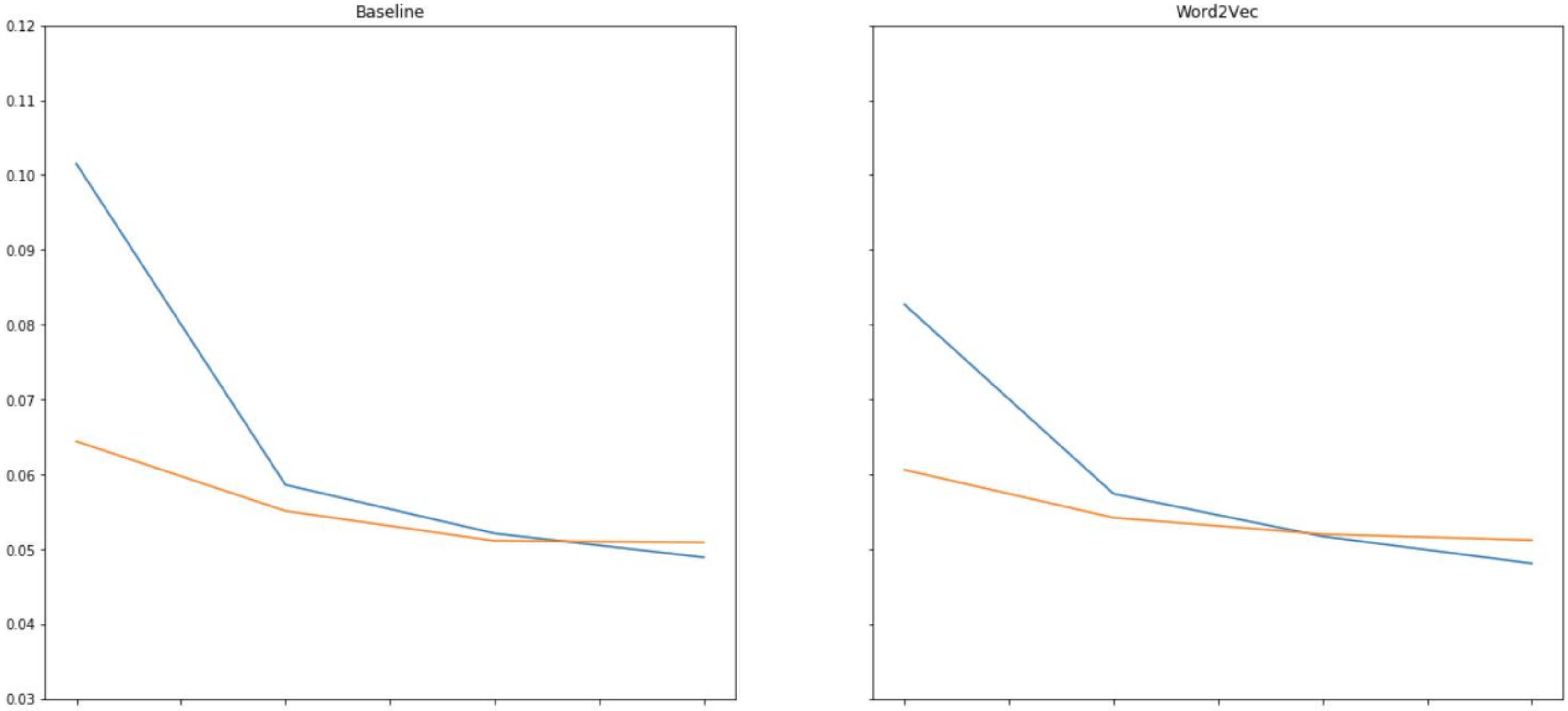


```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 200)               0
_________________________________________________________________
embedding_1 (Embedding)      (None, 200, 25)           5258425
_________________________________________________________________
bidirectional_1 (Bidirection (None, 200, 120)          41280
_________________________________________________________________
global_max_pooling1d_1 (Glob (None, 120)               0
_________________________________________________________________
dropout_1 (Dropout)          (None, 120)               0
_________________________________________________________________
dense_1 (Dense)              (None, 50)                6050
_________________________________________________________________
dropout_2 (Dropout)          (None, 50)                0
_________________________________________________________________
dense_2 (Dense)              (None, 6)                 306
=================================================================
```

Comparing Different Embedding Methods



Baseline vs Word2Vec



Hyper-parameter Tuning for Baseline Model

1. Optimizer
   a. Adam, SGD, Adagrad, Adadelta, **Adamax**, Nadam, RMSprop
2. Learning Rate
   a. 0.001, **0.01**, 0.1, 0.2, 0.3
3. Batch size
   a. 32, 128, **142**

Results

	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss
Epoch 1	0.9898	0.0256	0.9857	0.0407
Epoch 2	0.9909	0.0230	0.9856	0.0437

Thanks!