Classifying Tweets Based on Climate Change Stance

Motivation

Climate change has become an increasingly polarizing topic in popular and political media. Using Twitter, we seek to develop a classifier that can discriminate between text that shows belief vs. disbelief in human-caused climate change. We use semisupervised classification with Quasi-Newton Semi-SVM implementation (QN-S3VM) and Multinomial Naive Bayes with Expectation Maximization (MNB-EM) into positive, neutral, and negative categories. Our results show that semisupervised learning does not offer significant improvements over supervised learning.

Data Set

- Three categories: positive, neutral, negative belief of human-caused climate change
- Manually labeled training/ validation/test set of ~14K tweets
- Unlabeled set of ~38K tweets related to climate change/global warming collected using TweePy
- Data split into 70/20/10 train/test/val

Tweet	Class
"climate change, a factor in texas floods, largely ignored"	1
"global warming on mars...."	0
"scientists were attempting the same global warming scam 60 years ago"	-1

Methods

Supervised Learning

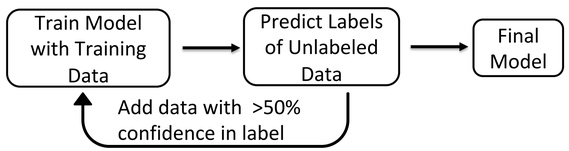
Multinomial Naive Bayes

$$L = \prod_{i=1}^{n}\left(\prod_{j=1}^{d_i} p\left(x_j^{(i)}\middle|y^{(i)};\phi_{k|y}\right)\right)p(y^{(i)};\phi_y)$$

- Unigram: $x_j^{(i)}$ is each word
- Bigram: $x_j^{(i)} = (z_j^{(i)}, z_{j+1}^{(i)})$ pair of words
- Assume independence of all words/pairs within and between each example

Semisupervised Learning

Self Training



QN-S3VM

- Maximum-margin classification algorithm
- Find the optimal **y** for the unlabeled data

$$\underset{f\in\mathcal{H},\mathbf{y}\in\{-1,+1\}^u}{\text{minimize}} \frac{1}{l}\sum_{i=1}^{l}\mathcal{L}(y_i', f(\mathbf{x}_i)) + \lambda'\frac{1}{u}\sum_{i=1}^{u}\mathcal{L}(y_i, f(\mathbf{x}_{l+i})) + \lambda||f||_{\mathcal{H}}^2$$

MNB-EM

$$L = \left(\sum_{i=1}^{n}\left(\sum_{j=1}^{d_i}\log p\left(x_j^{(i)}\middle|y^{(i)};\phi_{k|y}\right)\right) + \log p(y^{(i)};\phi_y)\right) * \alpha\left(\sum_{i=1}^{k}\log\left(\sum_{y^{(l)}} Q_l(y^{(l)})\frac{\prod_{m=1}^{d_l} p\left(x_m^{(l)}\middle|y^{(l)};\phi_{k|y}\right)p(y^{(i)};\phi_y)}{Q_l(y^{(l)})}\right)\right)$$

- E-step:

$$Q_l(y^{(l)}) = \frac{\prod_{m=1}^{d_l} p\left(x_m^{(l)}\middle|y^{(l)};\phi_{k|y}\right)p(y^{(i)};\phi_y)}{\sum_{y^{(l)}}\prod_{m=1}^{d_l} p\left(x_m^{(l)}\middle|y^{(l)};\phi_{k|y}\right)p(y^{(i)};\phi_y)}$$

- M-step:

$$\phi_{k|y} = \frac{1+\alpha\sum_{i=1}^{n}\sum_{j=1}^{d_i}1(x_j^{(i)}=k\wedge y^{(i)}=c)+\sum_{l=1}^{k}Q_l(y^{(l)})\sum_{m=1}^{d_l}1(x_m^{(l)}=k)}{v+\alpha\sum_{i=1}^{n}\sum_{j=1}^{d_i}1(y^{(i)}=c)+\sum_{l=1}^{k}Q_l(y^{(l)})\sum_{m=1}^{d_l}1}$$

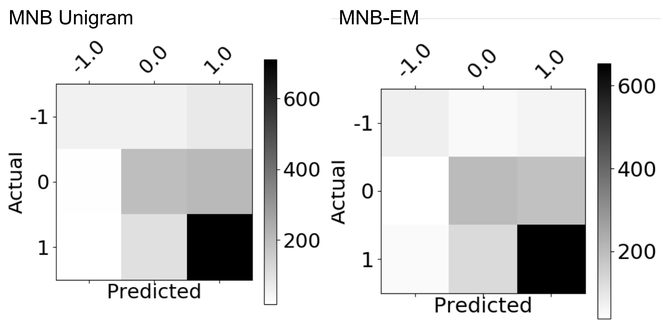
$$\phi_y = \frac{1+\alpha\sum_{i=1}^{n}1(y^{(i)}=c)+\sum_{l=1}^{k}Q_l(y^{(l)})}{C+\alpha n+k}$$

Results

Comparing Algorithms

Model	Training Acc	Validation Acc	Test Acc
MNB Unigram	76.8	67.6	66.4
MNB Bigram	72.8	62.5	60.6
Naive Bayes Self-Training	66.7	60.9	61.6
QN-S3VM	59.2	59.0	55.0
MNB-EM	81.9	62.5	63.9

Accuracy Per Class



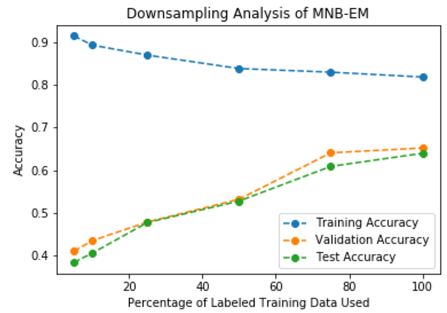
Model Tuning

Kernel	Training Acc	Validation Acc
RBF	59.2	59.0
Linear	52.9	49.2

Alpha	Training Acc	Validation Acc
1	67.3	60.7
5	76.8	62.3
20	81.9	65.2
50	82.6	65.1
100	82.9	64.5

Data Downsampling

Question: how much would unlabeled data help if we had less labeled data?



Conclusions

Conclusions

- Unigram Multinomial Naive Bayes model performs best
- Semisupervised learning with NB performs well, likely since underlying NB model fits data well
- More data would improve predictions

In future

- Incorporate more recently labeled data
- Attempt classification with DL models