Public Education Update via Reddit Post Classification

Wayne Chan

Agenda

- Executive Summary
- Methodology
- Findings
- Conclusion

Executive Summary

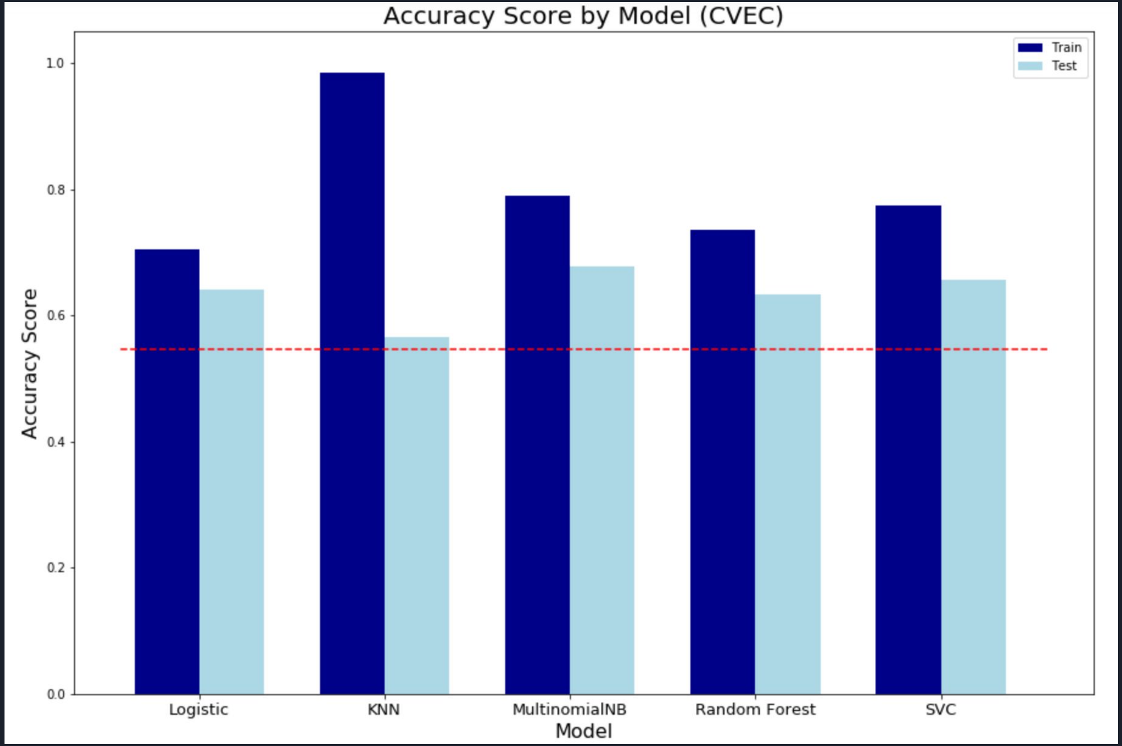
- With the continued disruptions of this educational year, we have an opportunity to augment or supplement the curriculum
- Reddit is a popular forum to go to for questions
  - NoStupidQuestions and TooAfraidToAsk are the two main subreddits where users go to ask questions without fear of being made fun of or embarrassed with TooAfraidToAsk being the more “serious” of the two
- Is it possible discern what topics should be added to curriculum by creating a model that separates posts between the two subreddits?

- The model was not able separate posts with a high degree of accuracy:
  - However, the weakness of the model is that it over rejects
  - Therefore, there is confidence in the topics that remain
  - These topics are those relating to health and relationships

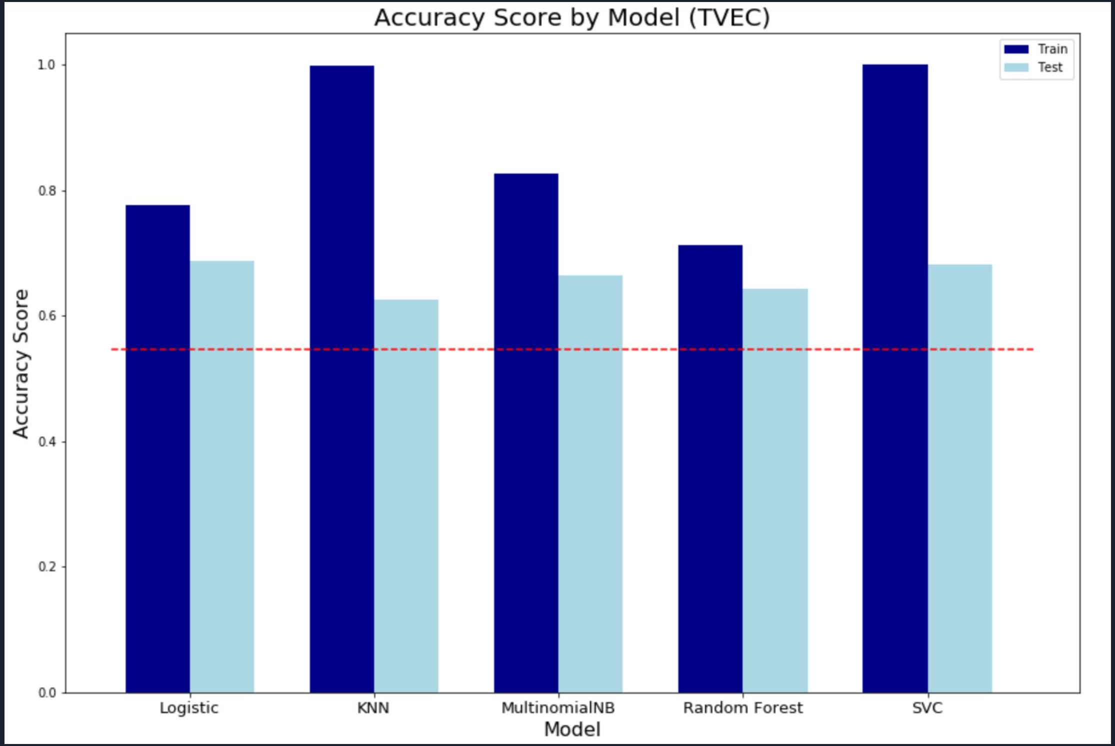
Methodology

1. Data acquisition
   - Used a pre-existing API to bring in posts from both subreddits
   - Check for deleted or removed posts
   - Check count of remaining posts and acquire additional if needed
   - Replaced blank text with post title
2. Setup Model
   - Initial model selections
   - Stemming
   - Searching for optimal parameters for models
   - Add new field combining post title and post body text
3. Score models and compare results

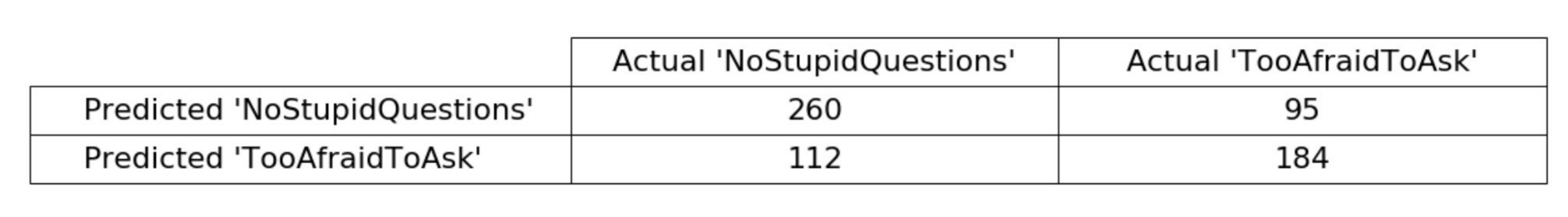
Findings - Score (Count Vectorizer)



Findings - Score (TF-IDF Vectorizer)

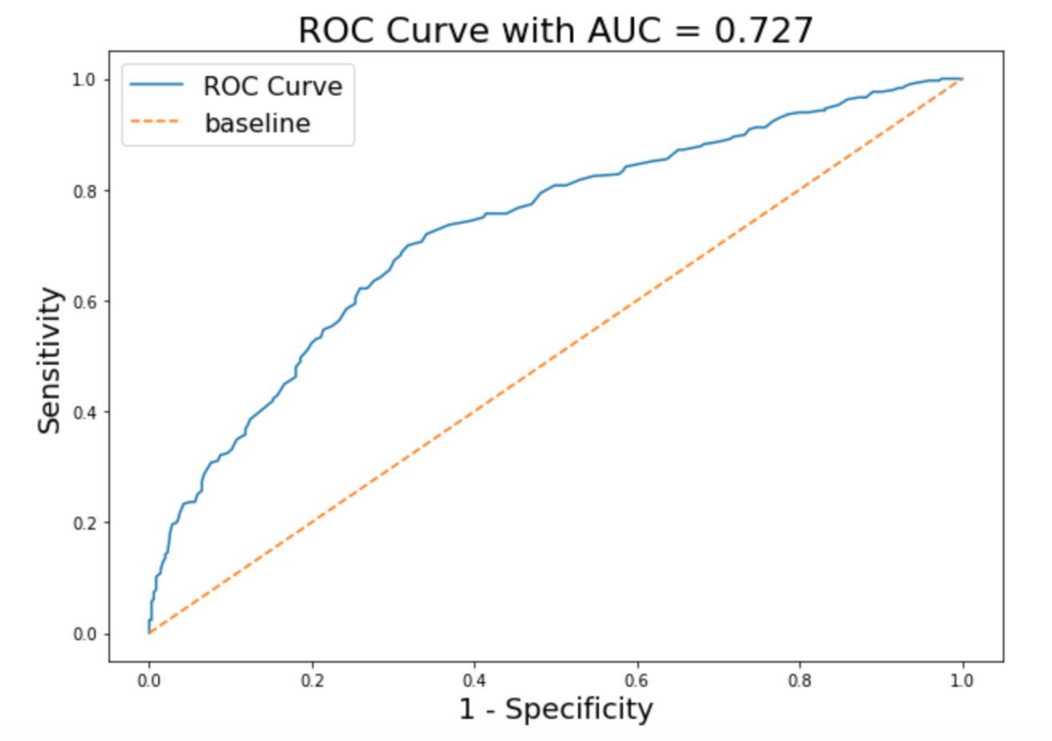


Findings - Model Predictions vs Actual



	Actual 'NoStupidQuestions'	Actual 'TooAfraidToAsk'
Predicted 'NoStupidQuestions'	260	95
Predicted 'TooAfraidToAsk'	112	184

Findings - ROC and AUC



Findings - Top Scoring Features

- ‘immedi’
- ‘been’
- ‘family’
- ‘surgeri’
- ‘mobil’
- ‘oper’
- ‘episod’
- ‘patient’
- ‘marri’
- ‘wife’
- ‘seriou’
- ‘date’
- ‘doctor’
- ‘certain’
- ‘anyway’
- ‘either’
- ‘sorri’
- ‘hospit’

Conclusion

- The model's predictive accuracy is 68%

What does this mean?

- We are not able to introduce posts into the model and classify between the two subreddits with high confidence because the model is overly sensitive
- Main topics of classified posts are relating to **health/medical** and **family/relationships**
- Rather than feed posts/topics into the model to guess whether it should be added or not, we can use the identified topics as areas to improve
- Reasonable initiatives that can be undertaken include:
  - expansion and promotion of TeleHealth (perhaps an online version)
  - relationship helpline and counselling (like 7cups, theSpark, or ginger.io)
  - creation/support/expansion of family health conversations (as per study from Nursing Research & Practice journal)[1]
  - directly adding to curriculum coping methods and tactics for discussion of sensitive matters

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3995177/

Questions