



**HAL**  
open science

## Can we predict self-reported customer satisfaction from interactions ?

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

### ► To cite this version:

Jeremy Auguste, Delphine Charlet, Geraldine Damnati, Frédéric Béchet, Benoit Favre. Can we predict self-reported customer satisfaction from interactions ?. International Conference on Acoustics, Speech and Signal Processing, May 2019, Brighton, United Kingdom. <hal-02134252>

**HAL Id: hal-02134252**

**<https://hal.science/hal-02134252v1>**

Submitted on 20 May 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

(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 × 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* ?

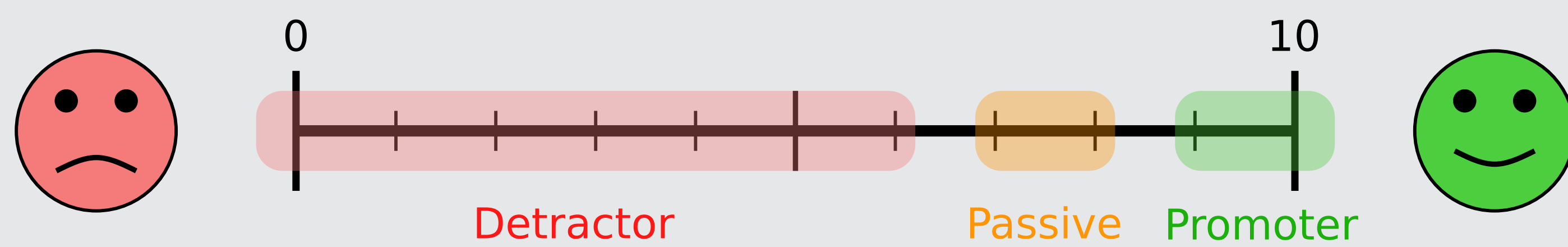
## Orange conversation corpus

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 ?

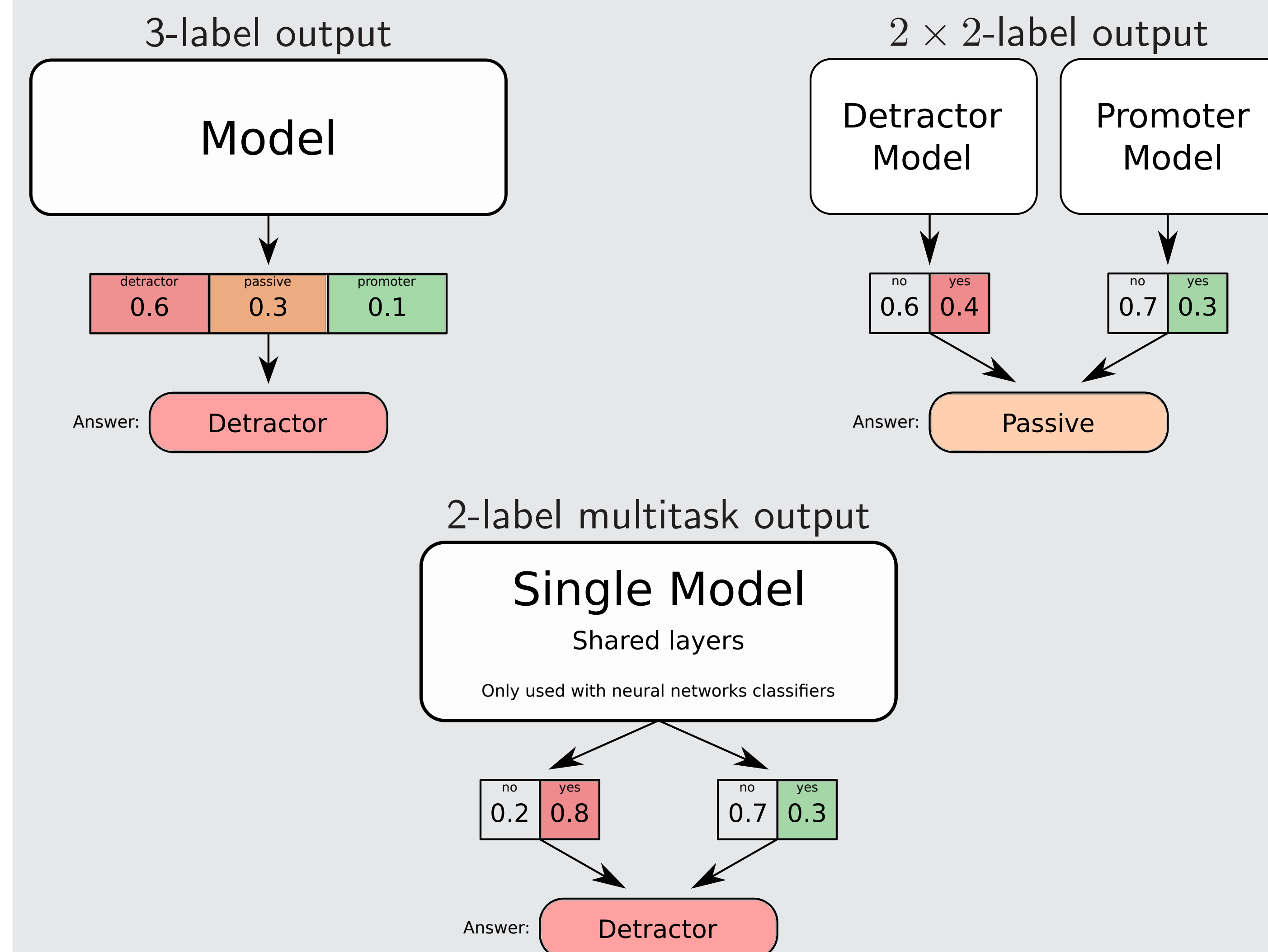


Following Customer Relationship Management conventions, appreciations are grouped into 3 categories: **detractor**, **passive** and **promoter**.

## 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).

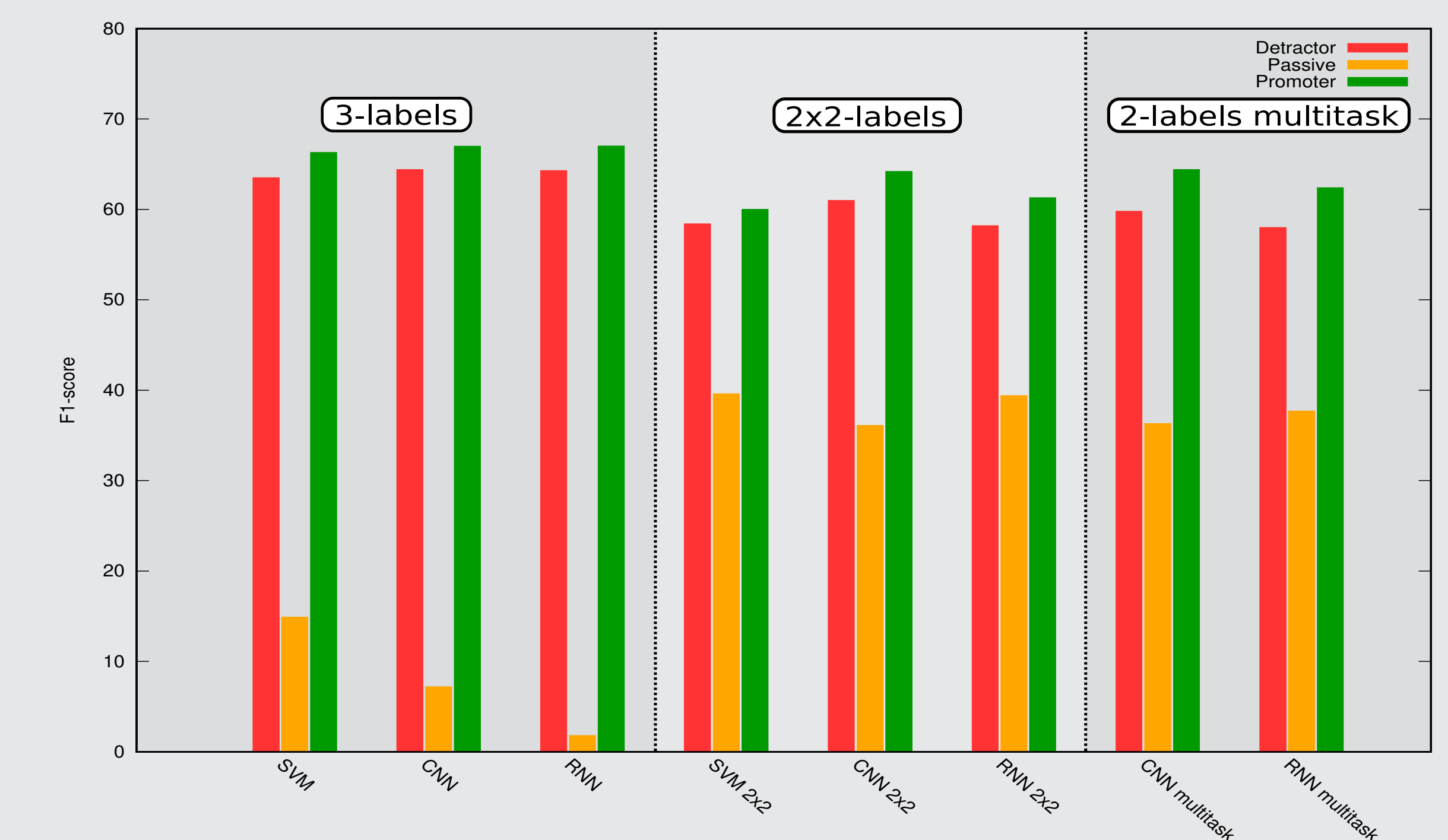
## Evaluation metrics

Use of 3 different metrics:

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

## Results

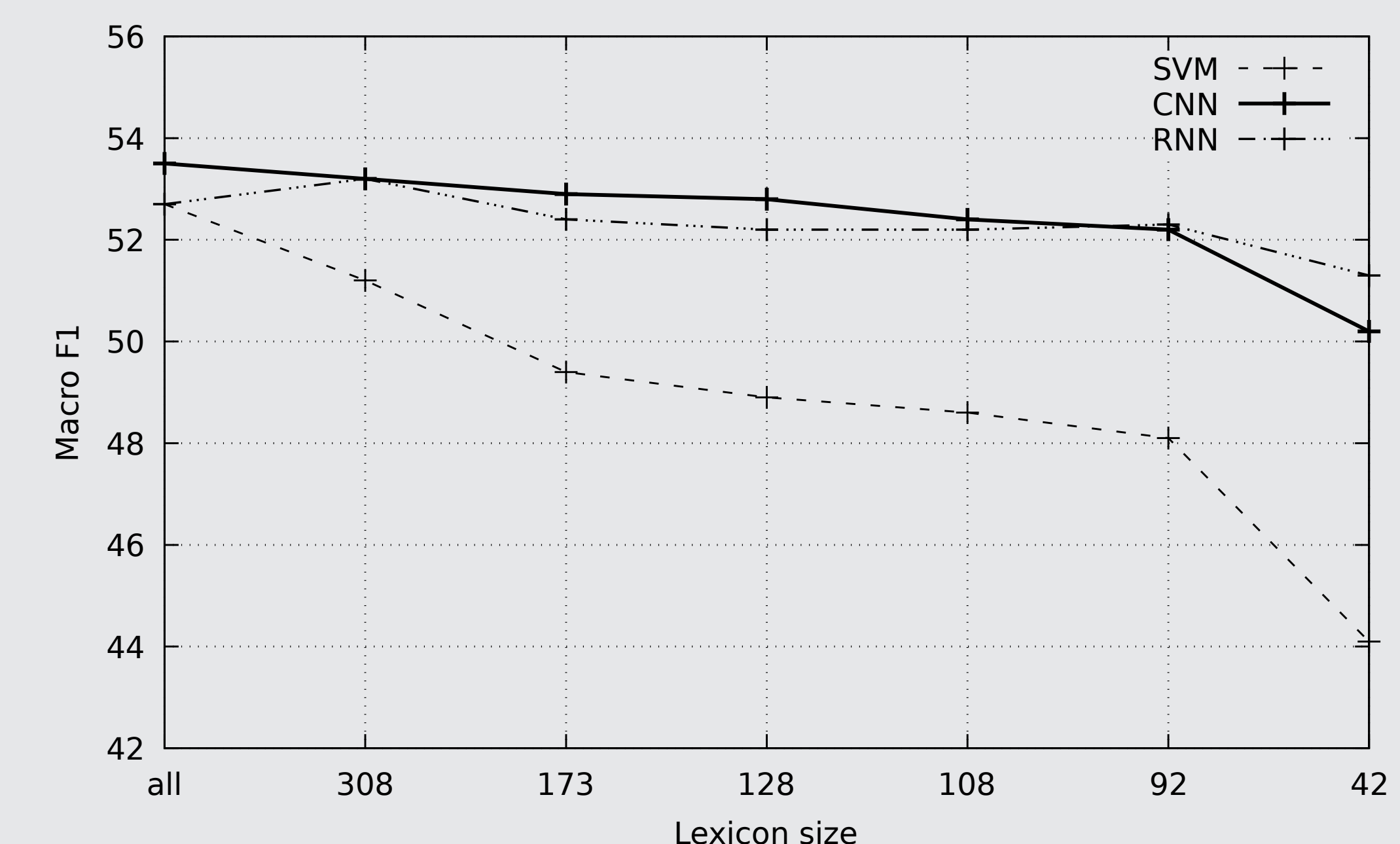
Model	Accuracy	Serious Error Rate
<i>3-labels classification scheme</i>		
majority class	42.7	30.9
SVM	56.9	<b>14.7</b>
CNN	<b>57.5</b>	15.5
RNN	<b>57.5</b>	15.8
<i>2-labels+reject classification scheme</i>		
SVM 2x2 labels	52.7	<b>6.2</b>
CNN 2x2 labels	<b>55.2</b>	7.7
CNN 2 labels multitask	55.0	7.6
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 occurring at least 10K to 100K times.