Disparities in the Performance of Natural Language Processing Tools

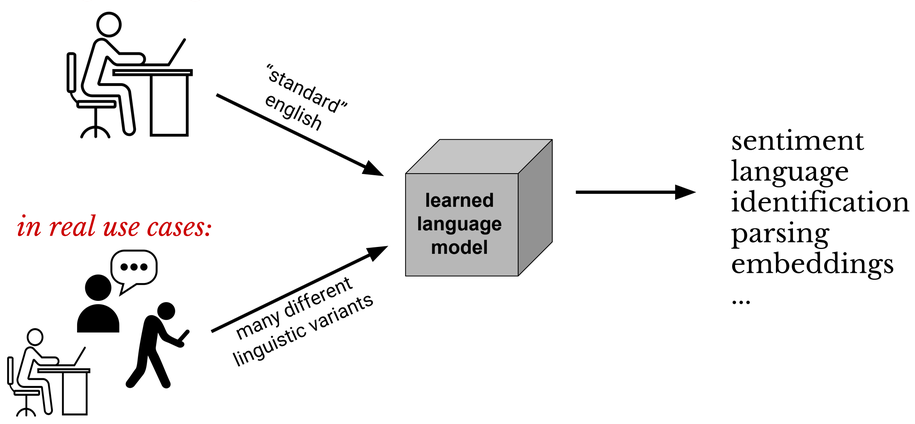
exploring and quantifying the problem

Harini Suresh, Divya Shanmugam, John Guttag

The Problem

Widely used NLP tools are applied to distributions that are quite different than those used in training.

during training:



negative impact of testing on different domains:

Tools trained on **formal english text** (like news articles) can have a **negative social impact** for users who communicate with different dialects.

- underexposure of language from minority groups
- bias in viewed and accessed information
- demographic misrepresentation
- exclusion of specific groups

for example:

- Youtube auto-captioning has a higher error rate for women
- Language identifiers have lower accuracy on tweets from African-American users
- Speech recognizers struggle with minority dialects

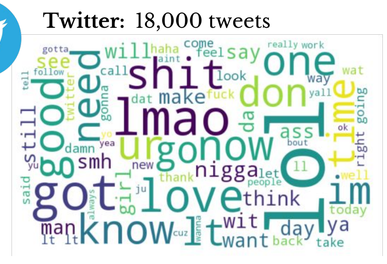
Approach

investigate the consequences of using off-the-shelf NLP tools trained primarily on structured, formal english for classifying messy text containing:

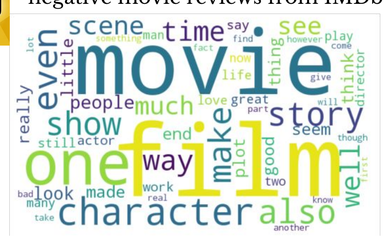
- slang ("whatup")
- abbreviations ("lmao")
- different english dialects ("i'm finna go")

Datasets

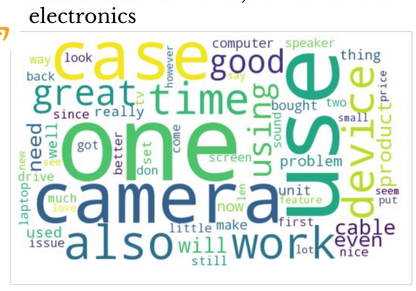
Twitter: 18,000 tweets



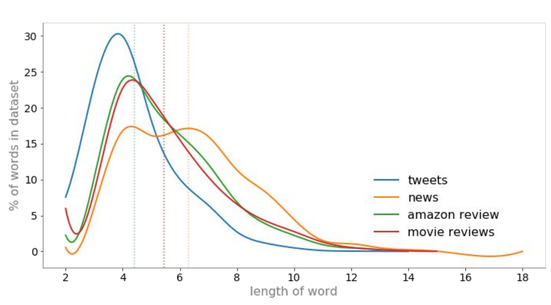
Movie Reviews: 18,000 positive and negative movie reviews from IMDb

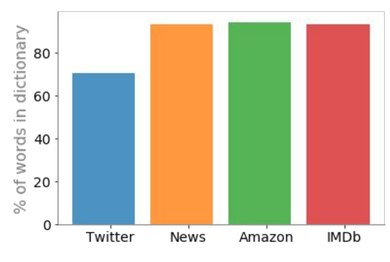


Amazon Reviews: 18,000 reviews of electronics



News: 18,000 articles from Wall Street Journal and Washington Post





Example: Language Identification with langid.py

what is type of english is actually recognized?

		Twitter	Amazon	IMDb	News
langid.py classifier error rate	**max length of 140 characters**	10.5%	0.16%	0.08%	0.06%
	total length$_1$	1.2%	0.14%	0.0%	0.0%
	average total length (# characters)	801.9	607.6	1277.4	1269.9

1 for Twitter, a longer sample is obtained by aggregating all tweets for a user

task: use *langid.py*, an off-the-shelf language identification tool, to predict whether text is in English

how is it trained?

- JRC-Acquis (mostly legal documents)
- ClueWeb 09 (webpages)
- Wikipedia
- Reuters RCV2 (news)
- Debian i18n (software documentation)

takeaways:

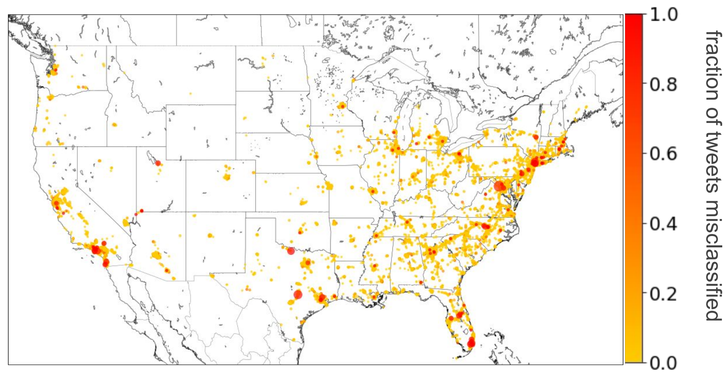
- performance on tweets is much worse than other mediums
- increasing tweet length by aggregating all tweets from one user reduces error from 10.5% to 1.2%
- however, there remains a large disparity in performance among short texts

what's going wrong with the tweets?

some misclassified tweets with their associated probabilities:

same sht diff day u kno, u been doin ite doe? How skool n sht? → Afrikaans (p = 0.99)
hmm wat made yo night so damn good → Breton (p = 0.42)
wat bottle u finna pop? → Maltese (p = 0.99)
lmao we gotta take turns → Finnish (p = 0.96)
we gotta make sumtin happen → Finnish (p = 0.99)

misclassified tweets are more prevalent in urban areas:



size of dot indicates number of tweets

- the performance disparity is not uniform across all tweets
- there is a bias against the type of language used mostly in urban areas

Example: Sentiment Analysis with NLTK Vader

how well is sentiment detected in different types of text?

	Twitter	Amazon	IMDb
NLTK VADER sentiment classification accuracy	59.0%	86.0%	70.3%

task: use *NLTK Vader*, an off-the-shelf sentiment analysis tool (trained on words from established treebanks, to classify positive vs negative sentiment in text

takeaways:

- transferring knowledge of sentiment is subject to differences in domain and structure
- more analysis is needed to quantify effect of changing domains

Current Work

bias quantification:

- quantify the generalizability of word embeddings across domains
  - do word embeddings trained with specific modalities exhibit distinct biases in the word similarities they learn?
- analyze tasks outside of sentiment analysis such as parse tree generation and topic sensing
- further examine how geography and socioeconomic status correlates with disparity in performance of NLP tools

bias correction:

- work towards models that utilize large corpuses of text, but *also* adapt well to specific domains
  - multitask learning, transfer learning
- develop a defined methodology for quantifying whether text is suitable for use with a given off-the-shelf tool
- before analysis, map text to a universal space that preserves semantic meaning but masks language choices that could lead to discrimination