Semantic Question Matching

Lihua Ma

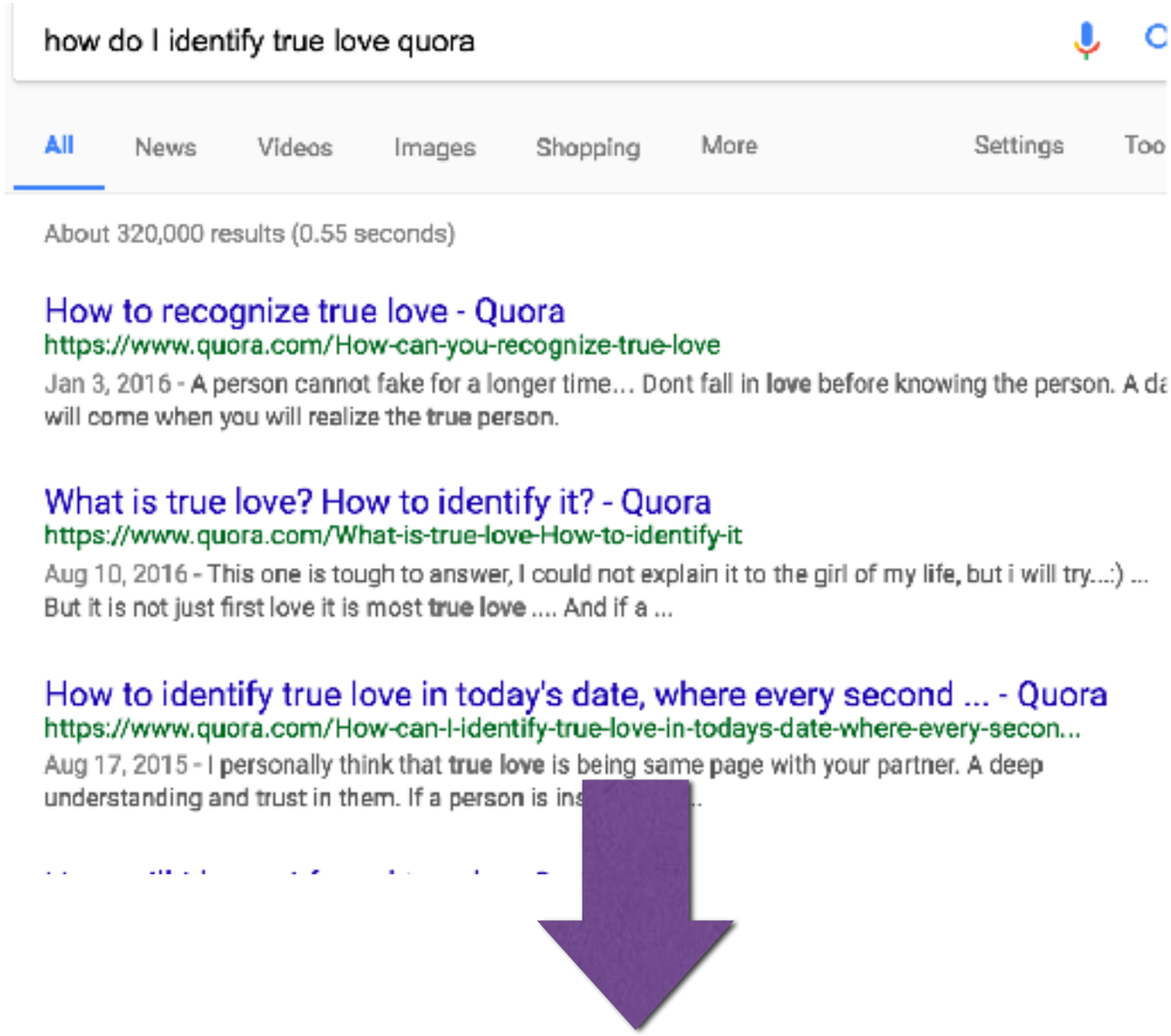
April, 2017

LinkedIn: https://www.linkedin.com/in/lihuama

Websites for Asking Questions

- Stack Overflow
- Stack Exchange
- Quora
- Wikihow
- Reddit
- Google Groups
- Yahoo! Answers
- FunAdvice
- Answers
- Blurt it

Search Results on One Question



Three pages results on quora, some of them are duplicated questions.

Dataset Sample

Question1	Question2	is_duplicate
How do I identify true love?	How can you recognize true love?	1
How do I identify true love?	How do I find true love?	0
How do I identify true love?	How do I identify true friendship?	0

404,290 question pairs from Quora Data

Method One: Regular Feature Engineering

Q1: How do I identify true love?

Q2: How can you recognize true love?

- Remove ‘what, when, where, which, why, how’ from regular stop_words list
- Not using stemming or lemmatization

Q1_unigram:[how, identify, true, love]

Q2_unigram:
[how, recognize, true, love]

Common_unigram_ratio:0.6

Q1_bigram:[(how, identify), (identify, true), (true, love)]

Q2: [(how, recognize), (recognize, true), (true, love)]

Common_bigram_ratio: 0.2

Q1_trigram:[[(how, identify, true), (identify, true, love)

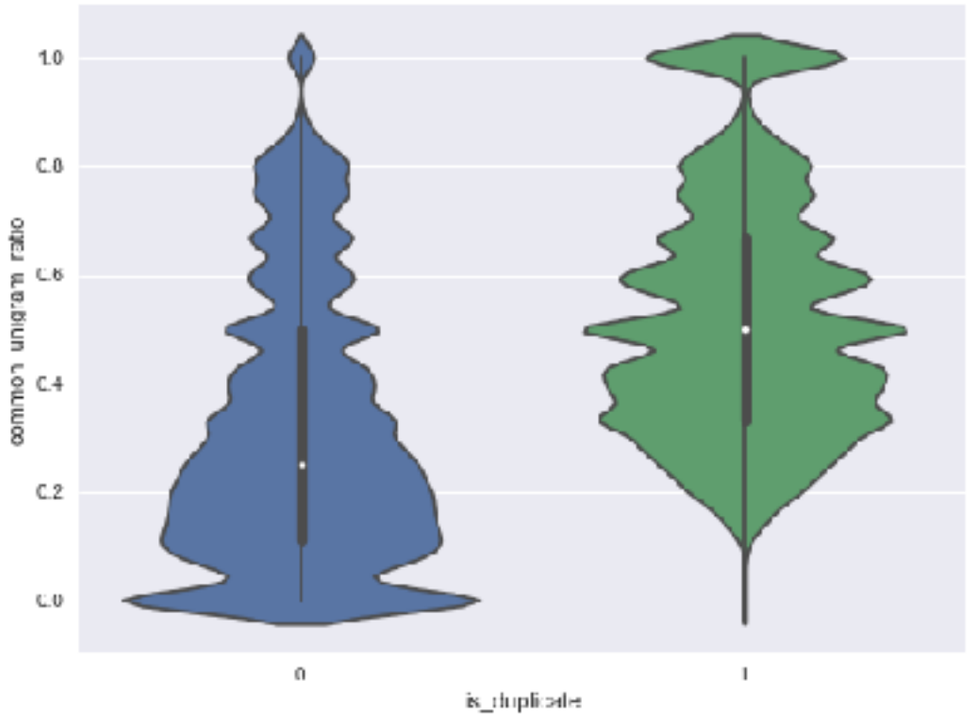
Q2_trigram:[(how, recognize, true), (recognize, true, love)]

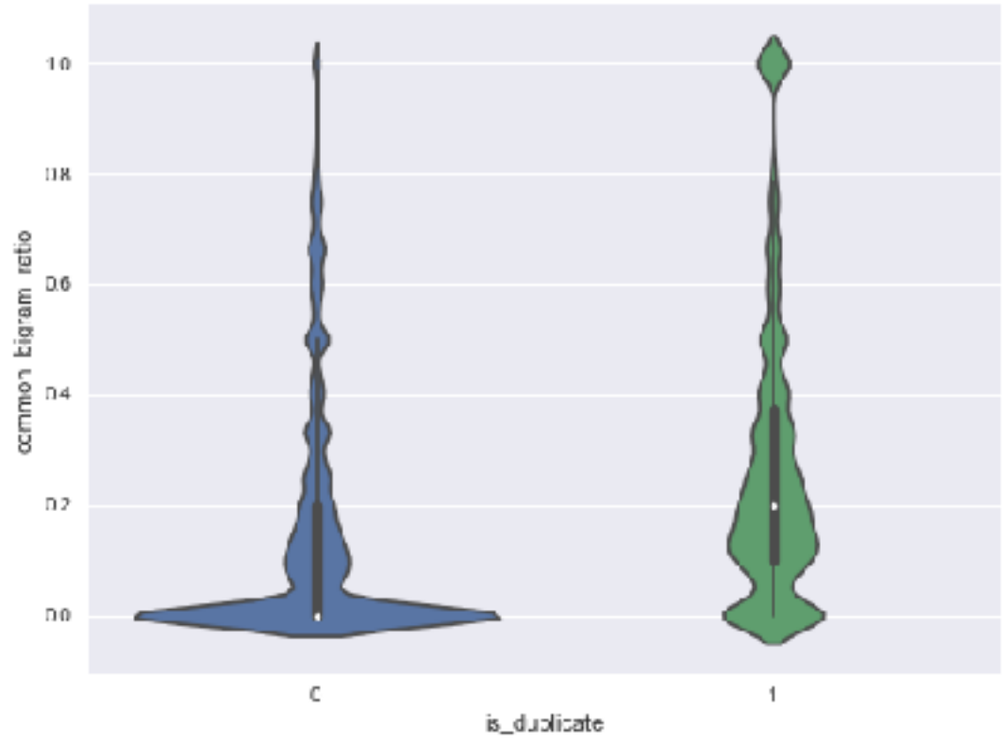
Common_trigram_ratio: 0.0

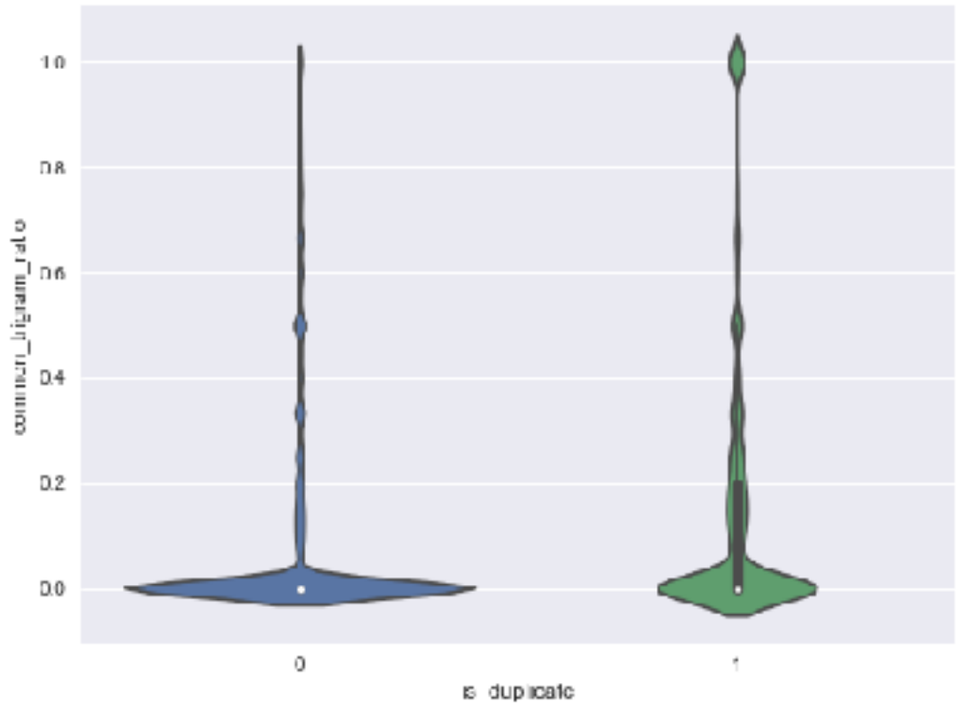
TF_IDF
Cosine_Similarity between Q1 and Q2

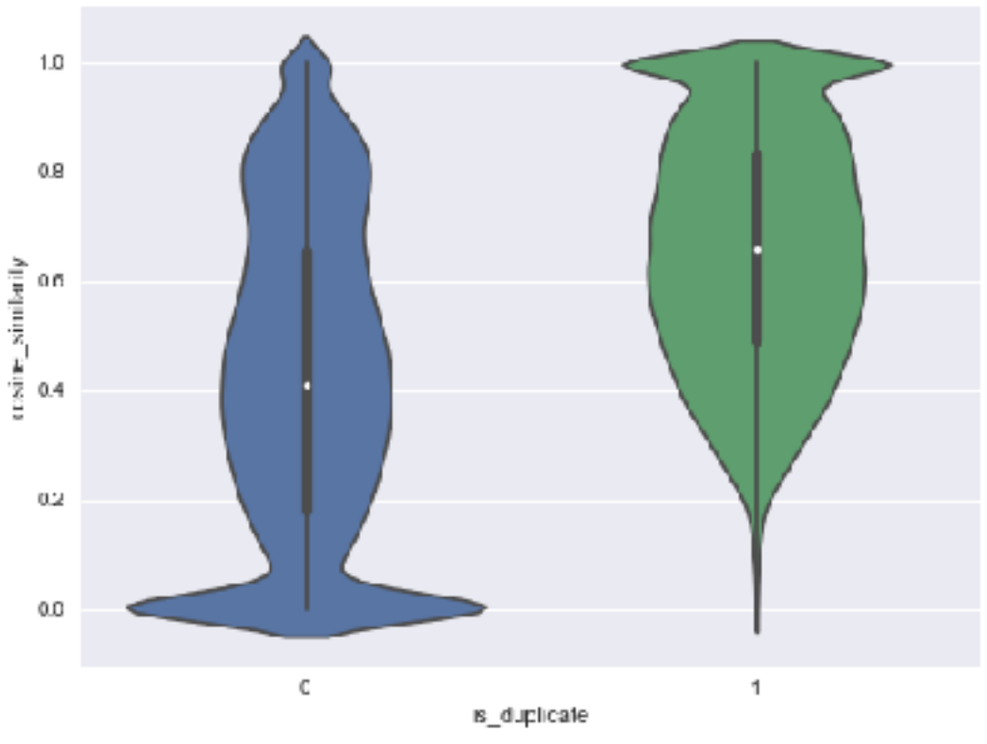
Cosine_Similarity: 0.499300

Difference Between Two Classes

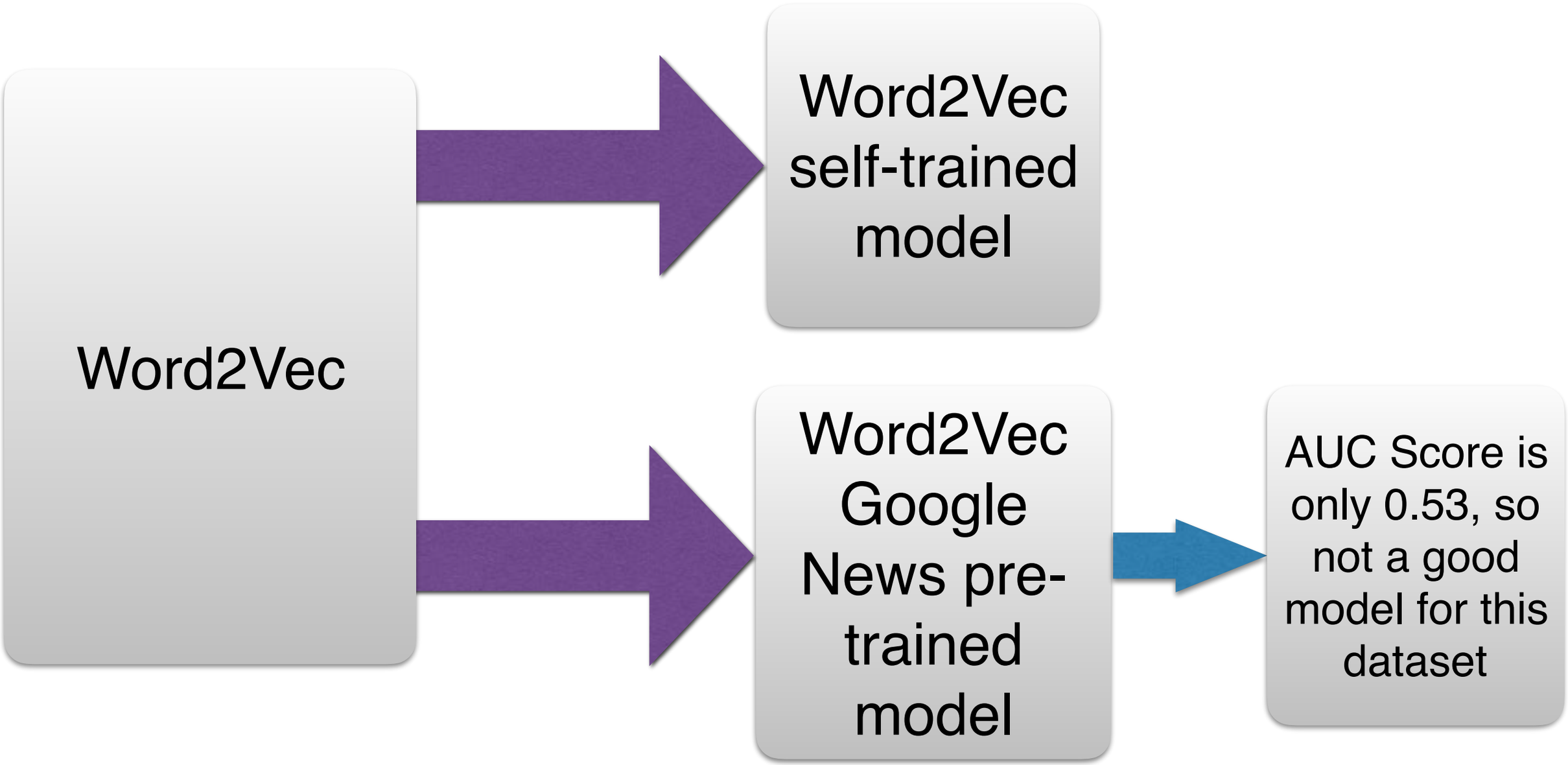








Method Two: Word2Vec



Methods Comparison

	Accuracy Score	Precision Score	AUC score
Method One (Regular Feature Engineering)	0.71	0.58	0.72
Method Two (Word2Vec on self_trained Model)	0.77	0.76	0.83

Conclusions

- It’s hard for machine to understand short sentences without context.

- Word2Vec improves AUC score a lot, because Word2Vec tries to learn the relationship between words in your dataset or in real life.

- Self_trained model on your own dataset tends to perform better than other available models especially if you have a specialized dataset.