Research Presentation

Translating Naive Bayes

Patrick Martin

February 12, 2018

Motivation 1

- Imagine we've got a lot of documents, only some of which are interesting
- But we don't know which ones!
- We can have a human look at them, but there are really too many documents to look at all of them
- Can we learn what documents are interesting under these constraints?
- Current answer: kinda

Motivation 1

- Imagine we've got a lot of documents, only some of which are interesting
- But we don't know which ones!
- We can have a human look at them, but there are really too many documents to look at all of them
- Can we learn what documents are interesting under these constraints?
- Current answer: kinda

The Data

- Reddit comments from January - March 2017 from French- and Spanish-speaking subreddits
- Our metric of “interesting” will be “controversiality”

	Spanish	French
# Comments	163,057	207,348
# Controversial	5,243	9,449

Bootstrapping

1. Pick 5,000 comments at random, determine which are controversial
2. Train classifiers on those comments, Naive Bayes (NB) and Logistic Regression (LR)
3. Take the classifier that does the best, and run it over all the documents
4. Pick the 5,000 documents deemed by the classifier to be most controversial, excluding any previously seen ones
5. Repeat

Results - Spanish

	# Docs	# Yes (repeats)	# Yes (total)	F-score	
				NB	LR
Init	5000	172 (3.4%)	172	0	**0.087**
Round 1	5000	399 (8.0%)	437 (+265)	0	**0.077**
Round 2	5000	574 (11.5%)	683 (+246)	0	**0.16**
Round 3	5000	710 (14.2%)	928 (+245)	0	**0.097**

Results - French

	# Docs	# Yes (repeats)	# Yes (total)	F-score	
				NB	LR
Init	5000	216 (4.3%)	216	0	**0.10**
Round 1	5000	571 (11.4%)	621 (+405)	0.017	**0.12**
Round 2	5000	883 (17.7%)	1034 (+413)	0	**0.17**
Round 3	5000	1124 (22.5%)	1432 (+398)	0.007	**0.16**

Naive Bayes

- The signal is too sparse for Naive Bayes to positively identify anything
- If we restrict the training data to be 33% controversial messages, will that help?
- Turns out, yes

Results - Spanish

	# Docs	# Yes (repeats)	# Yes (total)	F-score	
				NB	LR
Init	5000	173 (3.5%)	173	*0.16*	0.30
R1	5000	381 (7.6%)	429 (+256)	*0.36*	0.43
R2	5000	553 (11.1%)	702 (+273)	*0.31*	0.40
R3	5000	724 (14.5%)	1000 (+298)	*0.38*	0.47

Results - French

	# Docs	# Yes (repeats)	# Yes (total)	F-score	
				NB	LR
Init	5000	243 (4.9%)	243	*0.42*	0.49
R1	5000	533 (10.7%)	637 (+394)	*0.35*	0.44
R2	5000	828 (16.6%)	1095 (+458)	*0.36*	0.44
R3	5000	1041 (20.8%)	1523 (+428)	*0.38*	0.44

Conclusion 1

- Restricting the training data does help Naive Bayes get on the scoreboard, but Logistic Regression still wins out

Motivation 2

- Now imagine we've got a fairly good classifier on the Spanish comments, but we don't have anything for the French
- Can we somehow translate our Spanish classifier to French?

The Plan

- Assume our good classifier is Naive Bayes with word tokenization Ours will be about 30% accurate
- Take 500 words that strongly indicate either ‘yes’ or ‘no’ to controversiality in this classifier 250 from each side
- Translate those words to French Google Translate for now...
- Build a new classifier from these words <— This is the hard part
- Evaluate the quality

The Most Controversial Words

Spanish	English	French
allende	Allende	allende
retorno	return	revenir
pib	?	pib
ingrediente	ingredient	ingrédient
imbecil	imbecile	imbécile
⋮	⋮	⋮
03	03	03
data	data	données
vayas	go	aller
súper	super	super
sorry	sorry	désolé

Building a Classifier

- Naive Bayes needs to know $p(word|class)$ for each word.
- For the words we translated, this is pretty easy Just average the probabilities if the same word shows up multiple times
- But we can't make a classifier with only 500 words

Inferring Probabilities

- As mentioned before, this is the potentially interesting part
- First attempt will be to get word similarities from Word2Vec on all the French comments, and

$$p(class|word) \approx \sum_{w \ N_k(word)} p(class|w) sim(word, w)$$

  where $N_k(word)$ represents the k-nearest "anchor" words
- For now, $k = 5$
- Use this to estimate probabilities for all words that occur at least 100 times (8,554)

Result - Translated Model

	# Docs	# Yes (repeats)	# Yes (total)	F-score	
				NB	LR
Init	5000	213 (4.3%)	213	0	**0.05**
R1	5000	537 (10.7%)	599 (+386)	0.03	**0.09**
R2	5000	803 (16.1%)	979 (+380)	0	**0.14**
R3	5000	1069 (21.4%)	1383 (+404)	0.008	**0.17**

Initial Conclusions and Next Steps

- It doesn't *not* work
- Could improve translations:
  - Including the unaccented form of words when translating
  - Preserving English and other foreign words
- Could improve inference
  - Different corpus for Word2Vec?
  - Different formula for inference