Cluster 9: Machine Learning and Natural Language Processing, COSMOS, University of California, Santa Cruz

COSMOS UC Santa Cruz
California State Summer School for Mathematics and Science

Twitter Categorization

Richard Chen, Kevin Wang, Edmund Xin

Problem: Is a tweet related to a disaster event?

UNIVERSITY OF CALIFORNIA SANTA CRUZ

Abstract

Twitter provides a platform for human expression and discussion. The creation of a NLP classification model may be able to categorize the relevance of tweets. We developed multiple classification models, including a Neural Network, Logistic Regression, and Naive Bayes Classifier. Cross-validation of the model brought us over 80% accuracy on test data.

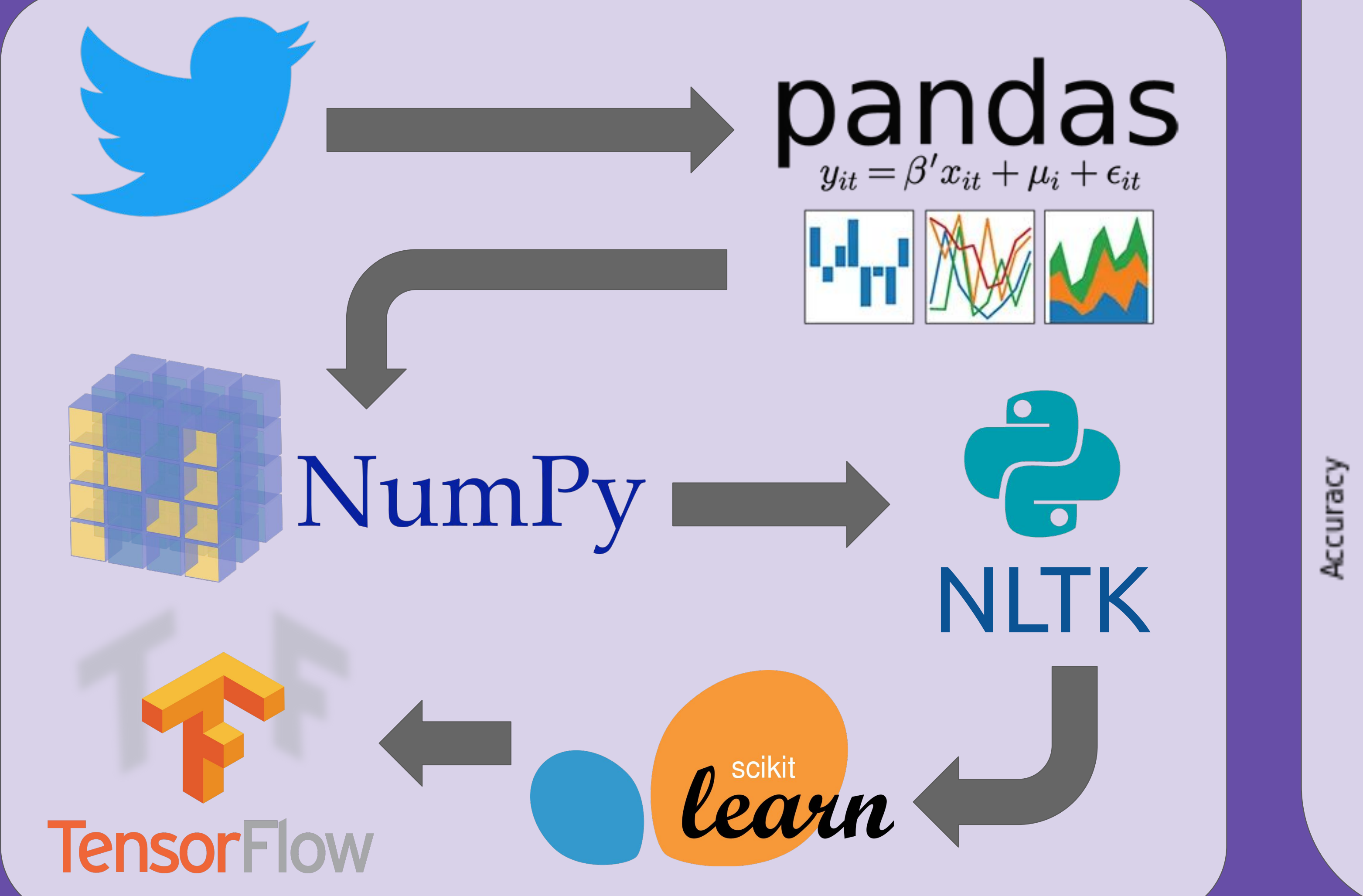
Data

Training	Testing
7416 Instances	1854 Instances

- # of classes: 2
  - Revelant = 1, Not relevant = 0
- Imbalanced training data: 4,305 (0) vs. 3,111 (1)

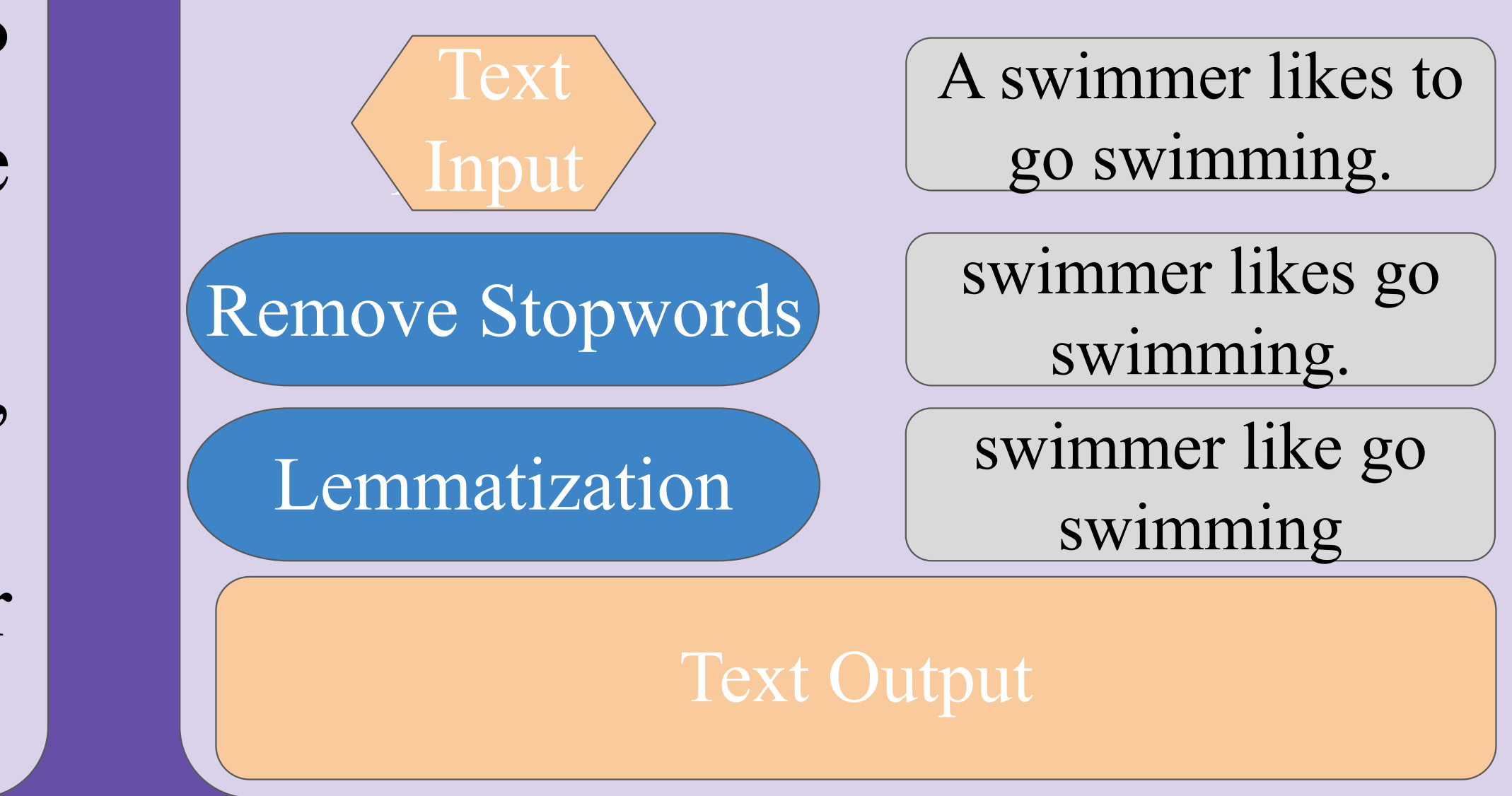
	index	class_label	text
0	8525	0	she keep it wet like tsunami
1	5008	1	when ur friend and u are talking about forest
2	8803	0	but i will be uploading these videos asap so y...
3	6795	0	i'm interested is it through yahoo?

Flowchart



Process

Preprocessing



TF-IDF Vectorization

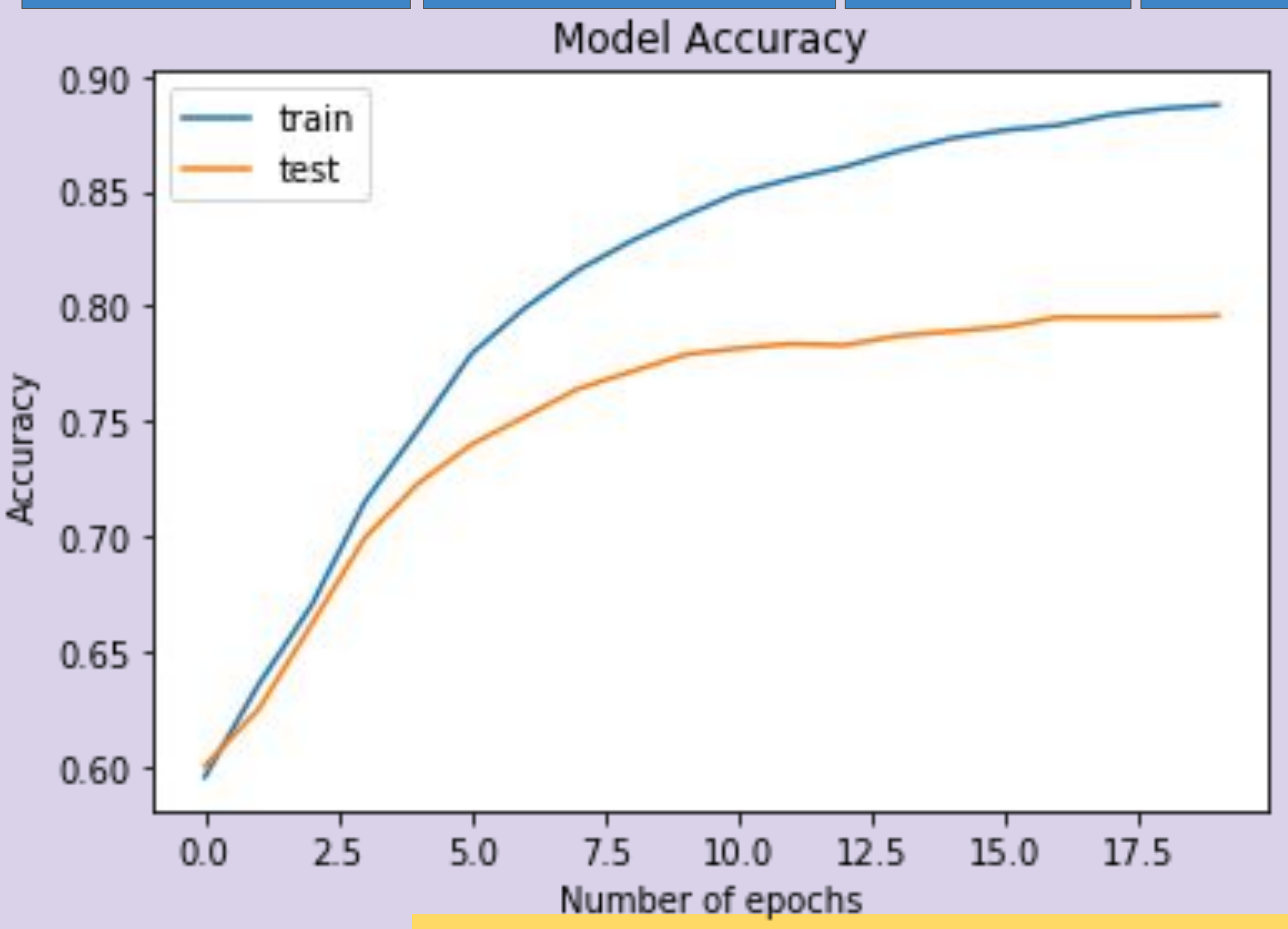
$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

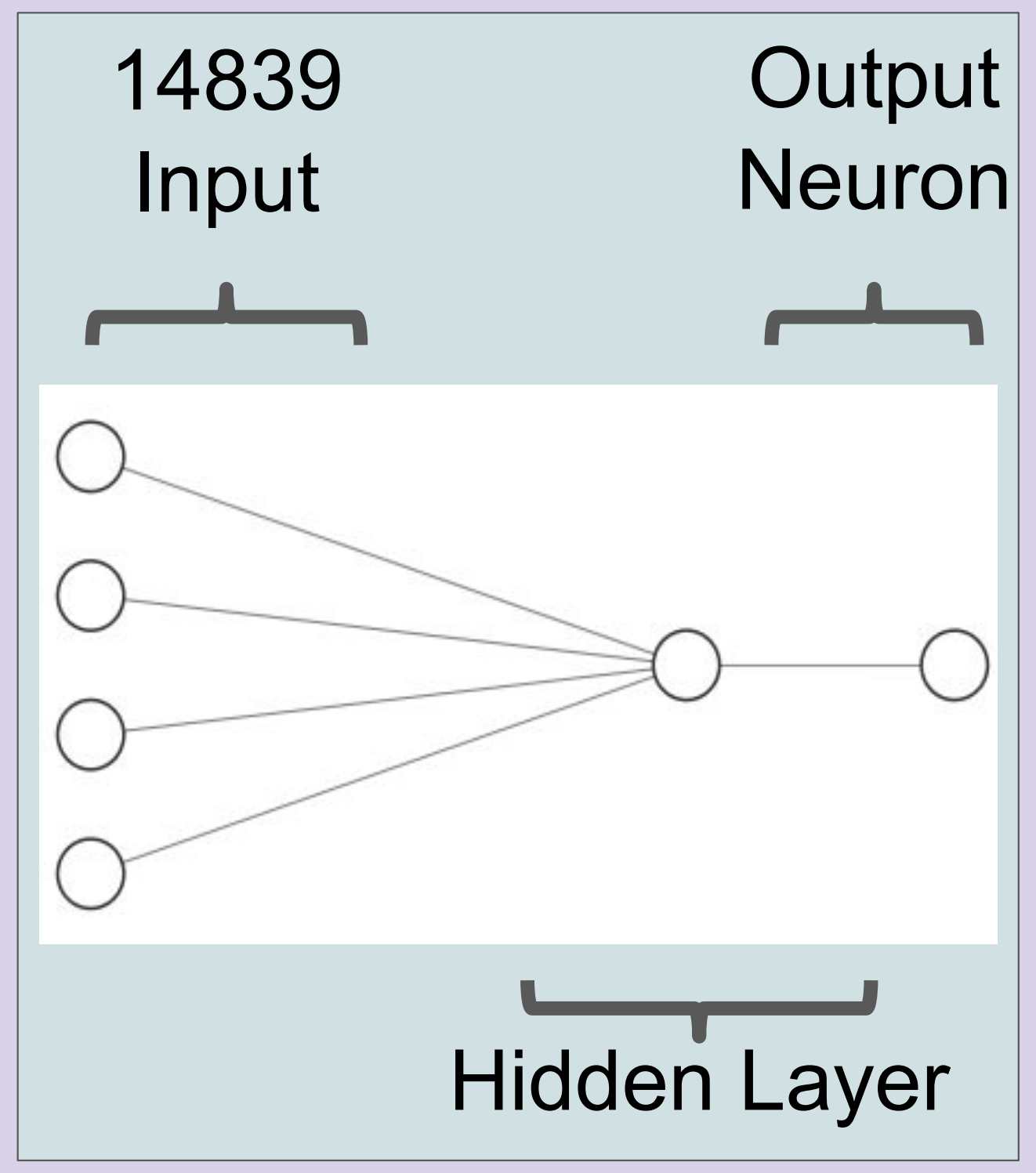
	The	TFIDF	Vectorization	Process	Is	Beautiful	Concept
Doc 1	1.90	1.30	0.84	1	2	0	0
Doc 2	1.92	1.39	0	0	2.06	0	1
Doc 3	1.97	1	0	0	1.97	0	0
Doc 4	2.02	0.95	1	0	2.17	0	0
Doc 5	1.84	0.30	0	0	1.90	0	0

Model

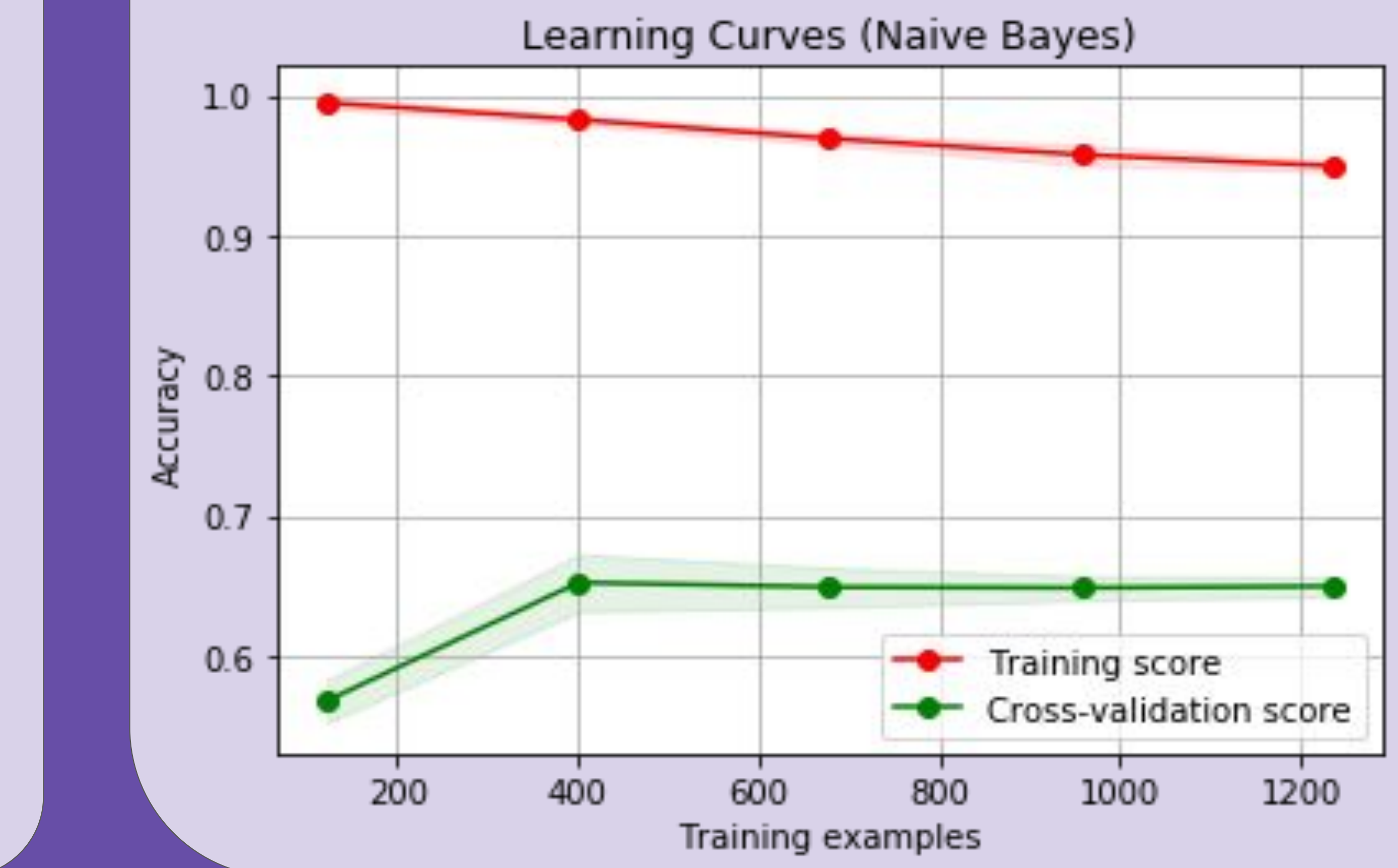
Artificial Neural Network

1 Define Network → 2 Compile Network → 3 Fit Network → 4 Evaluate Network → 5 Make Predictions





Bernoulli Naive Bayes Classifier



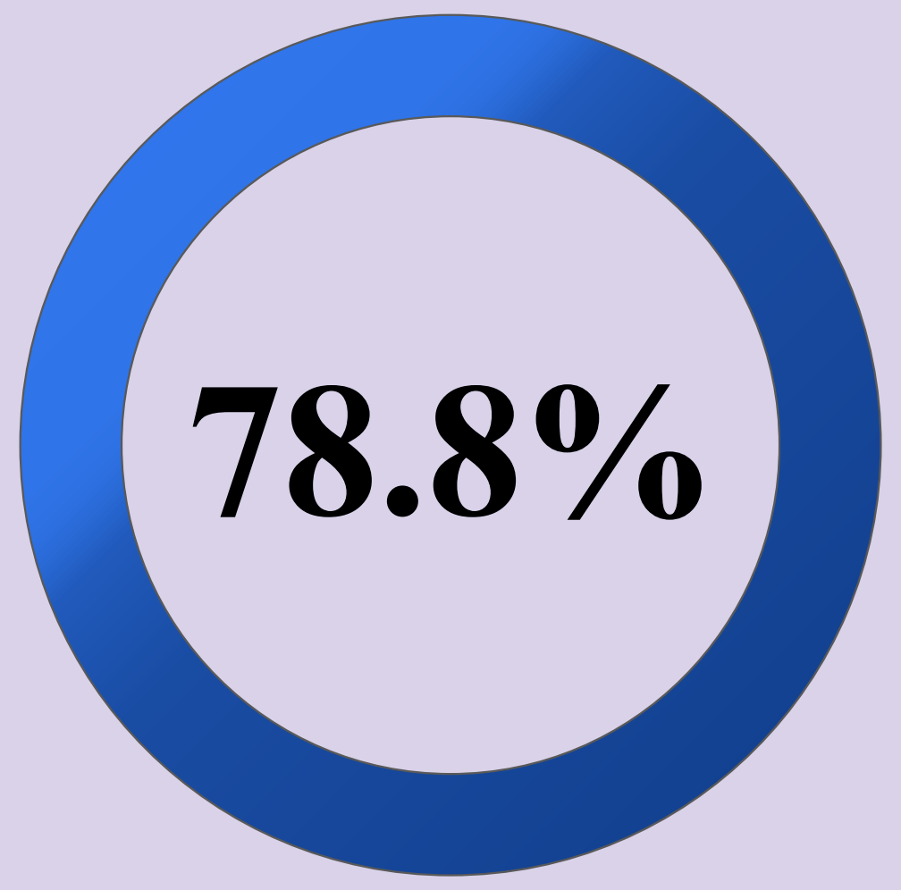
Naive Bayes Equation

$$y = \underset{c_i}{argmax}\ P(c_i) \prod_{j=1}^{n} P(x_j|c_i)$$

- Bernoulli Naive Bayes applies a Bernoulli distribution to the Naive Bayes classifier
- Each feature is a boolean outcome

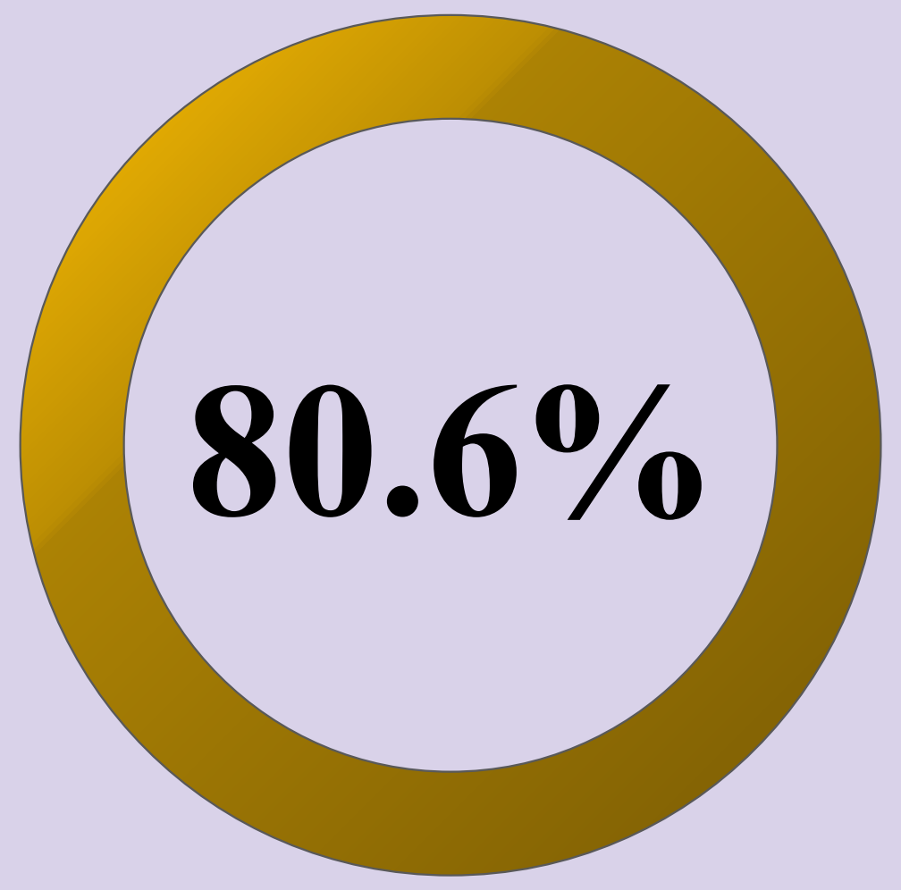
$$p(\mathbf{x} \mid C_k) = \prod_{i=1}^{n} p_{ki}^{x_i} (1 - p_{ki})^{(1-x_i)}$$

Results



Artificial Neural Network

- 62.8% Recall
- 81.5% Precision
- 0.70 F1 Score



Bernoulli Naive Bayes

- 78.1% Recall
- 81.8% Precision
- 0.78 F1 Score

Conclusion

- Bernoulli Naive Bayes achieved the highest accuracy of ~80.6%
- A neural network achieved an accuracy of ~78.8%
- Preprocessing is important for cleaning the data
- Nuances in individual texting habits make tweets inherently difficult to classify
- Tf-idf vectorization is better than countVectorizer as it values a word's importance to a document within a collection while countVectorizer simply performs frequency analysis

Acknowledgements

Thanks to Professors Narges Norouzi and Snigdha Chaturvedi for their patience and well-designed lectures. Thanks to their TAs as well as teacher fellow Matthew Rutledge for additional help both in understanding the content and for guidance throughout the project.

References

Mitchell, Tom M. *Machine Learning*. McGraw Hill, 2015.
Soni, Devin. "Introduction to Naive Bayes Classification." *Towards Data Science*, Towards Data Science, 16 July 2019