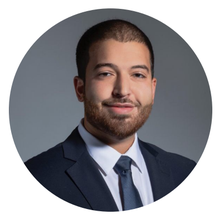
Explainability and Bias Removal in Natural Language

IBM Mentor: Mr. Rob Yates, DE, Executive
MIT Mentors: Prof. Don Sull & Prof. Carine Simon
Cambridge, MA, USA


Nader Hoballah


Annelise Steele

PROBLEM STATEMENT & SCOPE

WATSON ASSISTANT

- Watson Assistant is part of IBM Watson's suite of enterprise-ready AI business solutions
- Allows business users to build their own domain-specific chatbot — a customizable "Siri"

PROBLEM STATEMENT

- Machine learning models form relationships between features and outputs in a purely pragmatic manner, based on what produces the best overall performance
- The most powerful models currently available — including those used by Watson Assistant — do not offer interpretability, meaning the validity of the relationships a model forms cannot be confirmed
- **In order to increase end-user confidence and long-term performance of the end-user's Watson Assistant, we aimed to derive explainability from non-interpretable models to allow domain experts on the client side to assess and calibrate the validity of classifier relationships**

DATA OVERVIEW

UTTERANCES AND INTENTS

- Users request information from Watson Assistant in what is known as the user's utterance, and Watson Assistant classifies the objective of the user's request, which is known as the user's intent

Utterance Example: 'When do I need to submit my performance assessment?'
Corresponding Intent: 'AssessmentDueDate'

DATASETS: GROUND TRUTH & LOG DATA

32	User "intent" classifications
2,900	Labeled "ground truth" utterances, used for training multiclass model
14,300	Unlabeled log data utterances consisting of real-life user inputs, used for bias-removal and testing performance

FEATURE ENGINEERING

- Unigram and bigram embeddings to preserve interpretability
- Limited preprocessing to preserve outliers

10,662
UNIQUE FEATURES

MULTICLASS MODEL

LIGHTGBM
Selected as surrogate classifier due to speed and performance

76%
OUT-OF-SAMPLE ACCURACY

HYPEROPT
To attain optimal hyperparameters

78%
OUT-OF-SAMPLE PRECISION

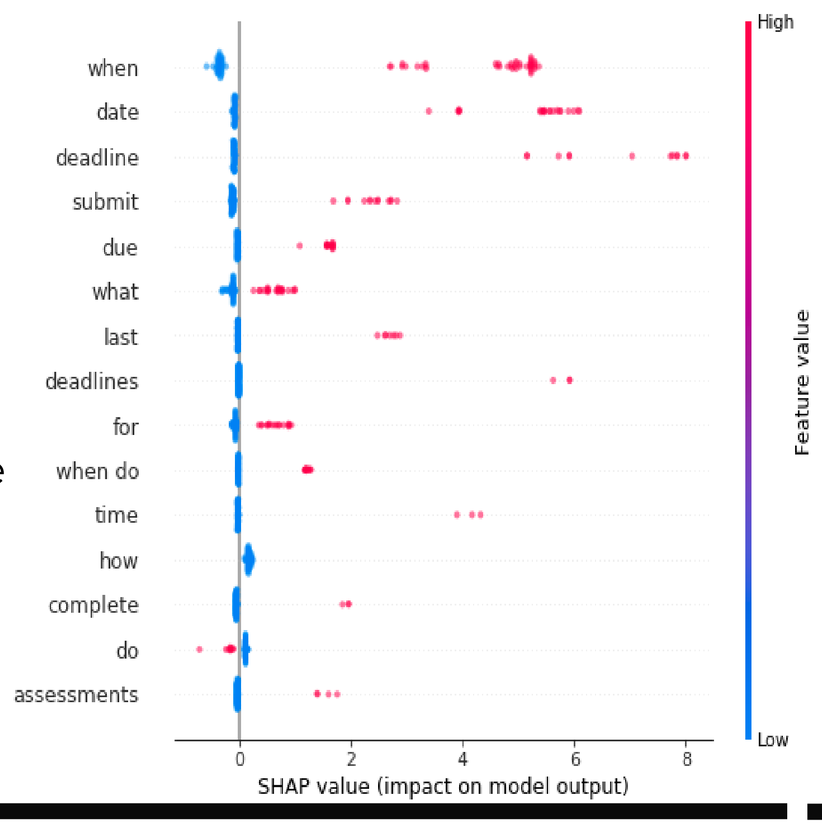
EXPLAINABILITY

APPROACHING EXPLAINABILITY

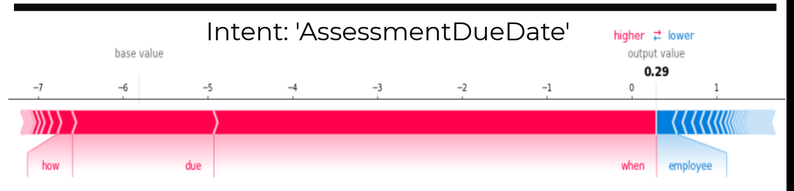
- Able to identify the most significant features (words) in each classification as well as their influence on the final confidence using SHAP values
- Quickly noticed a trend of intent-related keywords comprising the majority of the confidence for a particular classification
- Produced aggregate plots of feature influence across subsets of data (true and false positives) to identify most influential features, and if the classifier relationship appeared to be valid
- Presence of keywords propels the confidence of a classification quite high, while the absence of the same term does not reduce confidence by nearly as much

FEATURE INFLUENCE FORCEPLOT

- Red indicates presence of word in utterance; blue indicates absence
- Positive values add confidence; negative reduces

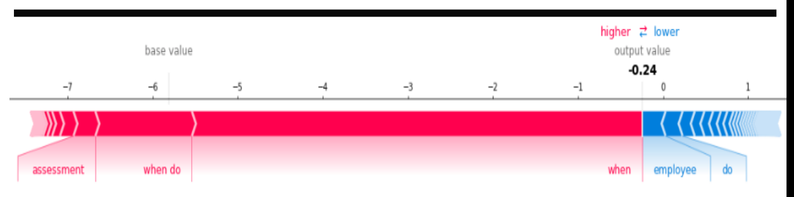


TRUE POSITIVE EXAMPLE



"When is the employee assessment due?"

FALSE POSITIVE EXAMPLE



"when do i communicate the assessment to the employee"

CHI-SQUARE ANALYSIS

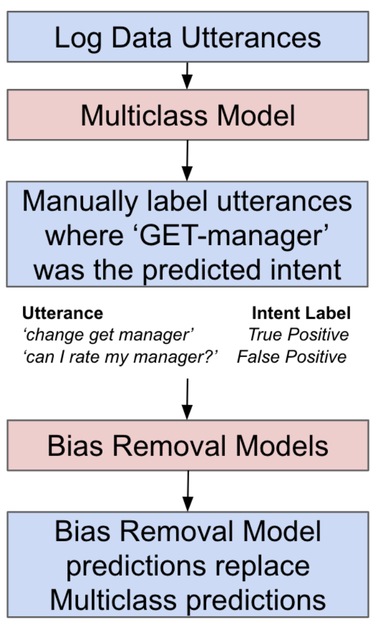
- Chi-Square feature selection produced similar results to SHAP forceplots
- Confirms statistical significance of highly correlated terms
- Also used to surface terms more prevalent in misclassifications

Term	Chi-Sq. Value	p-value
when	811.84	1.44e-178
date	471.61	1.43e-104
due	434.30	1.89e-96
deadline	352.74	1.07e-78
when do	280.38	6.22e-63
deadline for	158.34	2.61e-36
date for	157.47	4.03e-36

BIAS REMOVAL

BIAS REMOVAL IMPLEMENTATION

- Focus was on improving the precision of a specific intent using log data
- Objective: build a cascade with superior performance to baseline multiclass model in order to reduce bias

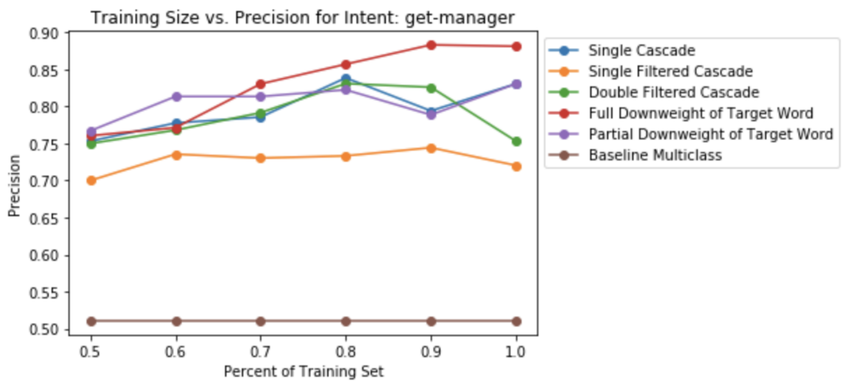


BIAS REMOVAL MODELS: CASCADING CLASSIFIERS

- The bias removal models implemented involved down-weighting or filtering the labeled log training data based on identified biased features
- The various bias removal models' precision across the labeled intents in log data test set can be seen below:

Model Type	GET-manager	AssessmentDueDate	Dimensions_Definition
Baseline Multiclass	**51.06%**	**86.42%**	**64.40%**
Single Cascade	83.07%	90.47%	87.05%
Single Filtered Cascade	69.07%	88.14%	65.21%
Double Filtered Cascade	81.42%	**91.12%**	**90.12%**
Full Downweight	**89.65%**	90.40%	86.90%
Partial Downweight	82.53%	90.40%	**90.12%**
Absolute Improvement	38.59%	4.70%	25.72%
Best Lift (%)	**75.57%**	**5.43%**	**39.93%**
Best Final Precision	**89.65%**	**91.12%**	**90.12%**

- The influence of cascade training data size on final precision can be seen below on intent 'GET-manager'



CONCLUSIONS

- All cascading classifiers produce a statistically significant performance lift over the multiclass model across intents
- Explainability is actionable by a domain expert to improve performance and reduce bias
- Improved performance occurs regardless of the size of the training set; cascade training data as low as 140 utterances can still offer significant lift
- Domain experts not required to invest significant time in labeling log data to see an improvement in precision, meaning the cascade solution is likely to be adopted by end-users
- Explainability manages to play an important role in certain cases where a problematic keyword needs to be down-weighted, eliminated, or used as a filter in the cascade model's training
- Different cascade variants may be more effective given characteristics of the intent, such as the percentage of occurrences of problematic terms