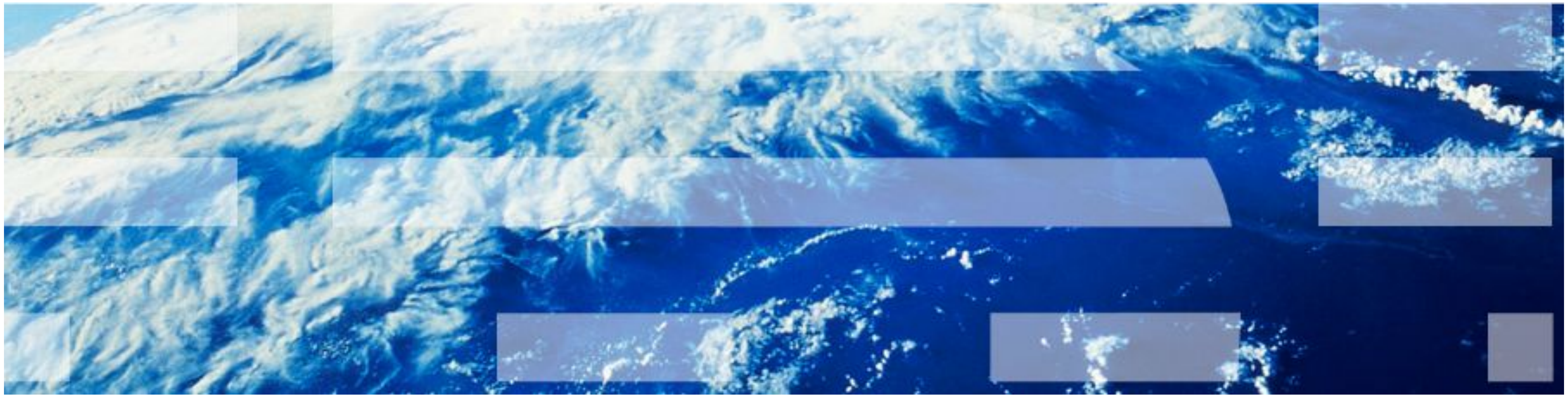
E6893 Big Data Analytics:

Advanced Sentiment Analysis

Team Members (with UNI):

Shan Guan (sg3506)
Yuehan Kong (yk2756)
Lin Jiang (lj2438)

Project ID: 201812-17



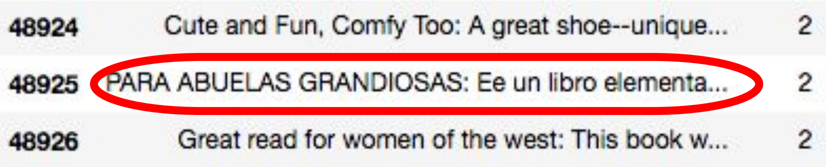
Motivation

- Sentiment analysis allows us to gain an overview of the wider public opinion behind certain topics.
- Applications of sentiment analysis are broad and powerful. eg: The Obama administration use this method to gauge public opinion to policy announcements and campaign messages ahead of 2012 presidential election.
- It is essential for market research and customer service. Not only you can see what people think of your own products or services, but also what they think about your competitors.
- However, we do not have enough labeled data to do classification(positive or negative sentiment) in real life and the prediction accuracy is not satisfying most time.

Dataset

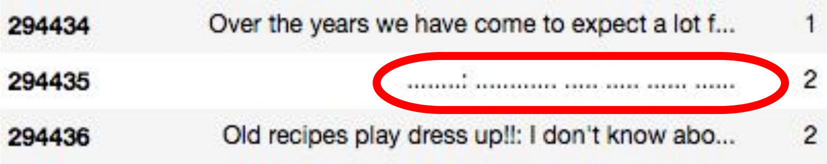
https://www.kaggle.com/bittlingmayer/amazonreviews

- Amazon Reviews for Sentiment Analysis
  - Input text: few million Amazon customer reviews
  - Output label: star ratings [1-negative review; 2-positive review]
- Data dimension
  - Training data: [3600000*2]
  - Testing data: [400000*2]



48924	Cute and Fun, Comfy Too: A great shoe--unique...	2
48925	PARA ABUELAS GRANDIOSAS: Ee un libro elementa...	2
48926	Great read for women of the west: This book w...	2

spanish!!!

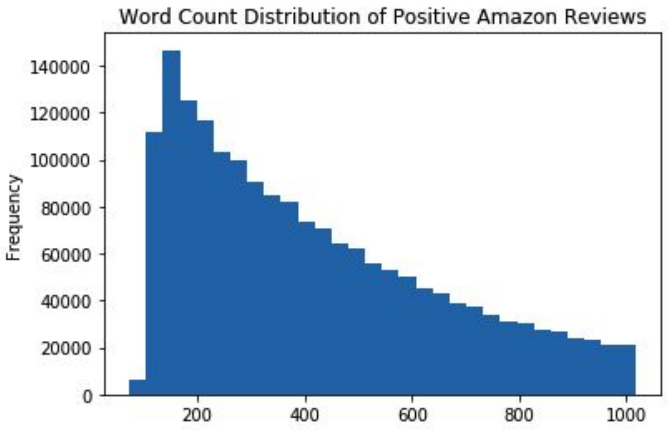


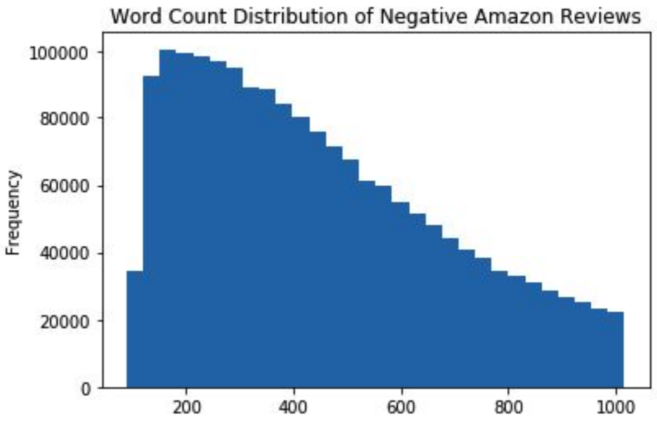
294434	Over the years we have come to expect a lot f...	1
294435	........:	2
294436	Old recipes play dress up!!: I don't know abo...	2

no text!!!

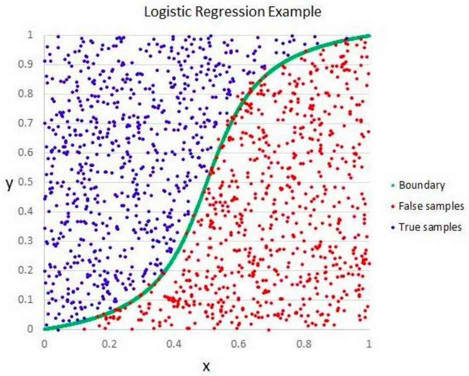
Interesting Finding

- Word count distribution of positive and negative Amazon reviews

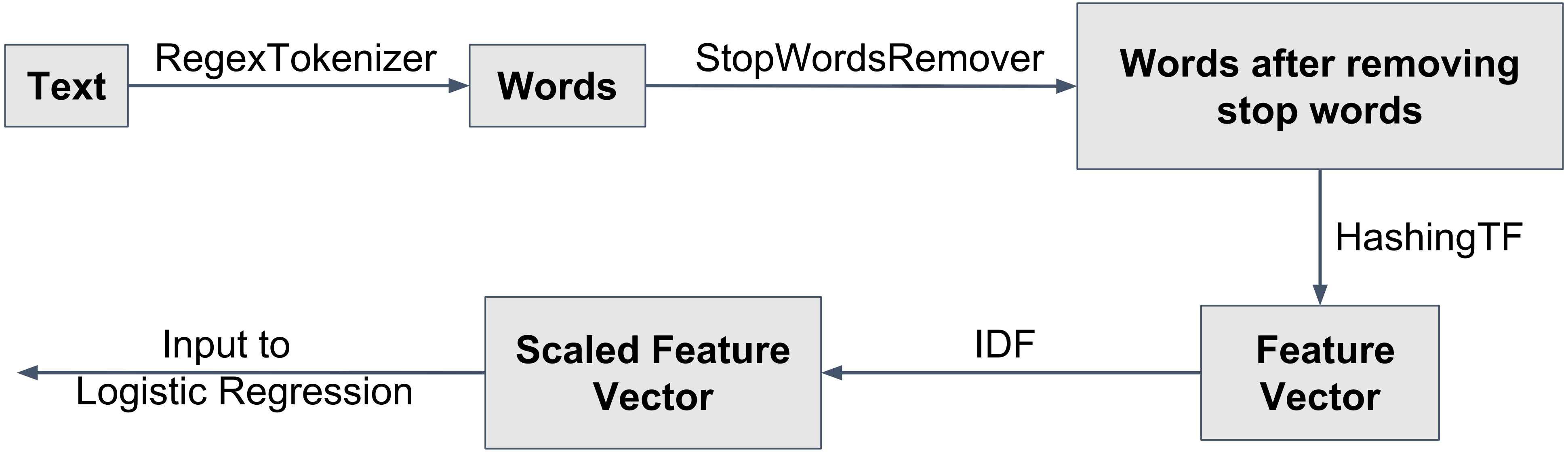




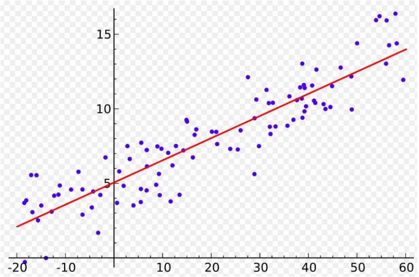
Algorithm



- 1st - TFIDF & Logistic Regression



Text → RegexTokenizer → Words → StopWordsRemover → Words after removing stop words → HashingTF → Feature Vector → IDF → Scaled Feature Vector → Input to Logistic Regression



Algorithm

- 2nd - FastText (gensim)
  - Unsupervised learning algorithm by Facebook
  - Preprocessing
    - Tokenizing: clean text and convert to word level
  - Word Embedding Vector
    - extension to Word2Vec. Instead of feeding individual words into the CNN, FastText breaks words into several n-grams (sub-words).
    - Obtaining vector representations for words
  - Model: CNN-LSTM in Keras

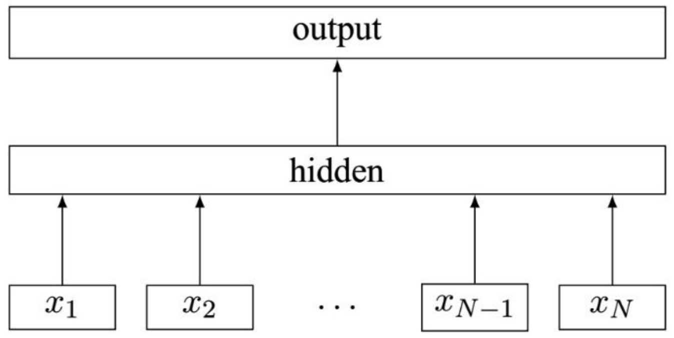
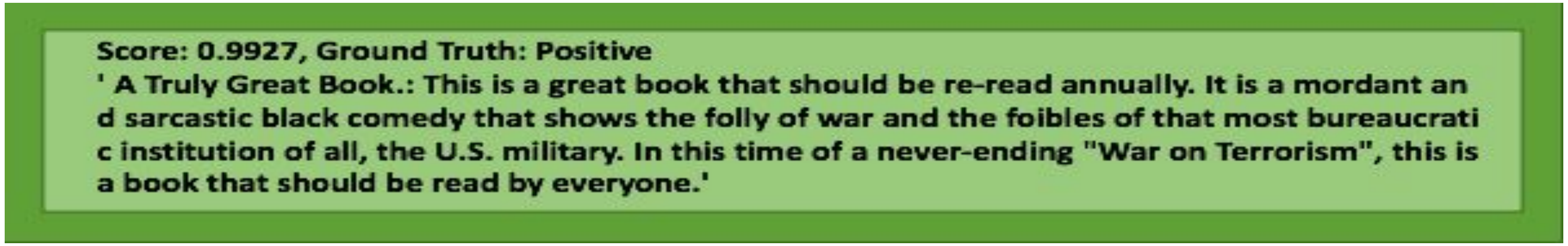


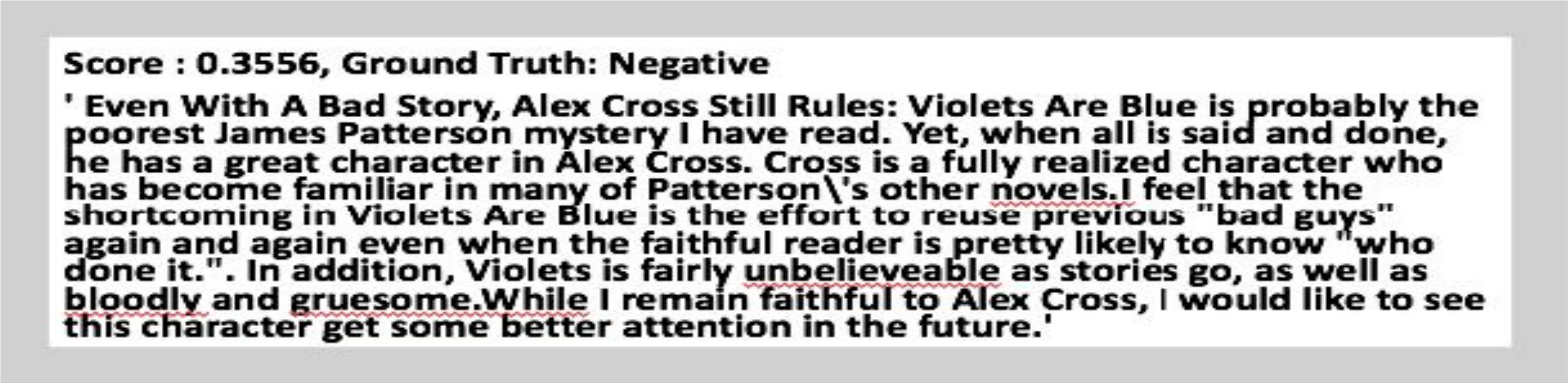
Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.

FastText Sentiment Analysis Results



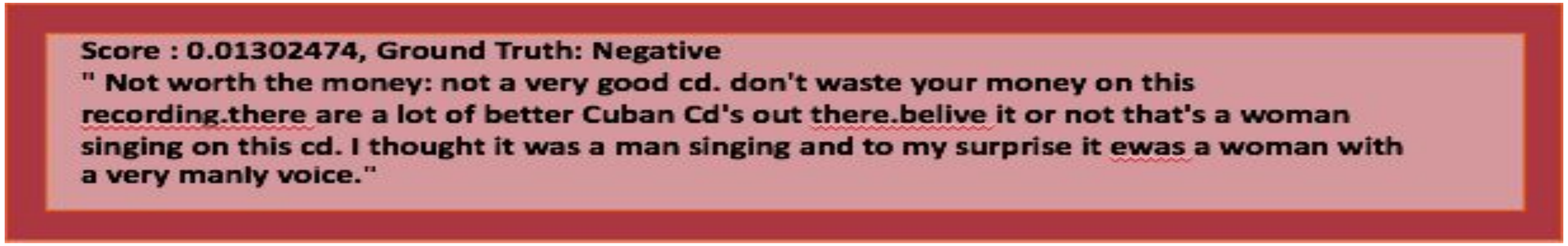
Score: 0.9927, Ground Truth: Positive

' A Truly Great Book.: This is a great book that should be re-read annually. It is a mordant and sarcastic black comedy that shows the folly of war and the foibles of that most bureaucratic institution of all, the U.S. military. In this time of a never-ending "War on Terrorism", this is a book that should be read by everyone.'



Score : 0.3556, Ground Truth: Negative

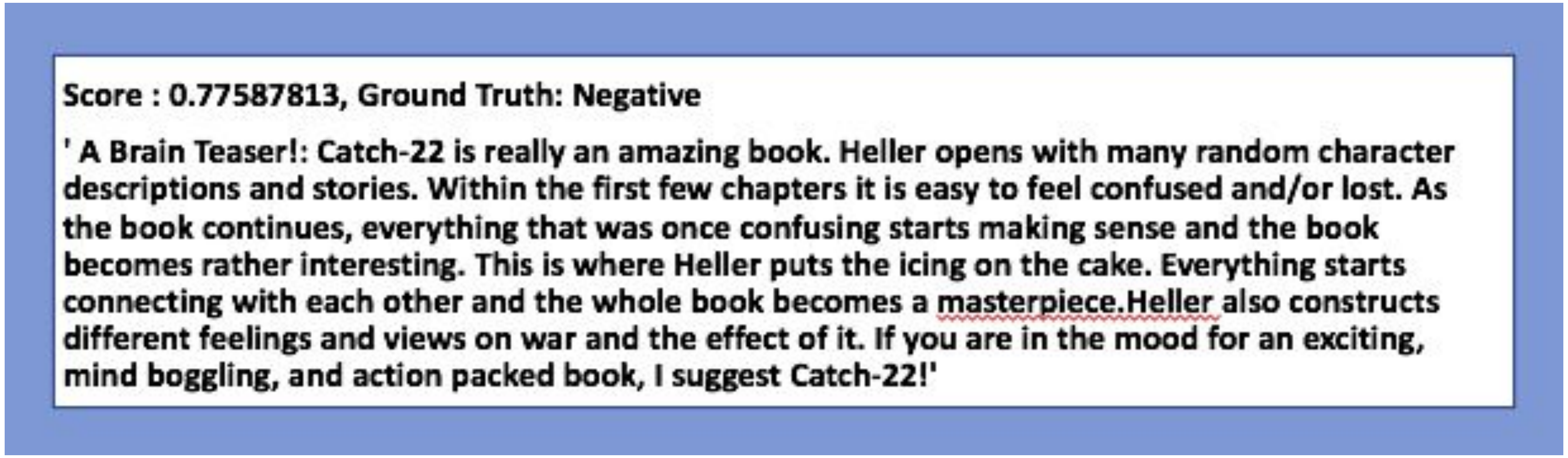
' Even With A Bad Story, Alex Cross Still Rules: Violets Are Blue is probably the poorest James Patterson mystery I have read. Yet, when all is said and done, he has a great character in Alex Cross. Cross is a fully realized character who has become familiar in many of Patterson\'s other novels.I feel that the shortcoming in Violets Are Blue is the effort to reuse previous "bad guys" again and again even when the faithful reader is pretty likely to know "who done it.". In addition, Violets is fairly unbelieveable as stories go, as well as bloodly and gruesome.While I remain faithful to Alex Cross, I would like to see this character get some better attention in the future.'



Score : 0.01302474, Ground Truth: Negative

" Not worth the money: not a very good cd. don't waste your money on this recording.there are a lot of better Cuban Cd's out there.belive it or not that's a woman singing on this cd. I thought it was a man singing and to my surprise it ewas a woman with a very manly voice."

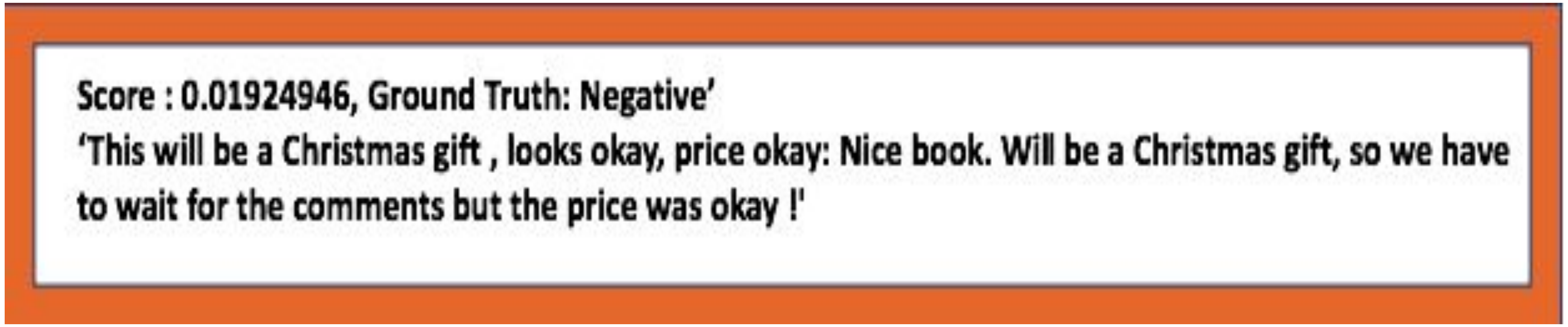
FastText Sentiment Analysis Results (Wrong Prediction)



Score : 0.77587813, Ground Truth: Negative

' A Brain Teaser!: Catch-22 is really an amazing book. Heller opens with many random character descriptions and stories. Within the first few chapters it is easy to feel confused and/or lost. As the book continues, everything that was once confusing starts making sense and the book becomes rather interesting. This is where Heller puts the icing on the cake. Everything starts connecting with each other and the whole book becomes a masterpiece.Heller also constructs different feelings and views on war and the effect of it. If you are in the mood for an exciting, mind boggling, and action packed book, I suggest Catch-22!'

Wrong label !!!



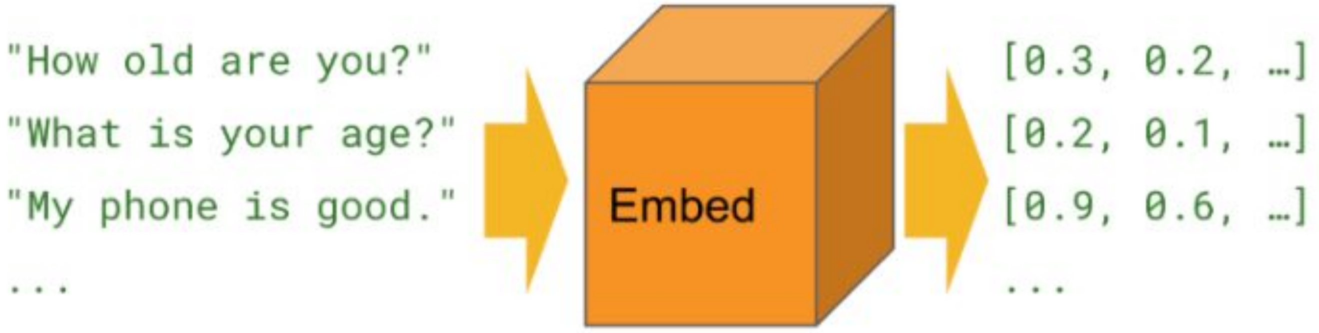
Score : 0.01924946, Ground Truth: Negative'

'This will be a Christmas gift , looks okay, price okay: Nice book. Will be a Christmas gift, so we have to wait for the comments but the price was okay !'

Wrong prediction

Too many negative words

Algorithm



- 3rd - Universal Sentence Encoder
  - What is universal sentence encoder? (encode text into vector)
    - Input: variable length english text
    - Output: a 512 dimensional vector
  - What could it do?
    - Text Classification
    - Semantic Similarity
    - Clustering
  - Why use it?
  - Surprisingly good performance(precision) with minimal amounts of supervised training data(first released in 29 Mar 2018)
  -

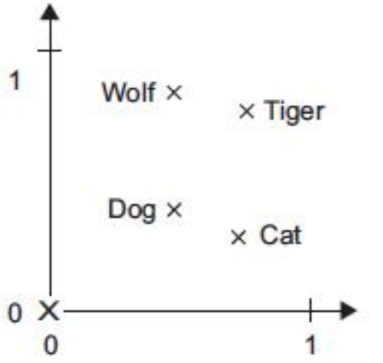
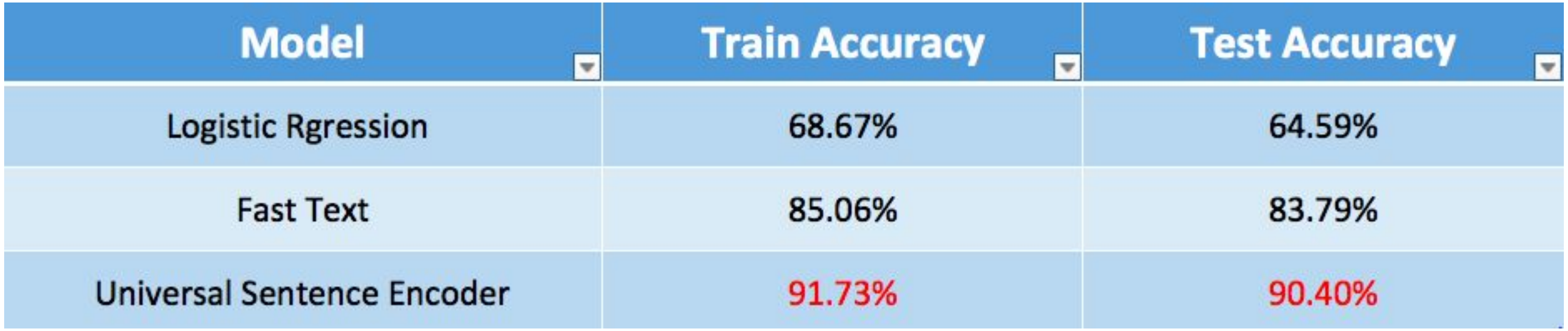


Figure 6.3 **A toy example of a word-embedding space**

Tools

- Python
- Spark / Pyspark
- Tensorflow / Keras
- Google Cloud Platform

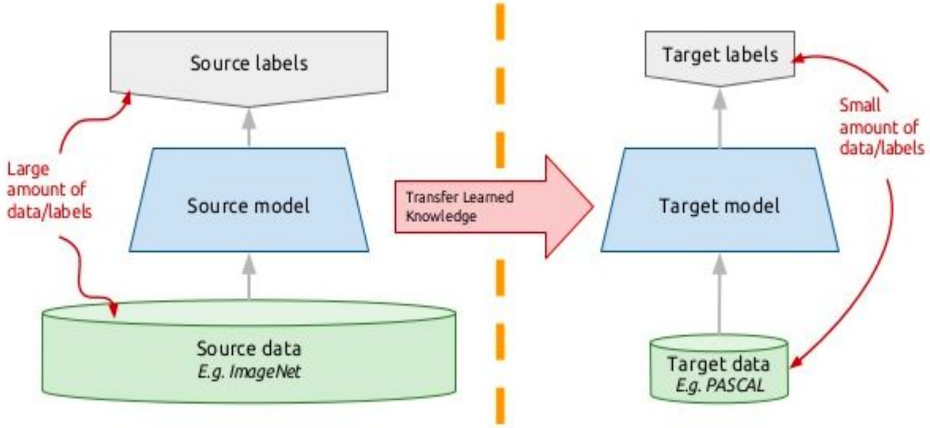
Model Comparison

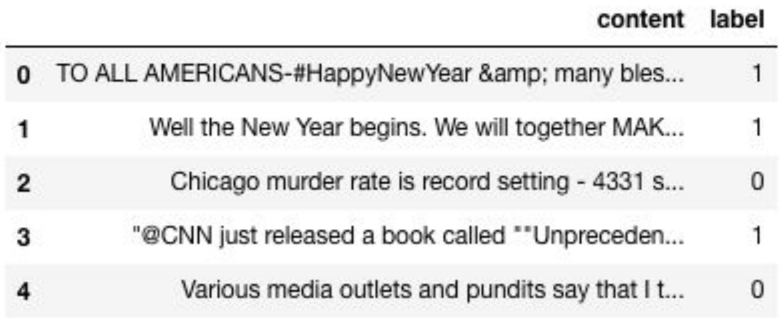


Model	Train Accuracy	Test Accuracy
Logistic Rgression	68.67%	64.59%
Fast Text	85.06%	83.79%
Universal Sentence Encoder	91.73%	90.40%

Introduction of Transfer Learning

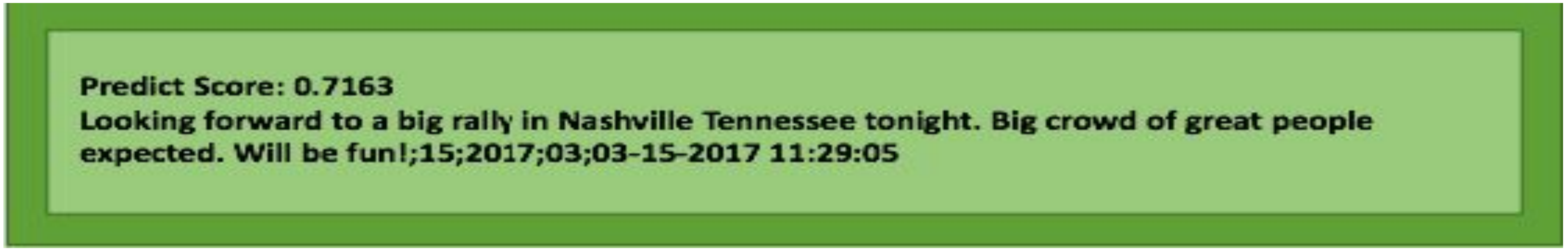
- Leverage prior knowledge from one domain and task into a different domain and task
- Inspired by us - humans who have inherent ability to not learn everything from scratch
- Deal with dataset has insufficient label
- World-class pre-trained model based on excellent big datasets like imagenet



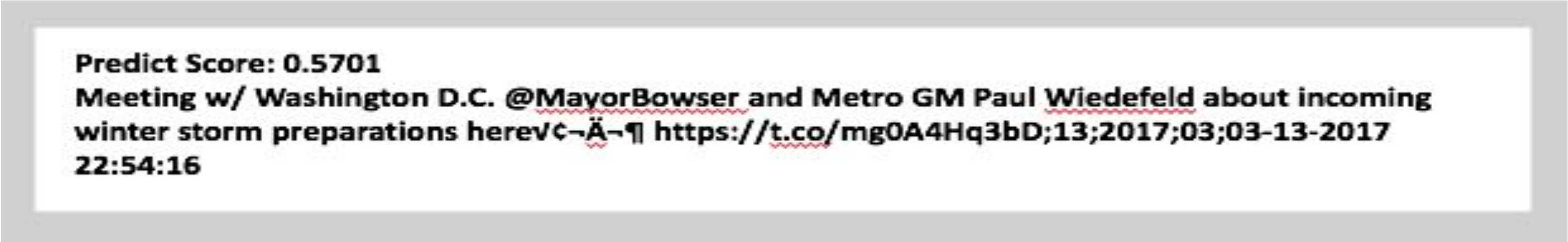


	content	label
0	TO ALL AMERICANS-#HappyNewYear &amp; many bles...	1
1	Well the New Year begins. We will together MAK...	1
2	Chicago murder rate is record setting - 4331 s...	0
3	"@CNN just released a book called ""Unpreceden...	1
4	Various media outlets and pundits say that I t...	0

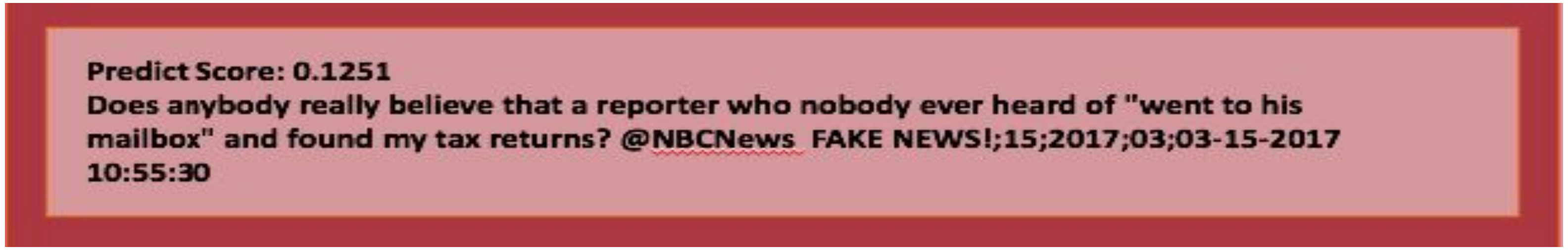
Transfer Learning Sentiment Result



Predict Score: 0.7163
Looking forward to a big rally in Nashville Tennessee tonight. Big crowd of great people expected. Will be fun!;15;2017;03;03-15-2017 11:29:05



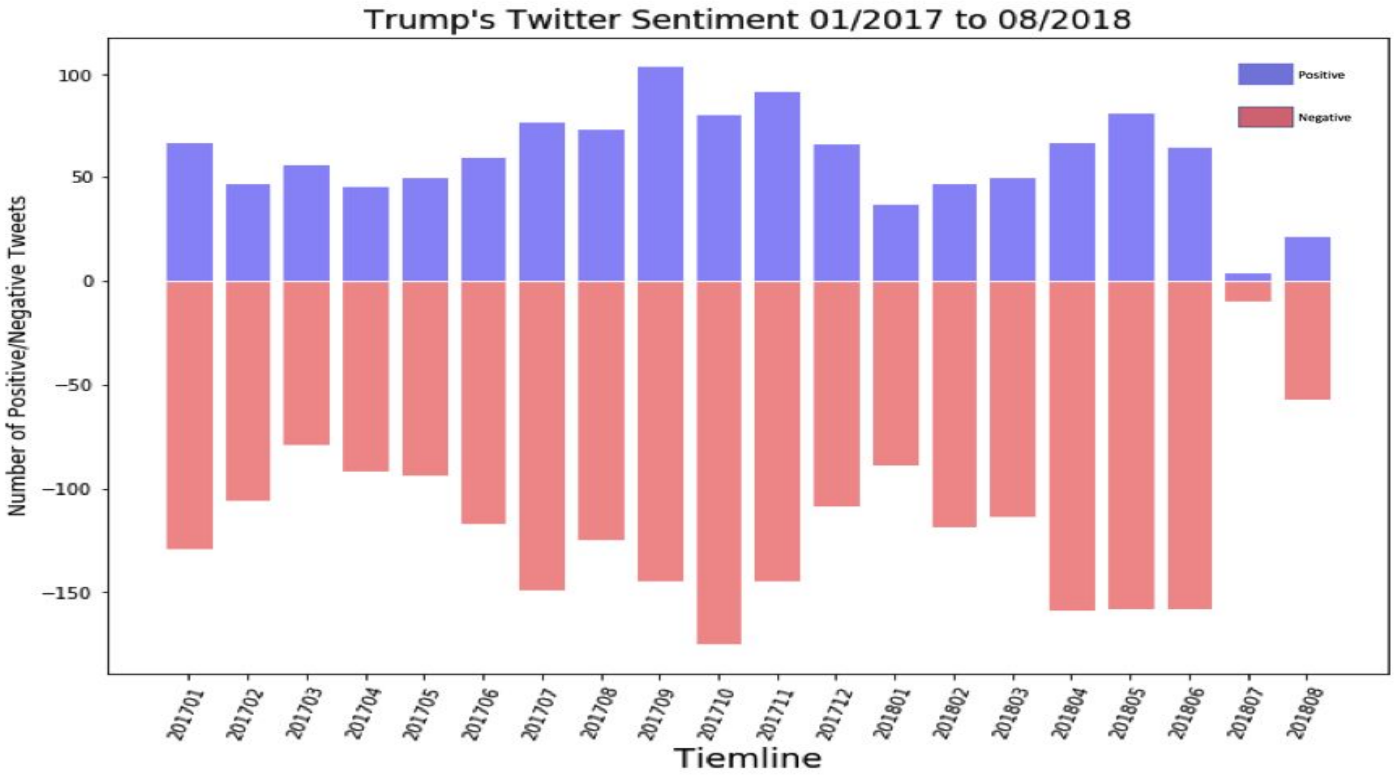
Predict Score: 0.5701
Meeting w/ Washington D.C. @MayorBowser and Metro GM Paul Wiedefeld about incoming winter storm preparations here√¢¬Ä¬¶ https://t.co/mg0A4Hq3bD;13;2017;03;03-13-2017 22:54:16



Predict Score: 0.1251
Does anybody really believe that a reporter who nobody ever heard of "went to his mailbox" and found my tax returns? @NBCNews FAKE NEWS!;15;2017;03;03-15-2017 10:55:30

Visualization by Transform Learning

- Dataset: Trump Twitter http://www.trumptwitterarchive.com/archive



Future Work

Transfer learning with Universal Sentence embedding could be very useful when we do not have enough labeled data but achieve high precision.

- want to know how does sentiment of China-US trade war in news change over time
- want to know how does people’s sentiment about Gene modified organism change in social media over time

Questions?