Automated Scoring of Chatbot Responses in Conversational Dialogue

Authors: Steven Kester Yuwono[1], Wu Biao[1], Luis Fernando D'Haro[2]

1. National University of Singapore, NUS
2. Institute for Infocomm Research, A*STAR, Singapore

Outline

- Background
- Challenge
- Methodology
- Result & Discussion
- Acknowledgement

Background

- Chatbots development is increasingly popular
- How to evaluate their performance?

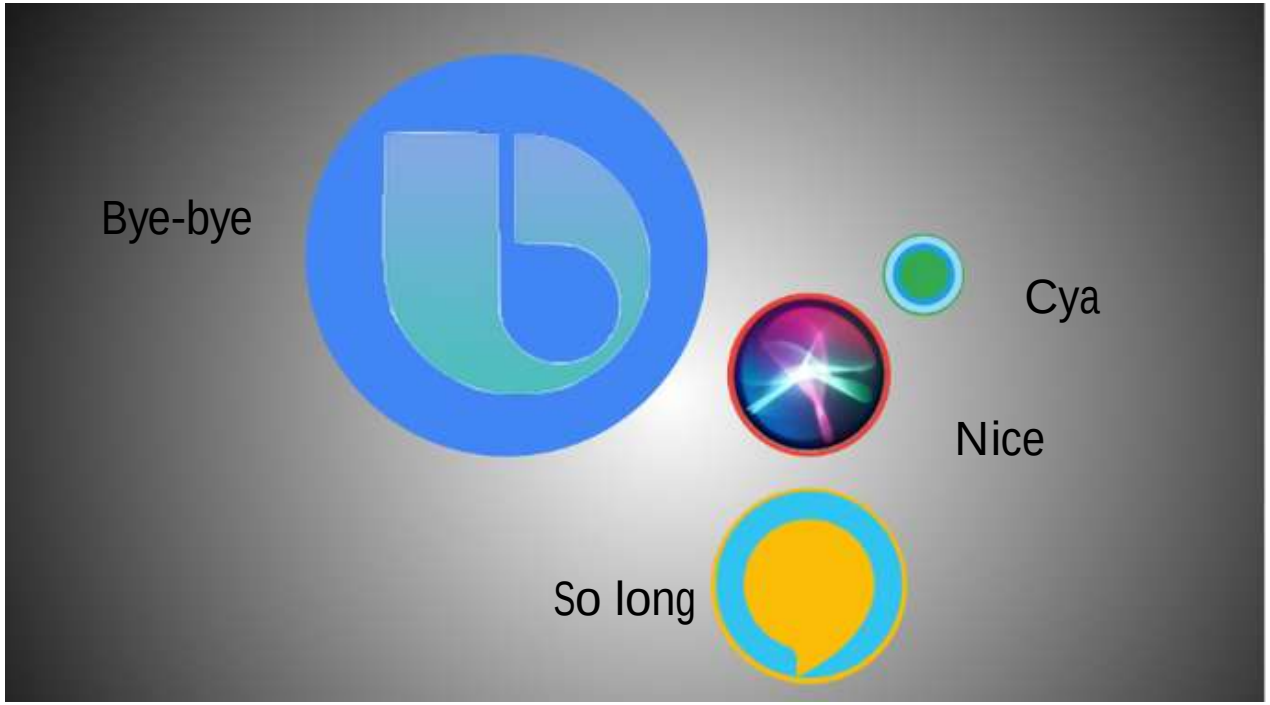


Illustration only, not actual response

Challenge

- Sequence based comparison to a reference response
  - BLEU, METEOR
- Even among human experts, inter-annotator agreement is low
  - For datasets used in the study, Fishers Interclass Correlation coefficient and Cohens Kappa coefficient are 0.5454 and 0.3736 respectively between experts.

Proposed Methodology

- Classification: pessimistic/optimistic
- Regression for averaging score
- Apply various word representations and machine learning models to evaluate their performance
  - Bag-of-words, Word Embeddings
  - SVM, Random Forest, CNN, RNN

Dataset

- WOCHAT: IRIS, TICKTOCK, and JOKER
- Annotated responses are either `valid, acceptable or invalid`
- Ground truth is constructed based on below three methods:
  - Optimistic
  - Pessimistic
  - Averaging

Sample Conversations

Turns	Annotation
User: yes, sorry to bother you with this Bot: It's fine, it's fine.	Valid
User: I am a geniuous Bot: My son pitched a no hitter!	Invalid
User:i'm not. i m going. bye-bye Bot: mrs florrick can i show you something	Acceptable

Ground Truth Construction

Data set	Annotator 1	Annotator 2	Annotator 3	Optimistic	Pessimistic	Averaging
Data1	Valid	Valid	Valid	1	1	1
Data2	Invalid	Invalid	Invalid	0	0	0
Data3	Valid	Valid	Invalid	1	0	2/3
Data4	Valid	Acceptable	Invalid	1	0	0.5

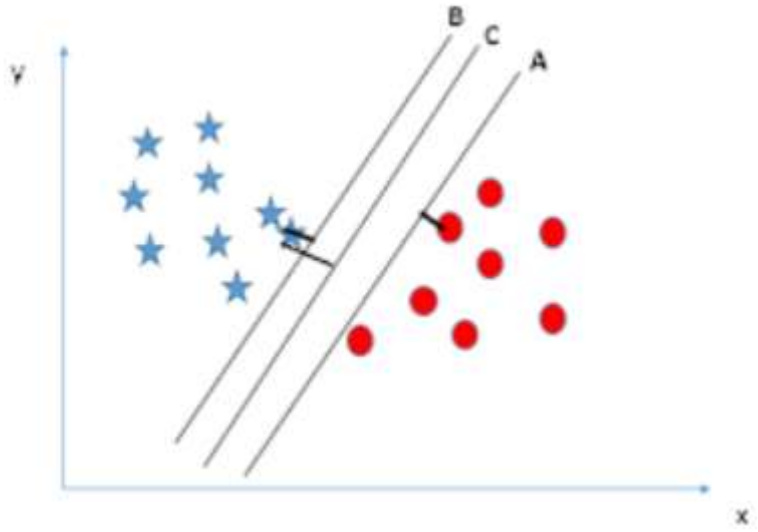
Ground Truth Statistics

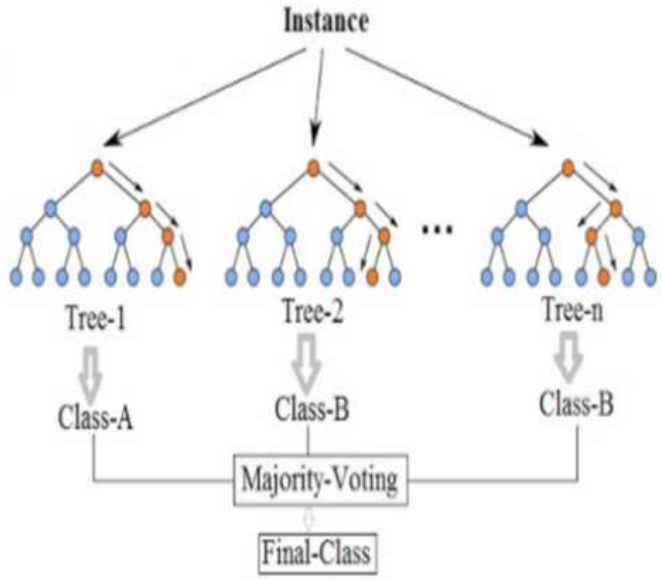
Dataset	Total	Ground Truth	Valid	Invalid
TickTock	2731	Optimistic	1786	945
		Pessimistic	940	1791
IRIS	790	Optimistic	661	129
		Pessimistic	244	546
Joker	535	Optimistic	362	173
		Pessimistic	142	393

The number of annotated chatbot turns for each dataset

SVM and Random Forest

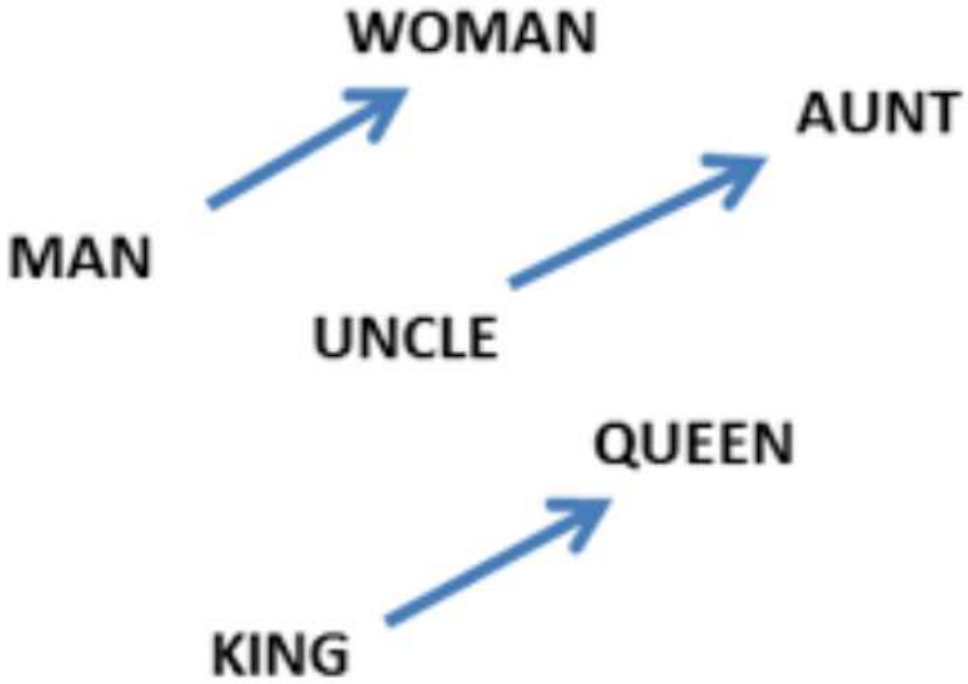
- Bag of words representation is used here, hence sequence information is lost
  - SVM: maximize margin
  - Random Forest (RF): bootstrap aggregation to reduce variance



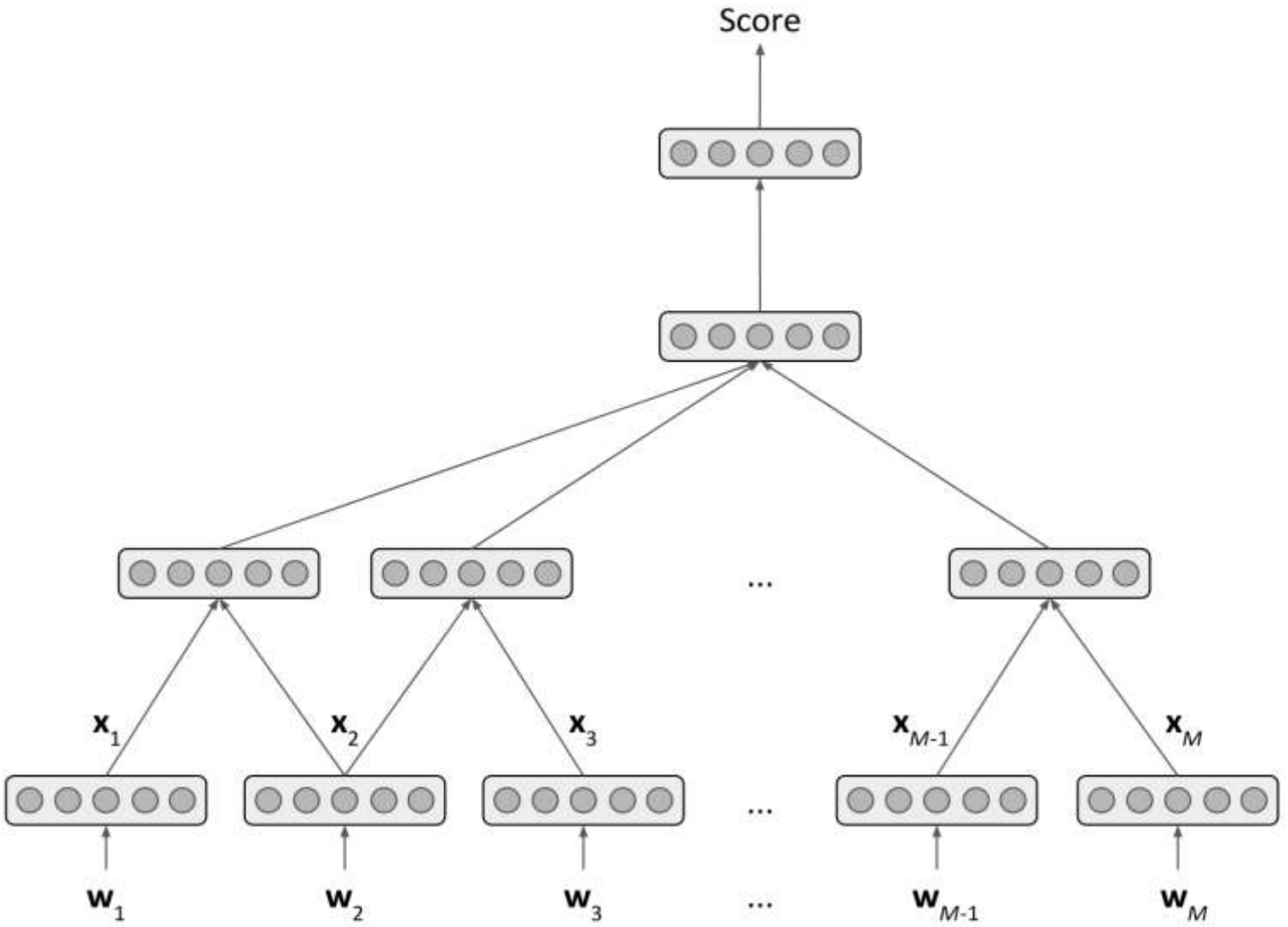


Neural Network

- Two representative neural network type are evaluated (CNN and RNN)
- Word Embedding (word vector)



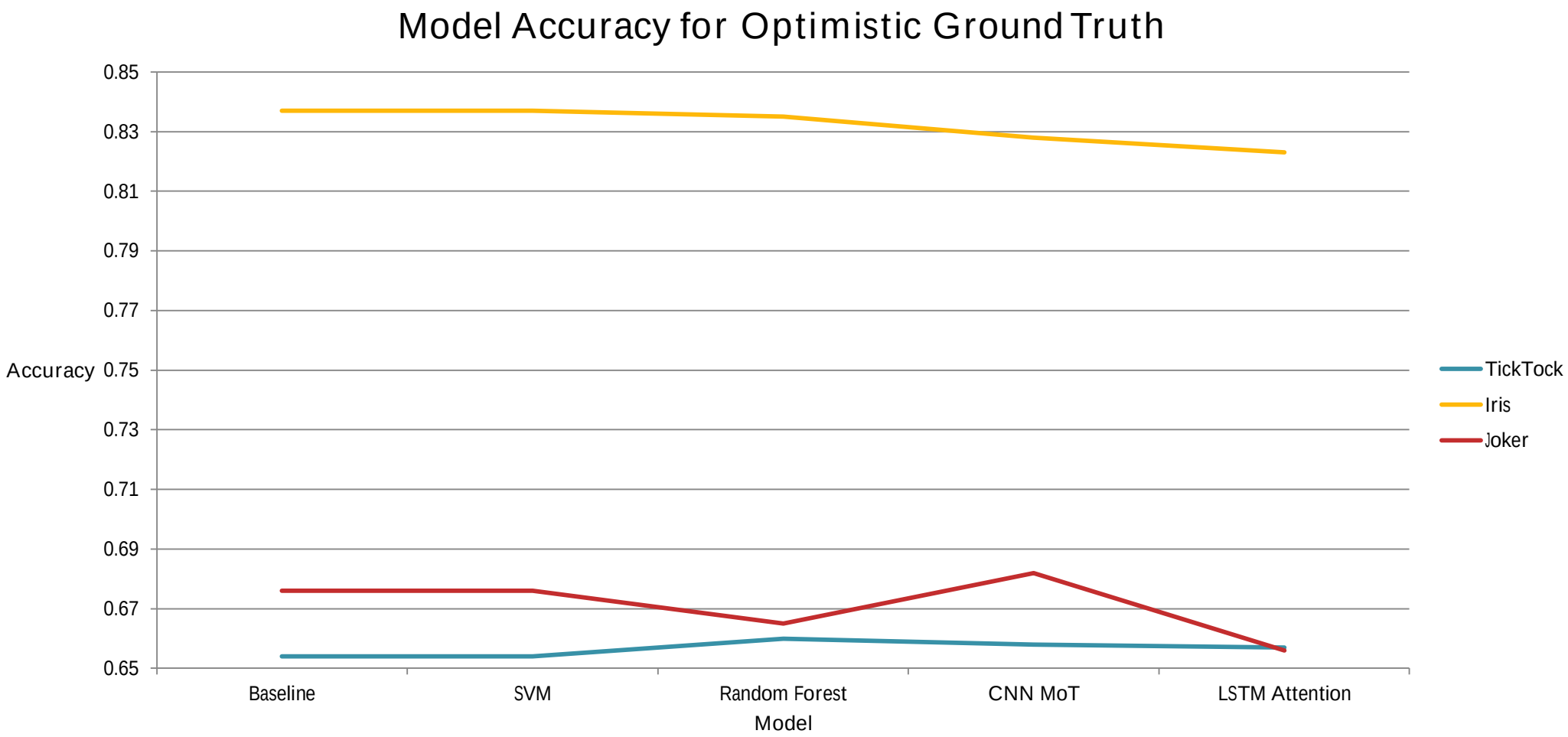
Neural Network Structure



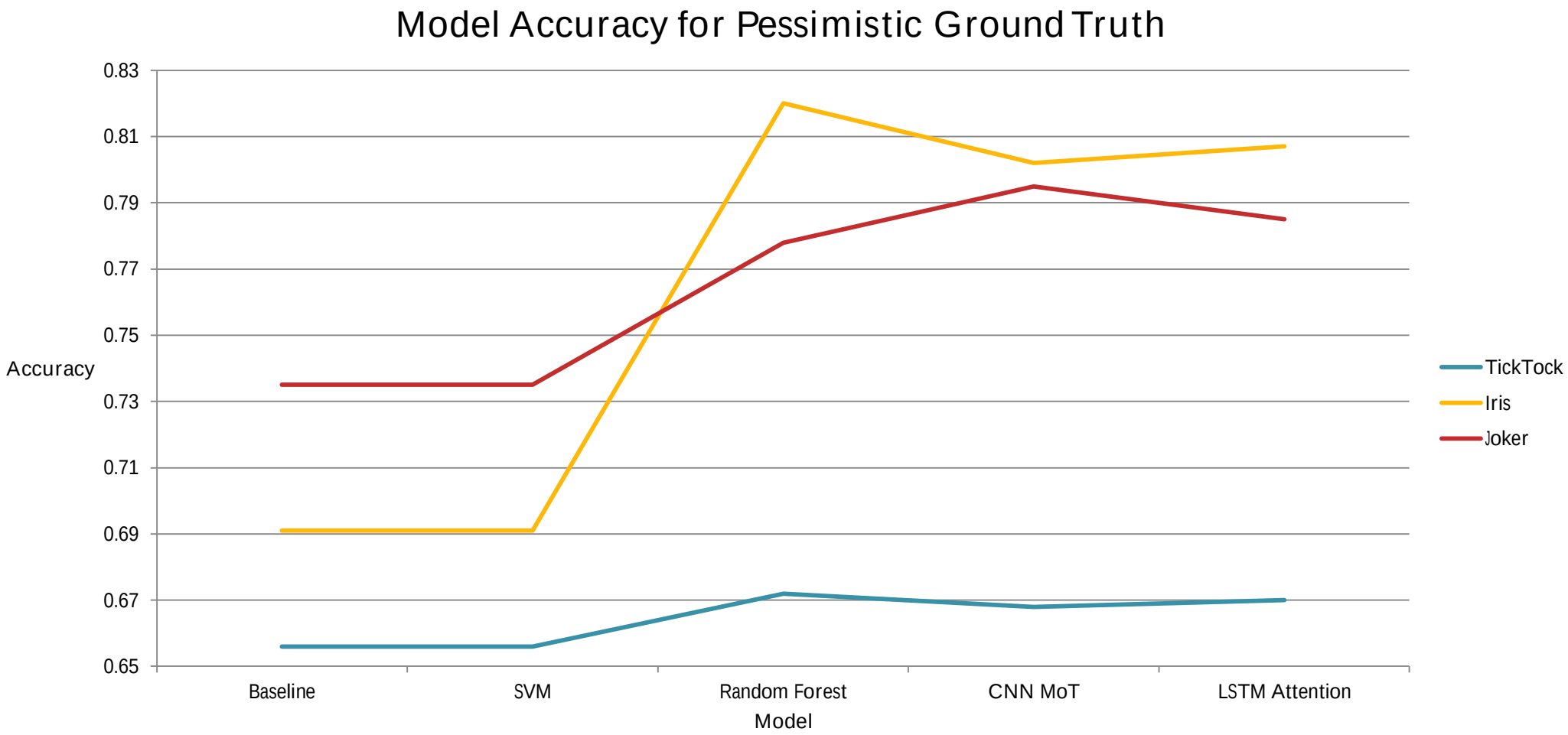
Result and Discussion

- Baseline predicts majority of the class
- In optimistic case, almost all models perform worse than baseline
- In pessimistic case, all models perform better than baseline

Result and Discussion



Result and Discussion



Result and Discussion

- All models outperforms baseline in regression case, as expected
  - Voting model of CNN performs well

Model	Pearson Correlation Coefficient		
	TickTock	IRIS	Joker
Baseline	0.0024 ± 0.061	0.0014 ± 0.104	0.0033 ± 0.138
SVM	0.225	0.457	0.333
Random Forest	**0.309**	0.464	**0.465**
CNN MoT	0.277	0.481	0.455
LSTM Attention	0.261	**0.505**	0.381
Voting CNN MoT	0.269	0.486	0.449

Averaged Pearson correlation coefficient with averaging ground truth

Result and Discussion

- Models perform well in pessimistic case because they can predict valid turn based on opening and closing remarks, which are highly similar in most valid responses
- Most valid responses are short as well
- Using large vocabulary may have better performance

Acknowledgement

- Thanks to Sunil Sivadas and Rafael Banchs from A*STAR for their meaningful guidance and comments

Resource

- Code used in this research is publicly accessible at https://github.com/yulonglong/ChatbotScorer


Q.
A.