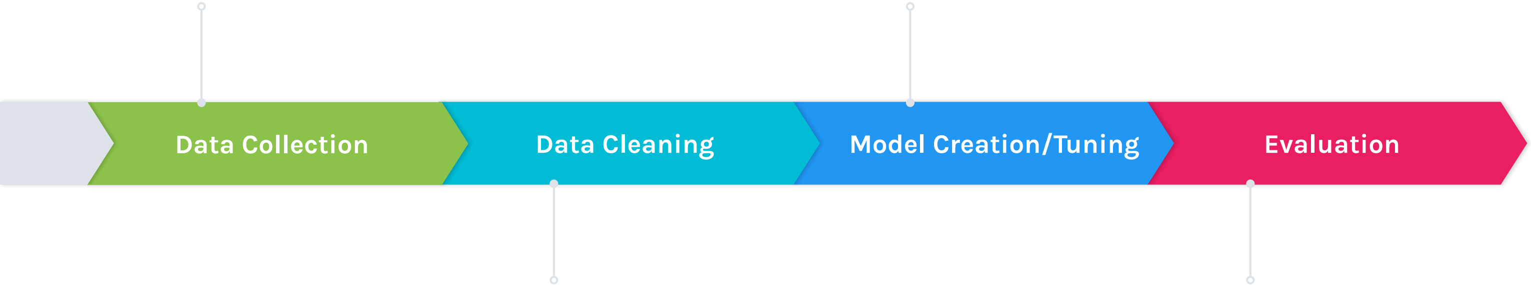
SMALLCO KNOWLEDGE BASE UPGRADE

OBJECTIVE:

- Knowledge base consists of many disconnected documents with no relationship to each other.
- Create a system that automatically makes label recommendations based document contents.

Approach



Data Collection

Data collected from Reddit was used to demonstrate the ability to effectively classify text

Data Cleaning

The collected text was scrubbed of HTML content and verified to be unique content for all samples

Model Creation/Tuning

Sample post text was broken down by words, and classification models were built around word frequencies

Evaluation

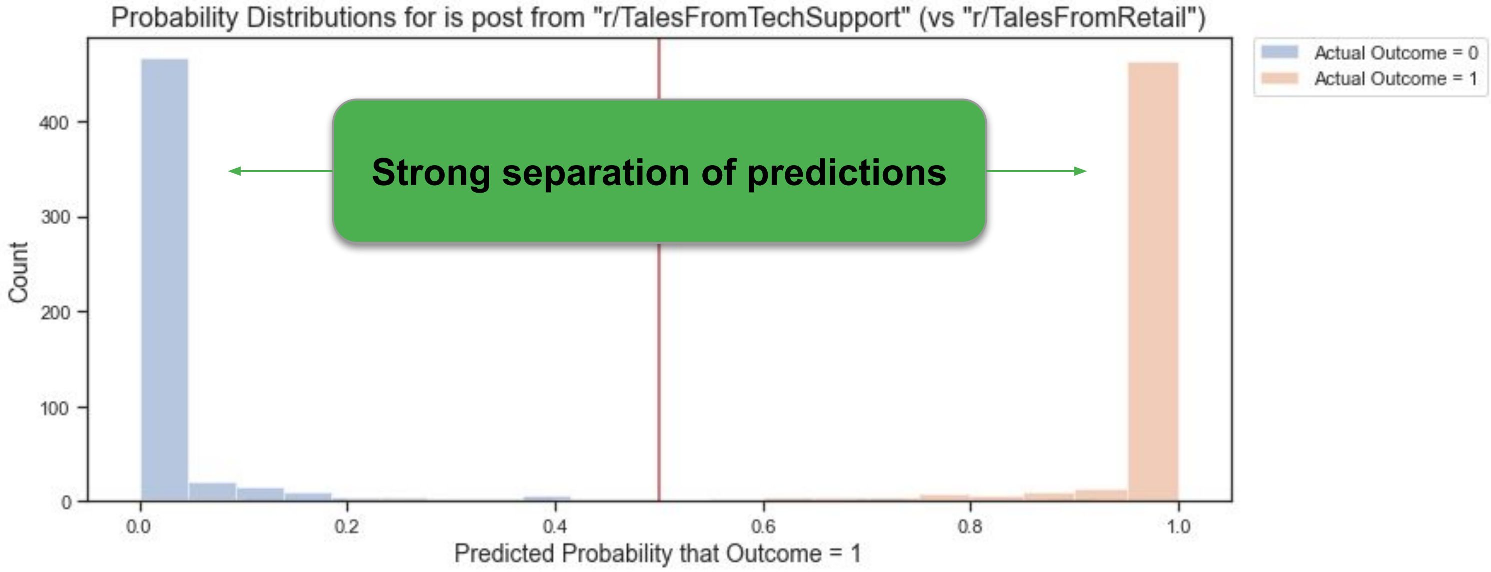
Models were compared by accuracy and std deviation, and the best model was replicated for a third data set for comparison

How did the model do?

- With **99% confidence**, we were able to correctly attribute the source of the content for **95% to 97%** of the test samples

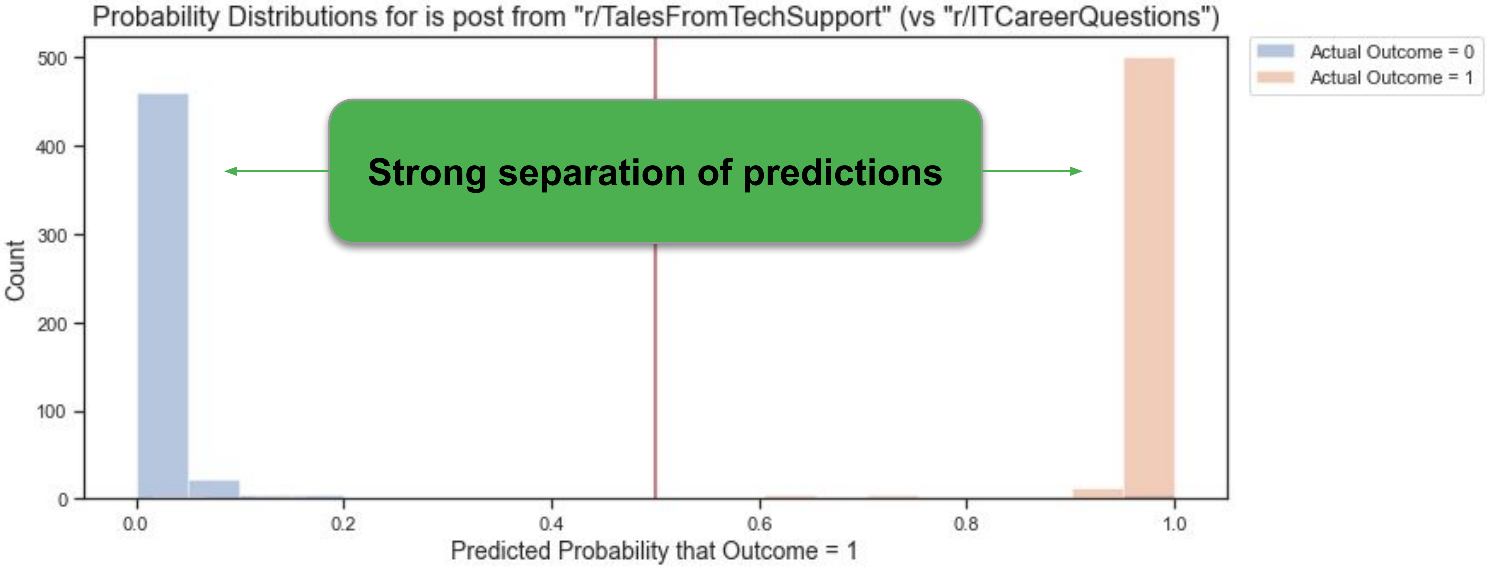
- On our particular set of test data, we saw:
  - Overall accuracy of **95.3%**
  - A false-negative rate of **6%**
  - A false-positive rate of **3%**

What does this look like?



Strong separation of predictions

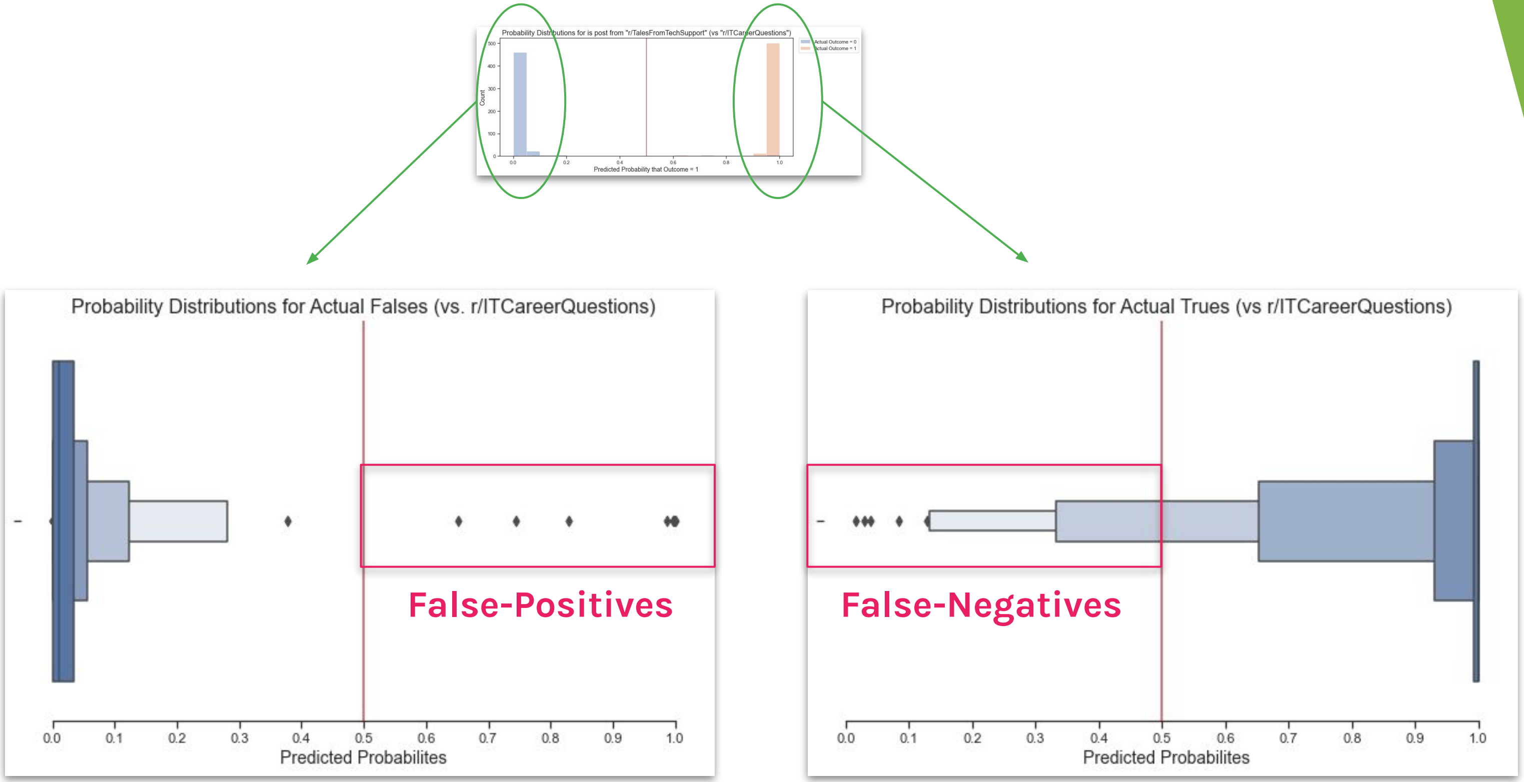
Does it work with other data?



Probability Distributions for is post from "r/TalesFromTechSupport" (vs "r/ITCareerQuestions")

Strong separation of predictions

What does this look like?



False-Positives

False-Negatives

A Few Points of Interest:

- False-positives tended to have an average post length nearly 2X longer than the average of the total data set

- False-negatives tended to have an average post length of about 1/2 the average length of the total data set

Conclusions:

- We know we can classify text origin with high accuracy

- Next Steps: Generalize Approach

Recommendations:

-

THANKS!

Any questions?

CREDITS

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival
- Photographs by Death to the Stock Photo (license)