Twitter Sentiment Evaluation

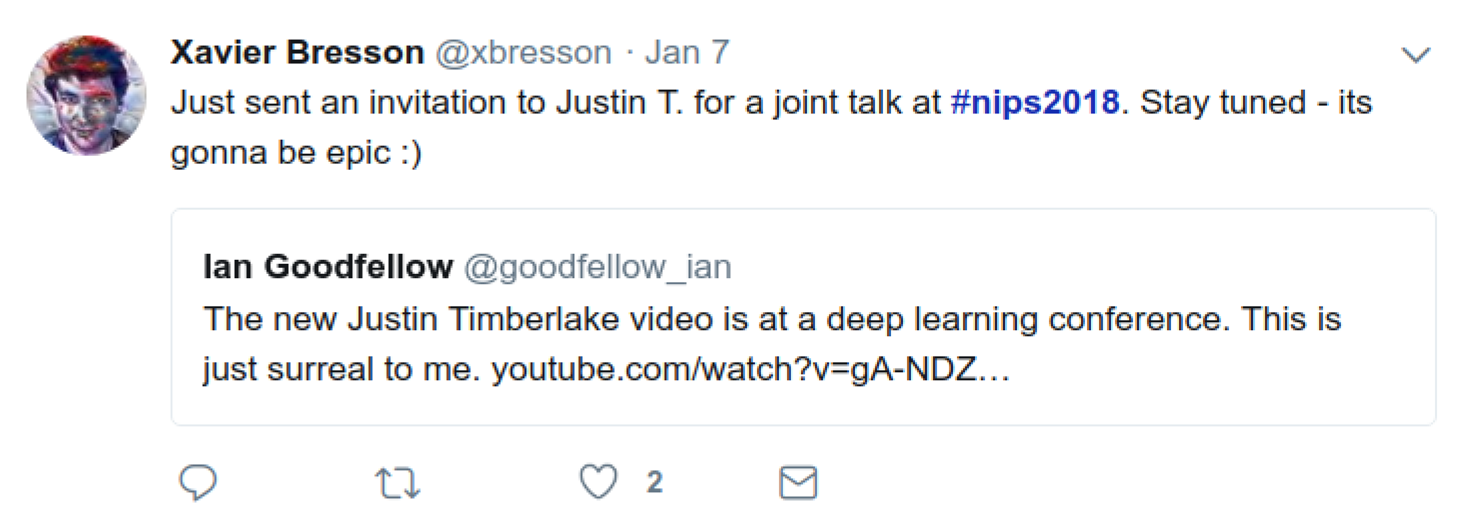
Georgios Balikas (@balikasg)

Short Bio

- A data science enthusiast!
- PhD from Univ. Grenoble-Alps (10/2017)
- Data scientist at Kelkoo

Definition: Sentiment Analysis

- “computationally identify and categorize the opinions expressed in a piece of text; determine whether positive/neutral/negative toward a topic/product..” [Oxford Dict.]



Sentiment analysis

- Business/Market understanding
  - Brand or product health
  - dealing with a crisis, ..
- Enables exciting applications
- Challenging problem

A (possible) taxonomy

- Binary
- Ternary
- Fine-grained

Why Twitter ?

- Lots of (real-time) data
- Public APIs
- Several languages
- Rich content (e.g., graph properties)

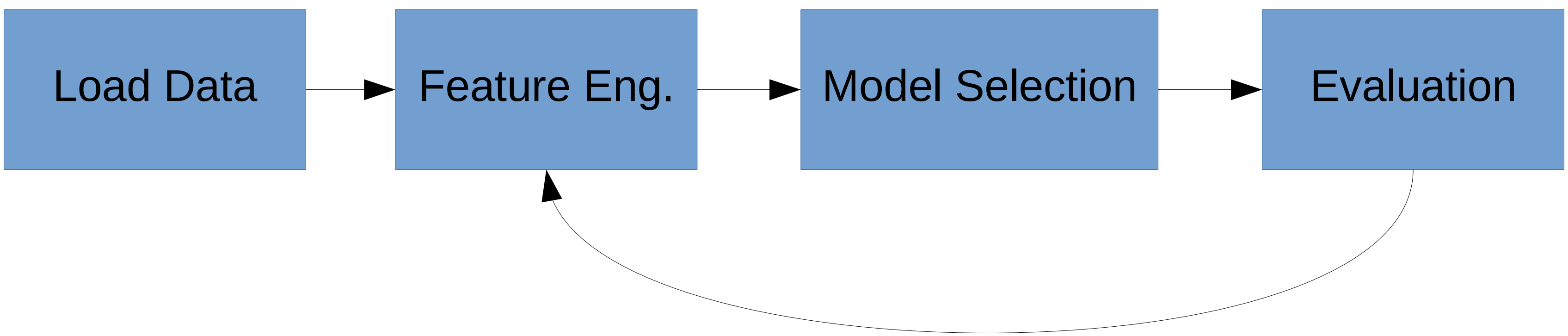
The fine-grained case

Very Negative	@Microsoft how about you make a system that doesn't eat my friggin discs. This is the 2nd time this has happened and I am so sick of it!
Negative	Thanks to @microsoft, I just may be switching over to @apple.
Neutral	@Microsoft the option should be in Windows update. You've got a month. Clean install may be better..
Positive	Microsoft, I may not prefer your gaming branch of business. But, you do make a damn fine operating system. #Windows10 @Microsoft
Very Positive	@Microsoft - congratulations on the 20th Birth Anniversary of @Windows 95. 20 years since we've come to love you (& backward compatibility)

The specificity of tweets

- Mentions (@balikasg)
- Short text: 140 chars, now 280
- Hashtags (#nips, #bad, #great)
- Emoticons ;-), :-D, :-(
- Punctuation
- Creative language: w8, 2morrow, w

The ML Pipeline

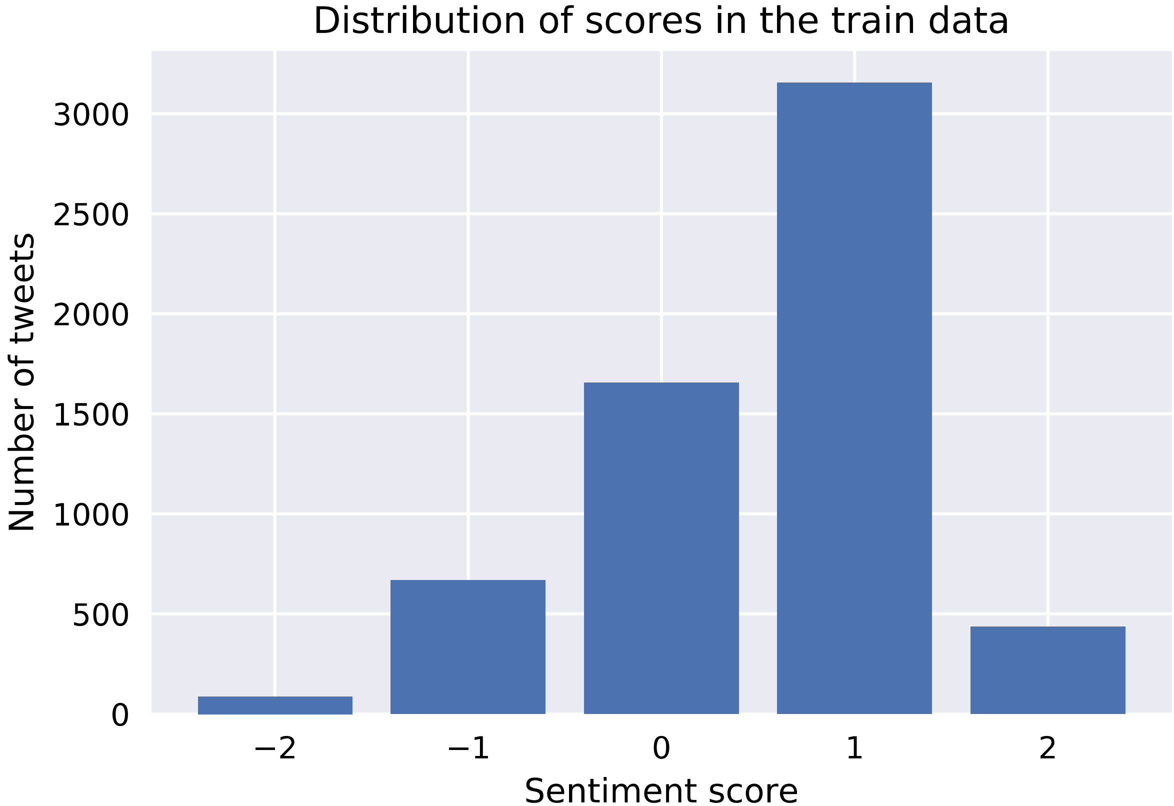


Outline

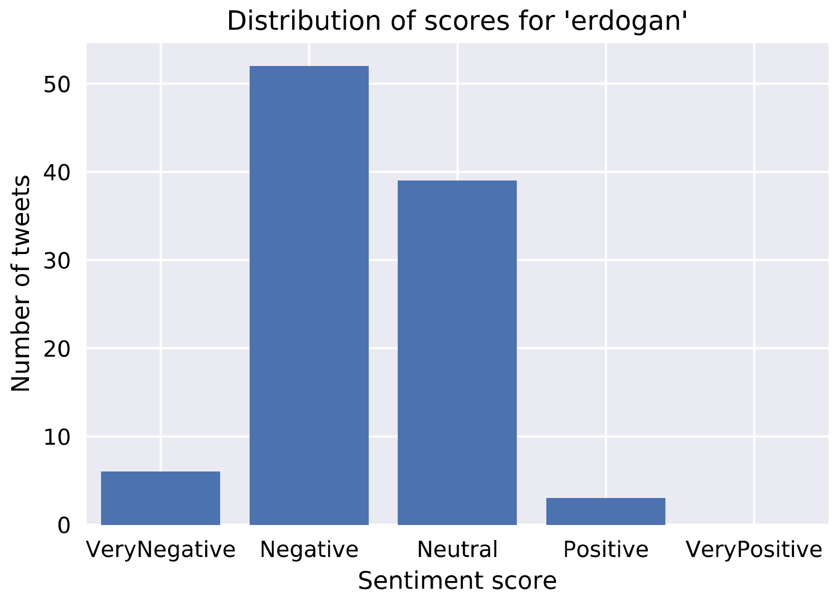
- The traditional approach
- An approach powered by word embeddings
- Quantification
- Research directions

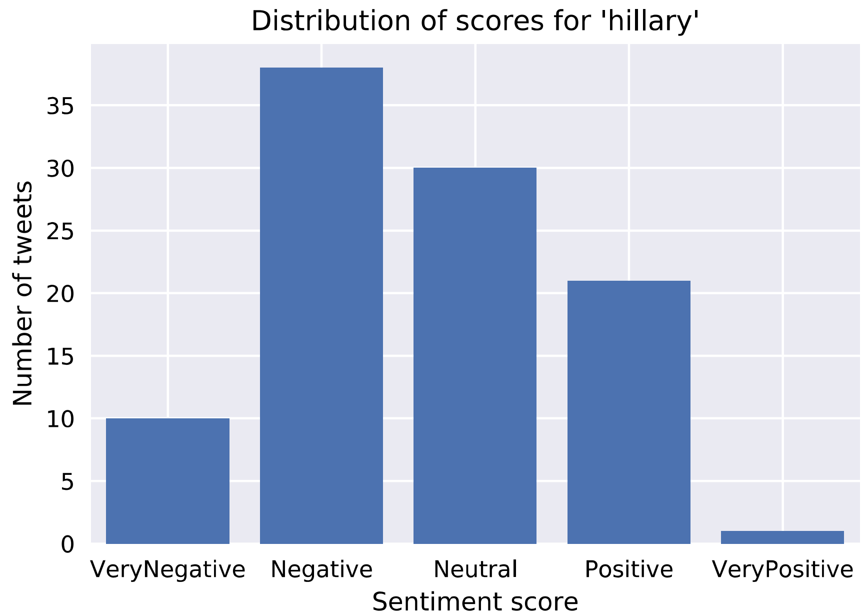
Our data

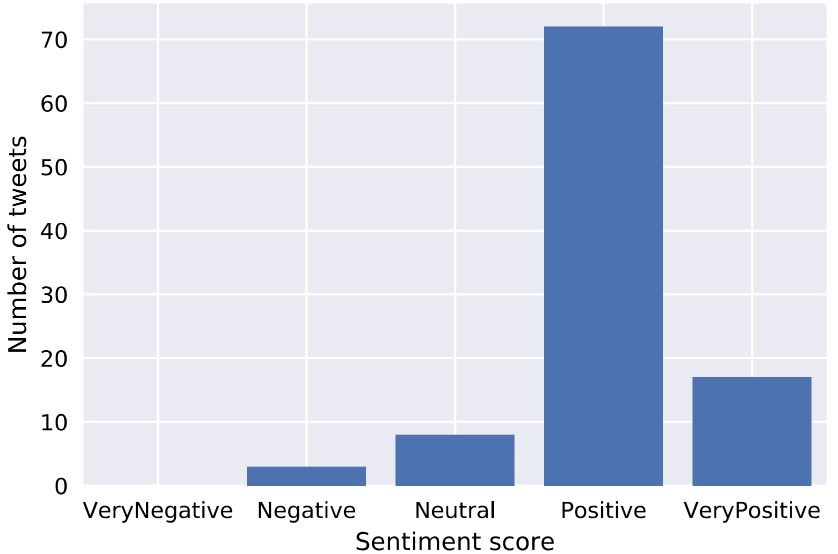
- SemEval 2017 (train: 6K, test: 2K)
- 5 classes

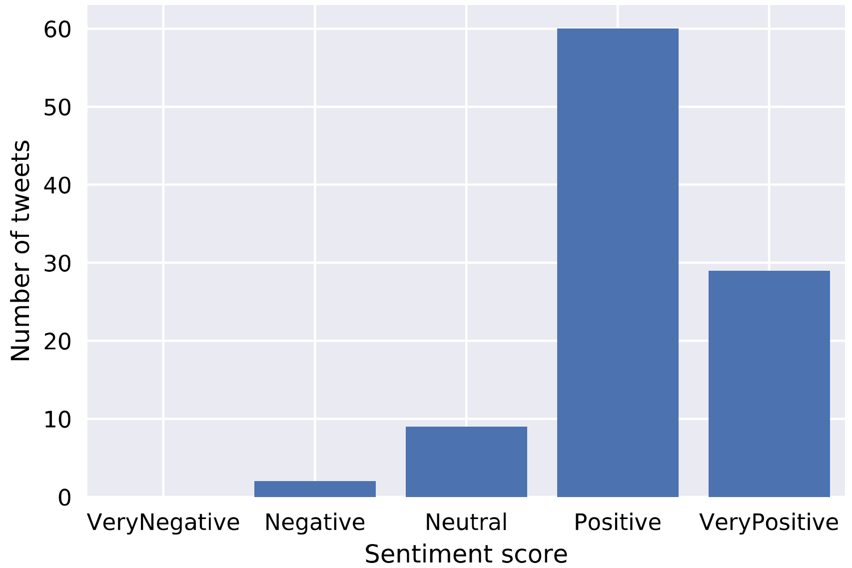


Our data (cont.)









Feature Extraction

- Tokenisation: split text in words
  - Do not split emoticons
  - Urls → <url>
  - Mentions @balikasg → <user>
- Normalization
  - 2morrow → tomorrow
- Vectorization
  - Term freq.
  - Term freq. Inverse document frequency
  - Hashing trick
  - On words or on n-character grams

"Thanks to @microsoft, I just may be switching over to @apple!! "

thanks, to, <user>, i, just, may, be, ...

Feature Engineering

- Counts of punctuation
  - !, ?, !+?
- Negative contexts
- “Not today.” → “Not today_NEG . ”
- POS tags
  - Verbs, Nouns, Adverbs,...
- Emoticons
  - Positive, Negative
- All caps
- Sentiment lexicons
  - Bing lius,..

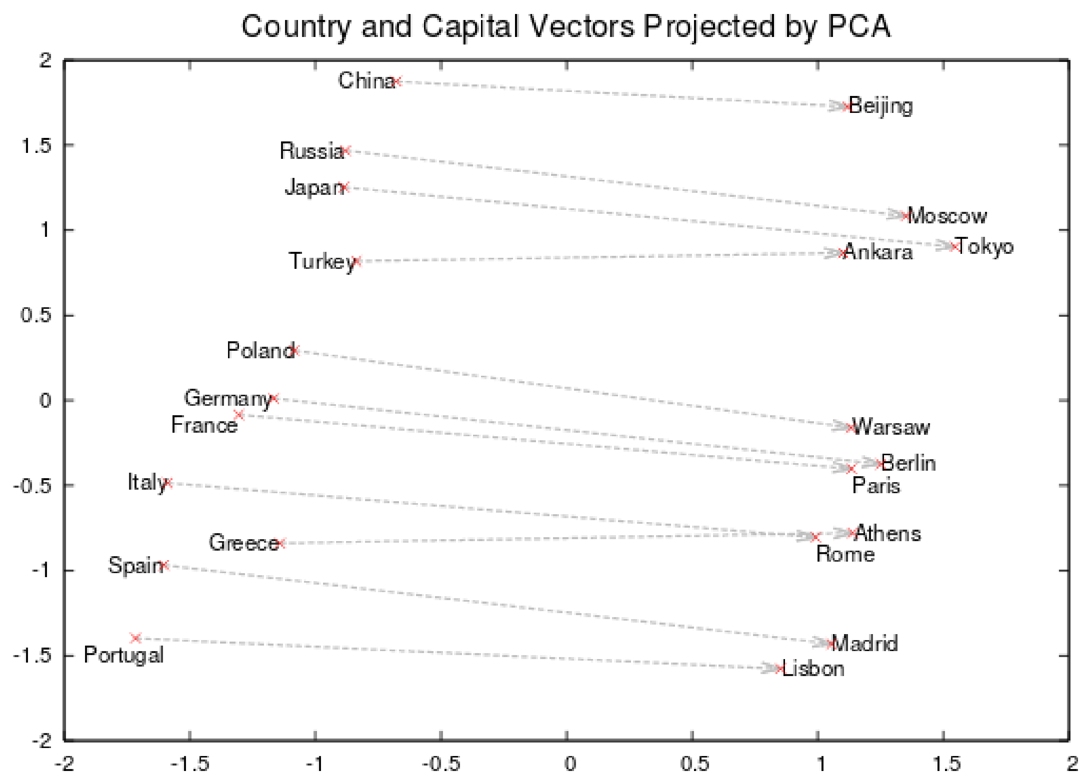
“Thanks to @microsoft, I just may be switching over to @apple!!”

thanks, to, <user>, i, just, may, be, ...

We refer to their concatenation as “special features” [jair14]

Word Embeddings

- Similar with skipgram + negative sampling
- Can handle OOV
- 294 languages



Model Selection

- Linear models
  - Logistic Regression
  - Support Vector Machines
- Trees
  - Random Forests
  - Gradient Boosted Trees
- Neural Nets
  - Recurrent Neural Networks
  - Convolutional NNs

Model Selection (cont.)

- Unbalanced problem
  - Dummy classifier may be competitive
- Tune for the measure we optimize
  - Grid search w stratified cross-validation
- Add weights to the classes
- Other strategies (e.g., sub/over-sampling)

Evaluation measures

- Macro-averaged f-measure
  - Harmonic mean of precision/recall
  - Does not penalize distance
- Macro-averaged absolute error
  - Distance matters

Results

- MaF_1
- Higher is better

Representation	Log. Reg	SVM
Freq. class baseline	0.127	0.127
tf	0.267	0.269
idf	0.260	0.262
tf+weights	0.288	0.292
idf+weights	0.319	0.303
cgrams-idf +weights	0.310	0.313
+special feat.	**0.336**	0.324
union: cgrams, ngrams	0.323	**0.327**

Results (cont.)

- MaF_1
- Higher is better

Representation	Log. Reg	SVM
Freq. class baseline	0.127	0.127
average	0.322	0.292
+weights	0.345	0.335
Only words w embeddings + weights	0.361	0.355
Best of previous slide	0.336	0.327
Union word embeddings + tfidf unigrams	**0.374**	**0.371**

Related work

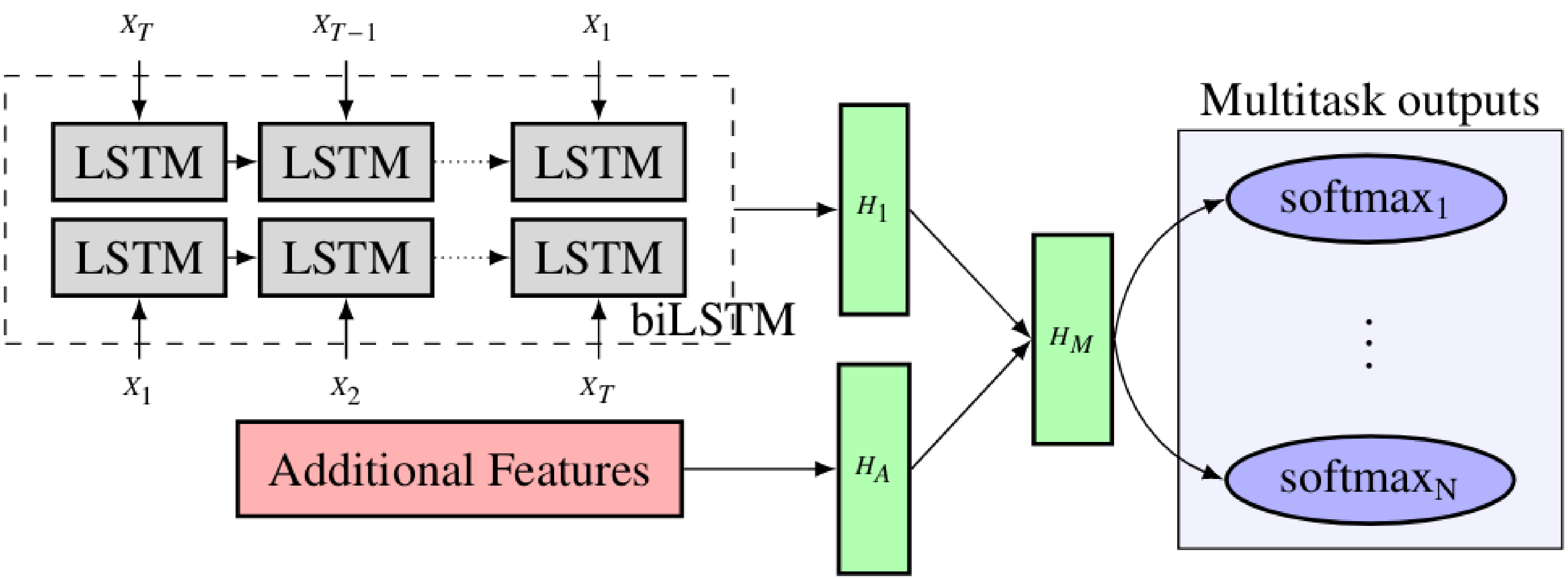
- State-of-the-art
- Research directions + quantification

BB_twtr @ SemEval2017

- Won every task
- 10 CNNs + 10LSTMs
- Distant supervision

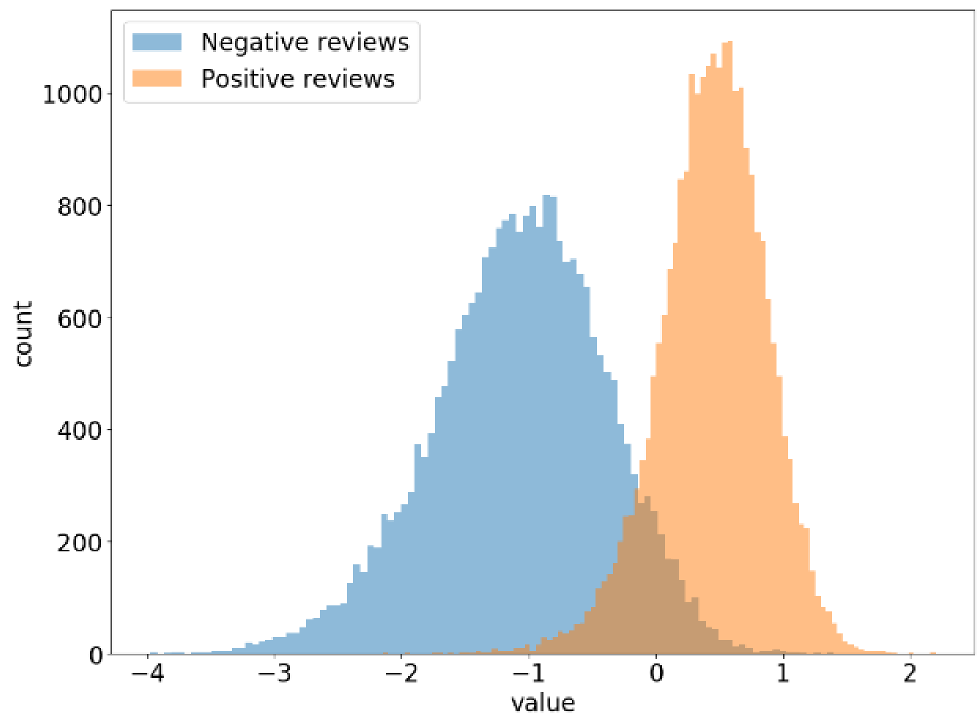
Multi-task learning [Sigir18]

- Correlation between domains/languages/problems



Language models [arxiv17]

- Unsupervised; model sequences
- A single neuron predictive of sentiment



Classification or Quantification?

- Classification: Given **tweet**, predict sentiment. Focus on each instance.
- Quantification: Given **tweets**, predict sentiment prevalence (~satisfaction studies). Focus on distribution.
  - iPhoneX: 20% Negative, 50% Neutral, 30% Positive

- A perfect quantifier may be a bad classifier but not vice versa.

Quantification [asonam15]

- Classify and Count (CC)
- Prob. Classify and Count (PCC)
- Adjusted CC/PCC
- Directly optimize KLD

References

- S. Kiritchenko, X. Zhu and S. Mohammad: Sentiment analysis of short informal texts. JAIR, 2014
- G. Balikas and MR Amini: TwiSE at SemEval-2016 Task 4: Twitter Sentiment Classification, SemEval, 2016
- G. Balikas, S. Moura, MR Amini: Multitask Learning for Fine-Grained Twitter Sentiment Analysis, SIGIR, 2016
- A Radford, R Jozefowicz, I Sutskever: Learning to generate reviews and discovering sentiment, Arxiv, 2017
- W Gao, F Sebastiani: Tweet sentiment: From classification to quantification, ASONAM, 2015
- S Rosenthal, N Farra, P Nakov: SemEval-2017 task 4: Sentiment analysis in Twitter

Thank you!

- Code + presentation @ github