Can we predict self-reported customer satisfaction from interactions ?

Jeremy Auguste, Delphine Charlet, Geraldine Damnati, Frédéric Béchet, Benoit Favre

Can we predict self-reported customer satisfaction from interactions ?

Jeremy Auguste[1], Delphine Charlet[2], Géraldine Damnati[2], Frédéric Béchet[1], Benoit Favre[1]

(1) {firstname.lastname}@univ-amu.fr
(2) {firstname.lastname}@orange.com

Summary

- **Objective:** Automatically evaluate the customer satisfaction from conversation logs.
- **Data:** Contact center chat conversations and the customers' satisfaction surveys.
- **Method:** Comparison of different classification schemes: 3-labels, $2 \times$ 2-labels or 2-labels multitask classification. Definition of the Serious Error Rate metric to focus on problematic confusions.
- **Results:** Considering the classification of extreme opinions as two distinct tasks greatly improves the results on the neutral class.

Task and Motivations

Task: Evaluate the customer satisfaction from the logs of a human-human conversation.
Possible evaluations: Direct supervision using surveys filled by the customers themselves and indirect supervision by experts.
Problem: Customer surveys are not mandatory and experts can't evaluate every conversation.
Question: Can we retrieve directly from conversation logs such subjective opinions as the *Net Promoter Score* ?

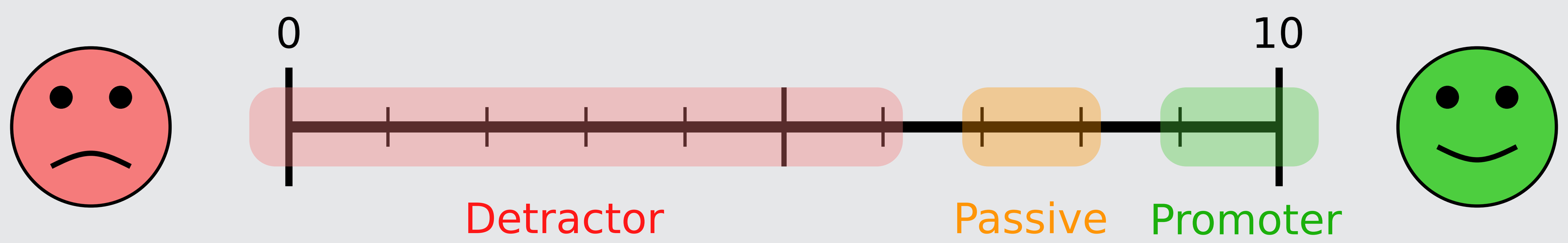
Orange conversation corpus

Chat data description:

- Technical and commercial assistance;
- 79,000 conversations with completed surveys;
- 140,000 unique tokens;
- Word Error Rate of 4.3% overall (10.1% for the *Customers*, 1.6% for the *Agents*)

Customer surveys

How likely would you be to recommend us to your family and friends ?

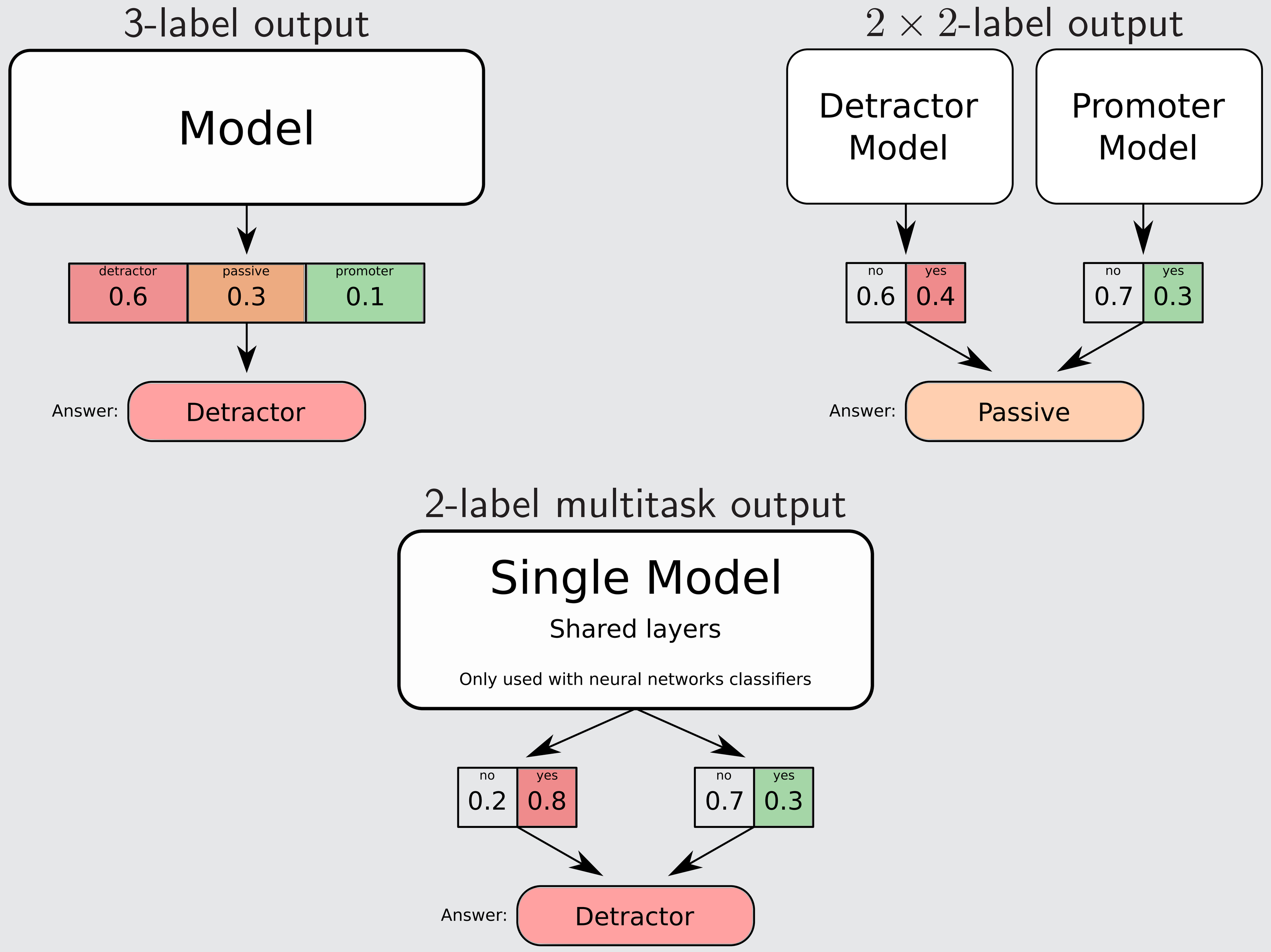


Following Customer Relationship Management conventions, appreciations are grouped into 3 categories: detractor, passive and promoter.

Acknowledgements

Research supported by grants ANR-15-CE23-0003 (DATCHA), ANR-16-CONV-0002 (ILCB) and ANR-11-IDEX-0001-02 (A*MIDEX).

Classification schemes



Goal: Avoid confusions between detractors and promoters.

Classifiers

Different classification methods that consider dialogues differently:

- Support-Vector Machine (SVM);
- Convolutional Neural Network (CNN);
- Long-Short Term Memory network with attention (RNN).

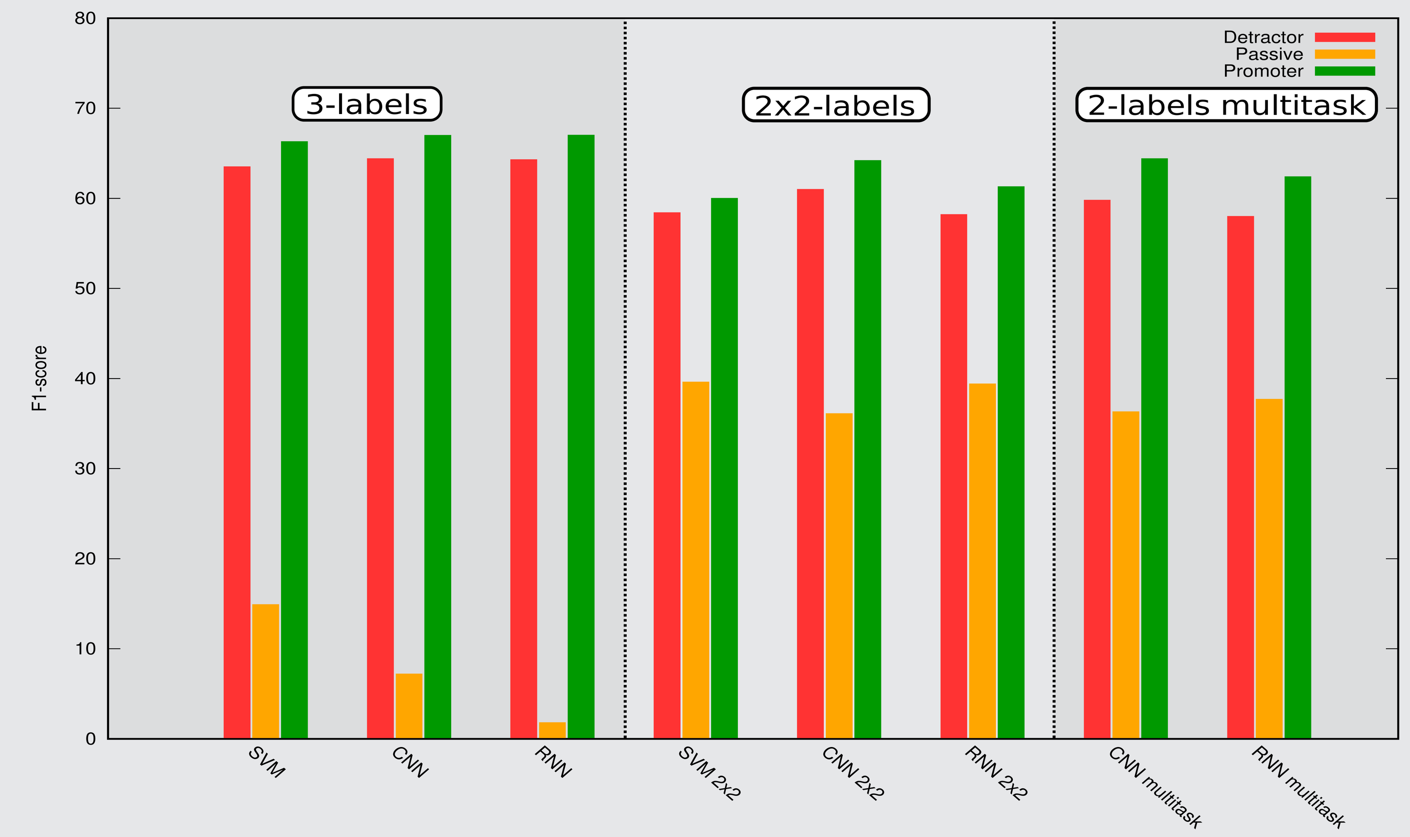
Evaluation metrics

Use of 3 different metrics:

- **Accuracy:** $\frac{\#\text{correct predictions}}{\#\text{samples}}$;
- **F1-score:** $\mathrm{F1}(l) = \frac{2 \times \mathrm{Precision}(l) \times \mathrm{Recall}(l)}{\mathrm{Precision}(l) + \mathrm{Recall}(l)}$;
- **Serious Error Rate**: Percentage of confusion between the Detractor and the Promoter classes.

Results

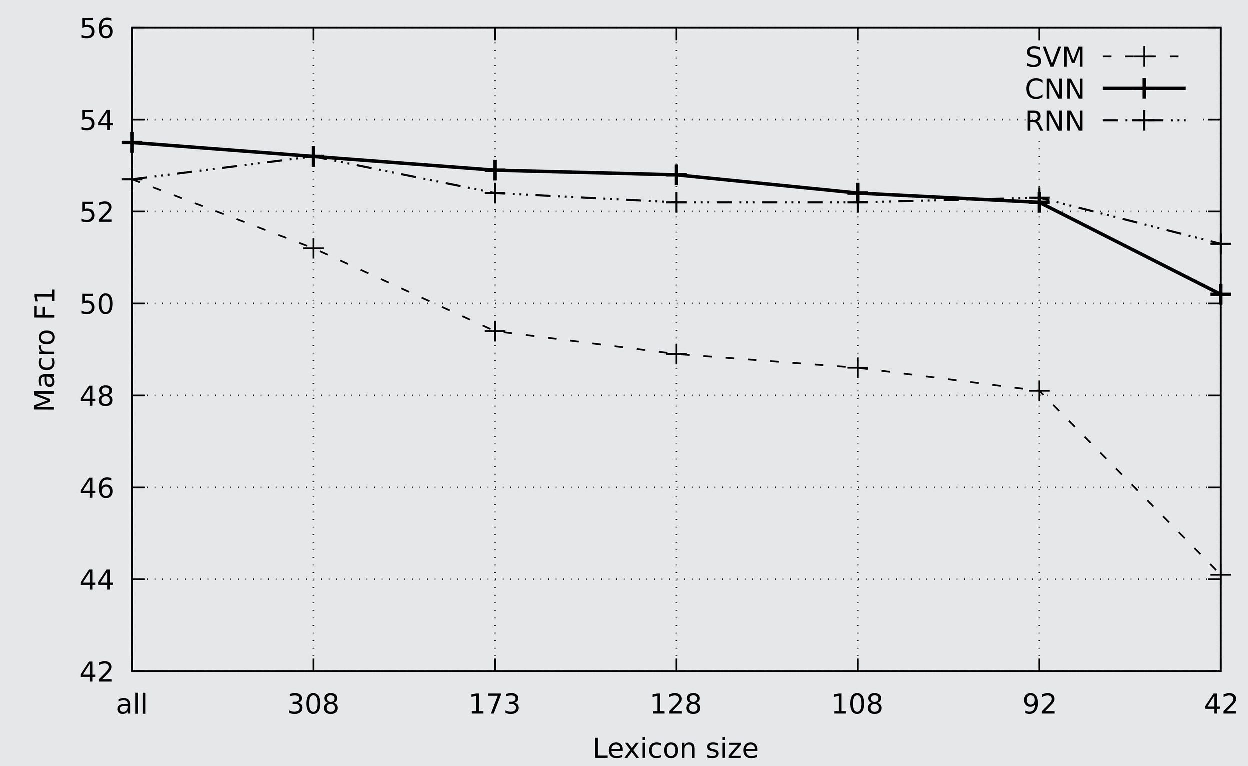
Model	Accuracy	Serious Error Rate
3-labels classification scheme		
majority class	42.7	30.9
SVM	56.9	**14.7**
CNN	**57.5**	15.5
RNN	**57.5**	15.8
2-labels+reject classification scheme		
SVM 2x2 labels	52.7	**6.2**
CNN 2x2 labels	**55.2**	7.7
CNN 2 labels multitask	55.0	7.6
RNN 2x2 labels	53.5	6.5
RNN 2 labels multitask	53.5	6.5



2-labels schemes greatly improve the prediction of the passive class and greatly reduce confusions between extreme classes.

Contrastive experiment

Reducing the lexicon size to evaluate domain robustness:



Lexicon reduced by selecting words occuring at least 10K to 100K times.