Classifying Healthcare Forum Posts

Oguzhan (Ouz) Gencoglu

Tampere Data Science Meetup - 5 April 2017

Outline

1 The Competition
- The Task
- The Data

2 My Approach
- Design Process
- Final Pipeline

3 Results

4 Reflections & Comments

5 Further Applications

Who is Ouz Gencoglu?

TAMPERE UNIVERSITY OF TECHNOLOGY

Personal Health Informatics Team, Tampere University of Technology

Who is Ouz Gencoglu?

Co-founder & Chief Data Scientist

Healthcare Data Analysis Challenge

TUT's doctoral student Oguzhan Gencoglu selected as finalist in machine learning competition

The finalists of the ICHI2016 Healthcare Data Analytics Challenge have been announced. Doctoral student Oguzhan Gencoglu from the Personal Health Informatics research group at TUT is one of the top five finalists. His one-man-team will represent TUT in Chicago in October 2016.

In the challenge, the participant teams were given healthcare forum posts and they needed to develop machine learning solutions for predicting the topic of each of these posts using seven different categories, such as disease, treatment and family support. For the task, thousands of healthcare discussion forum posts and their corresponding topic categories were provided in text format, in addition to 3,000 posts without the ground truth for the teams to predict.

"The informal and untidy nature of the text data poses the main challenge," Gencoglu says.

The competition has lasted for two months and altogether five teams were selected as finalists. They will present their solutions at the IEEE International Conference on Healthcare Informatics in Chicago in October.

Gencoglu's natural language processing and machine learning solution will be competing against those of IBM Watson Research Center, New York University Langone Medical Center, University of Waterloo and Humana Inc.

tekniikka &talous

TEKNIKAN HISTORIA metalli-tekniikka

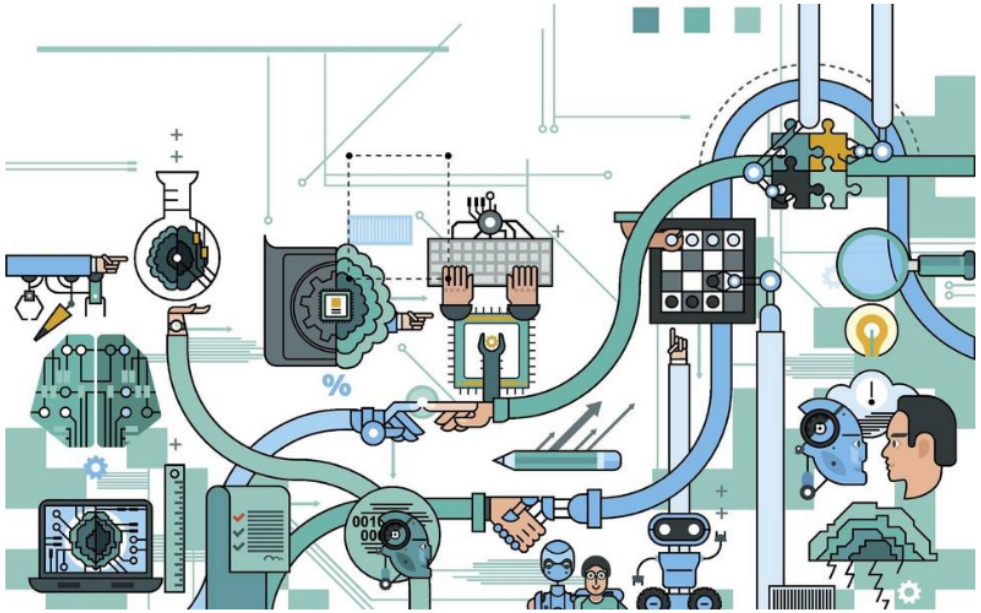
Tilaa lehti

Tilaa autokirje

ETUSIVU | UUTISET | PUHEENVUOROT | KUMPPANIBLOGIT | OTA YHTEYTTÄ

TIETOJENKÄSITTELY | Tapio Ikkala 23.8. klo 11:34

Koneoppimiskilpailun finalisti Tampereelta - Yhden miehen tiimi haastaa IBM:n Watsonin



Outline

1 The Competition
- The Task
- The Data

2 My Approach
- Design Process
- Final Pipeline

3 Results

4 Reflections & Comments

5 Further Applications

The Task

ICHI2016 Healthcare Data Analysis Challenge

Developing machine learning solutions to classify healthcare forum posts into 7 categories:

1. *Demographic*
2. *Disease*
3. *Family Support*
4. *Goal-oriented*
5. *Pregnancy*
6. *Socializing*
7. *Treatment*

Outline

1 The Competition
- The Task
- The Data

2 My Approach
- Design Process
- Final Pipeline

3 Results

4 Reflections & Comments

5 Further Applications

The Data

multiple prolapses and perineocele ??????

"I would appreciate anyone's opinionon a few matters. I have a cystocele, rectocele, enterocele, prolapsed uterus- Stage 1, and a perineocele. This is causing pain and problems so I need to have surgery. My doctor would like to use the DeVinchi robot. Has anyone done this?? How long is the recovery? Is it best to have uterus fixed, left alone or removed? I am 52 and not through menopause yet. Also, is it best to do all this at the same time? Thank you"

The Data

Day 28 :-(

"Well, the long wait is over. I woke up at 3:13 this morning. First thought was "TAKE YOUR FRER!" All I got was a BFN! Ugh. No sign of AF yet. DH doesn't believe the test. He said to wait a bit and take another one. Then, reminded me that I was almost 6 weeks with the last pregnancy before I showed positive POAS. So...maybe there's still a glimmer of hope. Thatś all for now ladies. Good luck to the rest.Katie"

The Data

- 8000 training, 3000 test examples

The Data

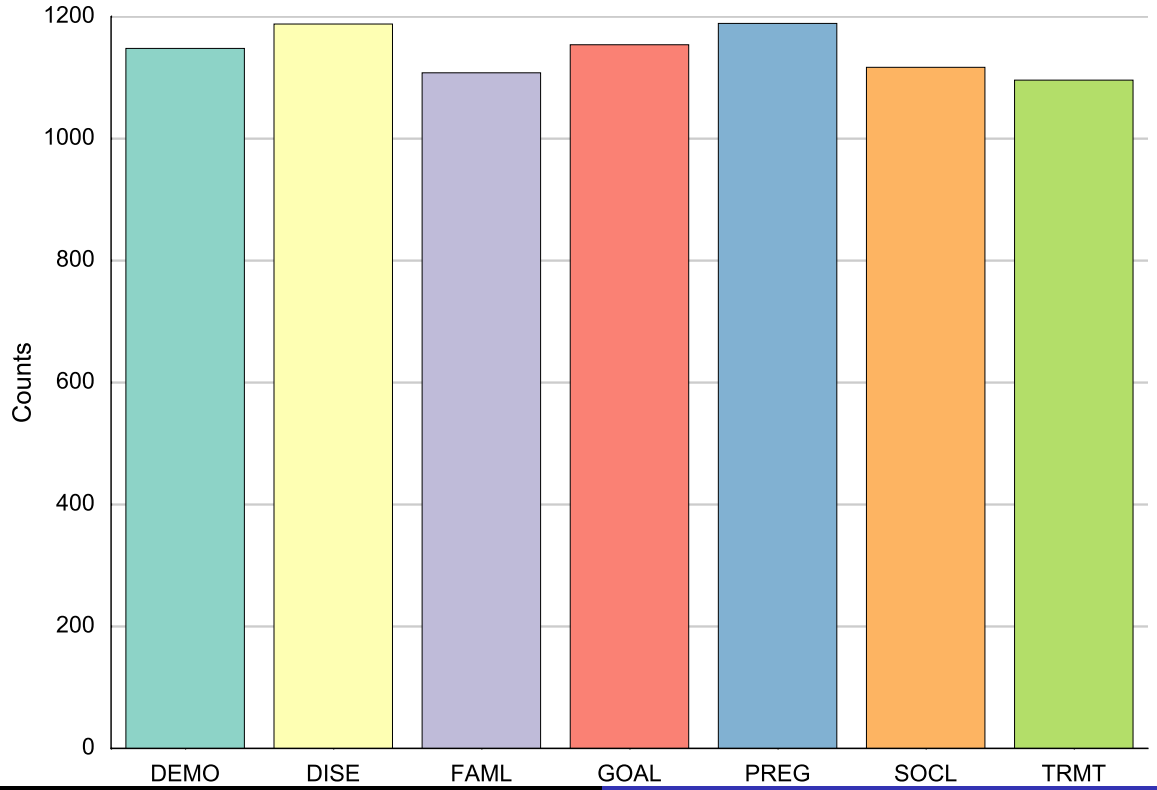
- 8000 training, 3000 test examples
- Each observation: Title + Post

The Data

- 8000 training, 3000 test examples
- Each observation: Title + Post
- Informal, untidy free text

The Data

- 8000 training, 3000 test examples
- Each observation: Title + Post
- Informal, untidy free text



Outline

1 The Competition
- The Task
- The Data

2 My Approach
- Design Process
- Final Pipeline

3 Results

4 Reflections & Comments

5 Further Applications

Cross Validation & Hyper-Parameter Optimization

10-fold CV

- All preprocessing actions

Cross Validation & Hyper-Parameter Optimization

10-fold CV

- All preprocessing actions
- Choice of classifier
  - kNN
  - Random Forest
  - SVM
  - Passive-Aggressive Classifier
  - Naive Bayes Classifier
  - Logistic Regression
  - LSTM & CNN variations

Cross Validation & Hyper-Parameter Optimization

10-fold CV

- All preprocessing actions
- Choice of classifier
  - kNN
  - Random Forest
  - SVM
  - Passive-Aggressive Classifier
  - Naive Bayes Classifier
  - Logistic Regression
  - LSTM & CNN variations
- Grid search over hyper-parameters, *e.g., SVM:*
  - C: [0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0]
  - kernel: ['linear', 'rbf', 'poly']
  - gamma: [0.01, 0.05, 0.1, 1.0]

Outline

1 The Competition
- The Task
- The Data

2 My Approach
- Design Process
- Final Pipeline

3 Results

4 Reflections & Comments

5 Further Applications

Preprocessing

For each observation in both training and test set:

1. *Title and post are merged with a whitespace in between.*
2. *All text is converted to lowercase letters/characters.*
3. *A whitespace is added after each '.' or ',' unless it is already there.*
4. *Any number of consecutive whitespace characters are transformed into a single whitespace character.*
5. *All non-ASCII characters are removed.*
6. *All urls are removed.*
7. *Word stemming is applied on each word with the help of WordNet lexical database*[1].

[1] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995.

Feature Extraction

Term Frequency - Inverse Document Frequency[2] (tf-idf)

+

2-grams & 3-grams

[2] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, 2003.

Feature Extraction

Term Frequency - Inverse Document Frequency[2] (tf-idf)

+

2-grams & 3-grams

Result

A sparse matrix of size **8000 x 1 182 516**

[2] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, 2003.

Classification

Online Passive-Aggressive Classifier[3]

[3] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *Journal of Machine Learning Research*, vol. 7, no. Mar, pp. 551-585, 2006.

Classification

Online Passive-Aggressive Classifier[3]

Parameters

- Loss Function: *Hinge Loss*

[3] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *Journal of Machine Learning Research*, vol. 7, no. Mar, pp. 551-585, 2006.

Classification

Online Passive-Aggressive Classifier[3]

Parameters

- Loss Function: *Hinge Loss*
- Regularization Coefficient: *0.3*

[3] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *Journal of Machine Learning Research*, vol. 7, no. Mar, pp. 551-585, 2006.

Results

	Accuracy
Local 10-fold CV	68.61%
Test Set	**67.47%**

Results

	Accuracy
Local 10-fold CV	68.61%
Test Set	**67.47%**

Remarks

- online learning

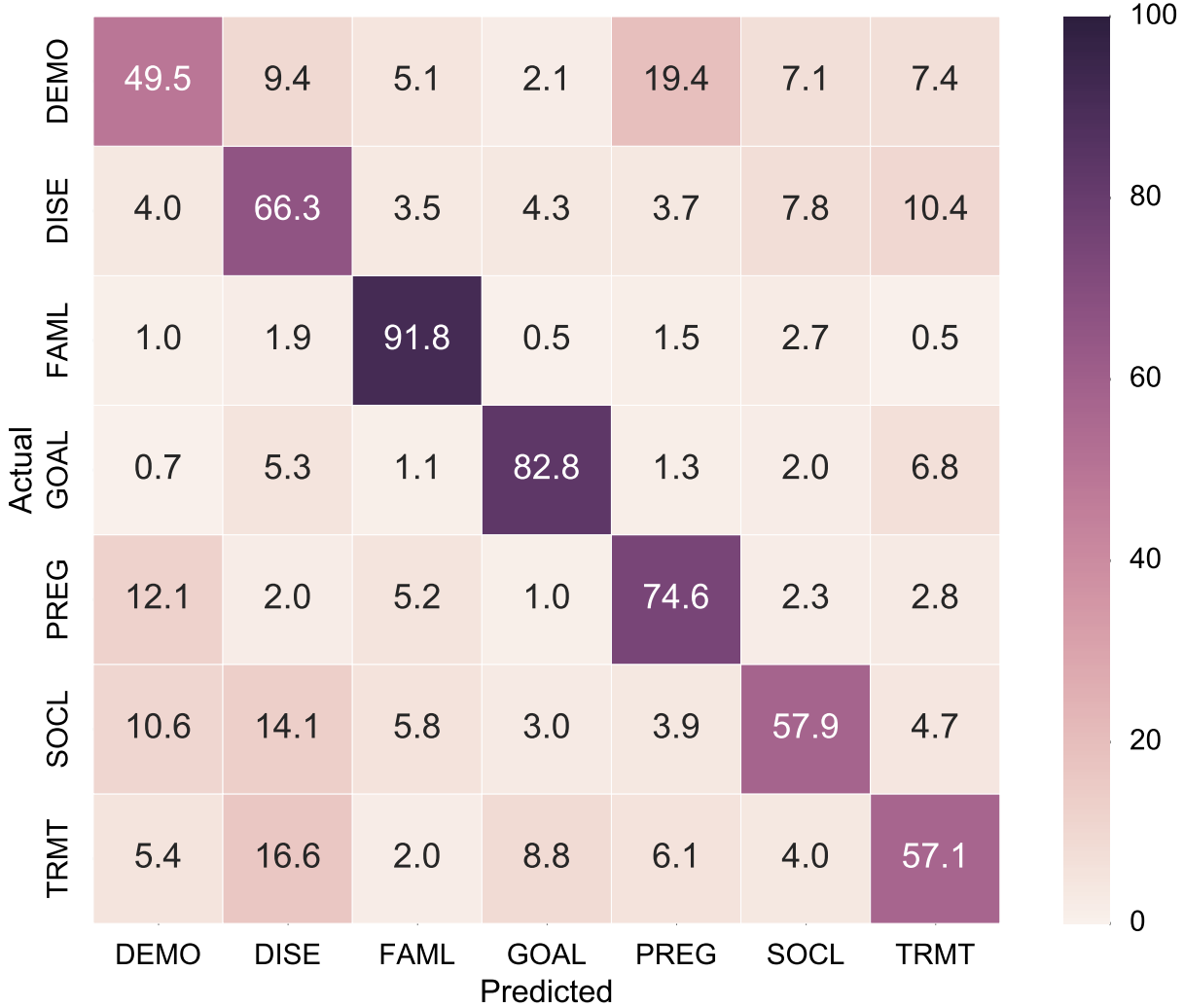
Results

	Accuracy
Local 10-fold CV	68.61%
Test Set	**67.47%**

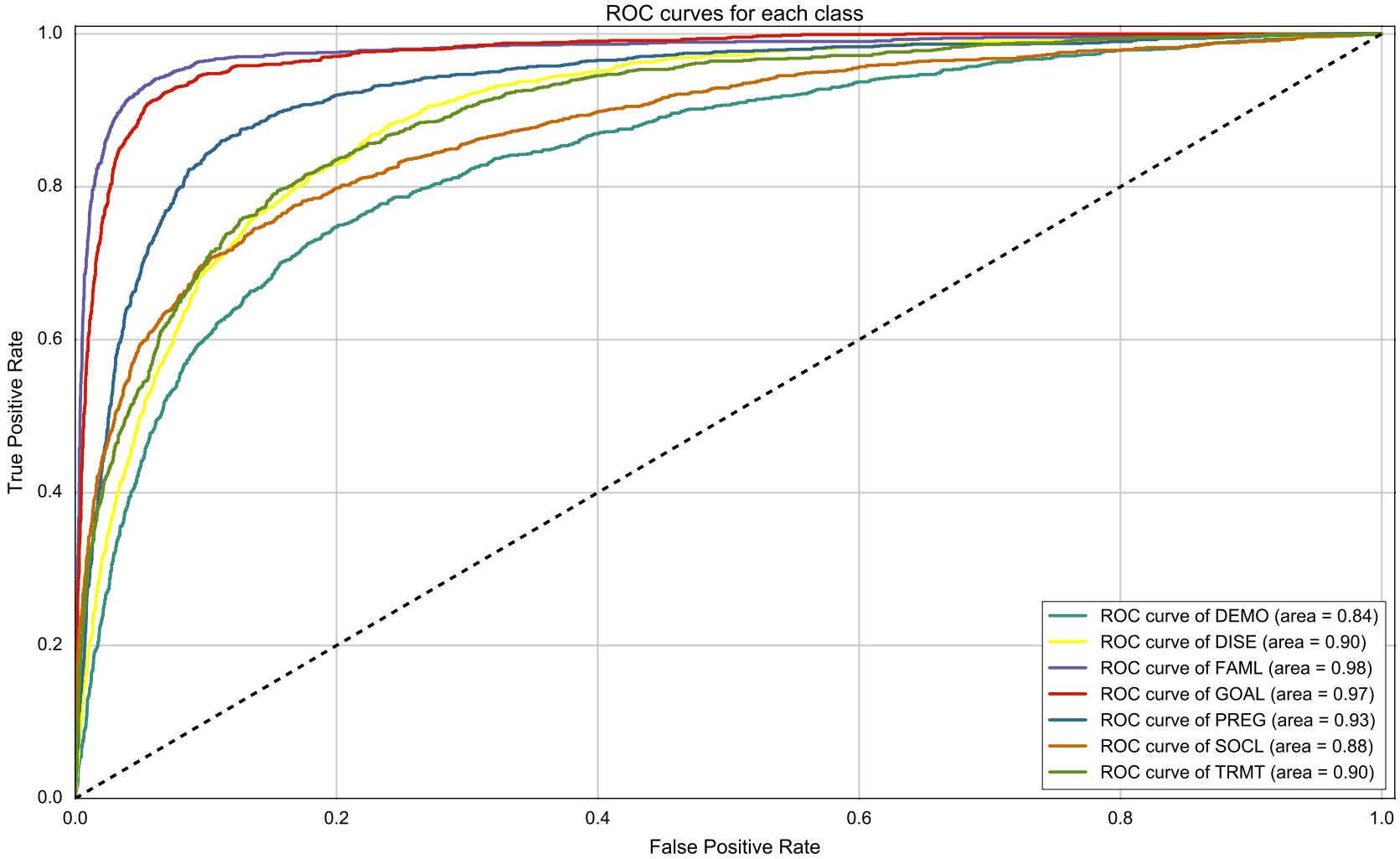
Remarks

- online learning
- parallel training

Confusion Matrix



ROC Curves


ROC curves for each class
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of DEMO (area = 0.84)
ROC curve of DISE (area = 0.90)
ROC curve of FAML (area = 0.98)
ROC curve of GOAL (area = 0.97)
ROC curve of PREG (area = 0.93)
ROC curve of SOCL (area = 0.88)
ROC curve of TRMT (area = 0.90)

Reflections & Comments

Reflections & Comments

Reflections

- Small dataset: Feasible to try different methods
- Preprocessing & hyper-parameter optimization is significant
- CV is crucial
- Deep learning didn't fly

Reflections & Comments

Reflections

- Small dataset: Feasible to try different methods
- Preprocessing & hyper-parameter optimization is significant
- CV is crucial
- Deep learning didn't fly

What Ifs

- Ensemble learning

Reflections & Comments

Reflections

- Small dataset: Feasible to try different methods
- Preprocessing & hyper-parameter optimization is significant
- CV is crucial
- Deep learning didn't fly

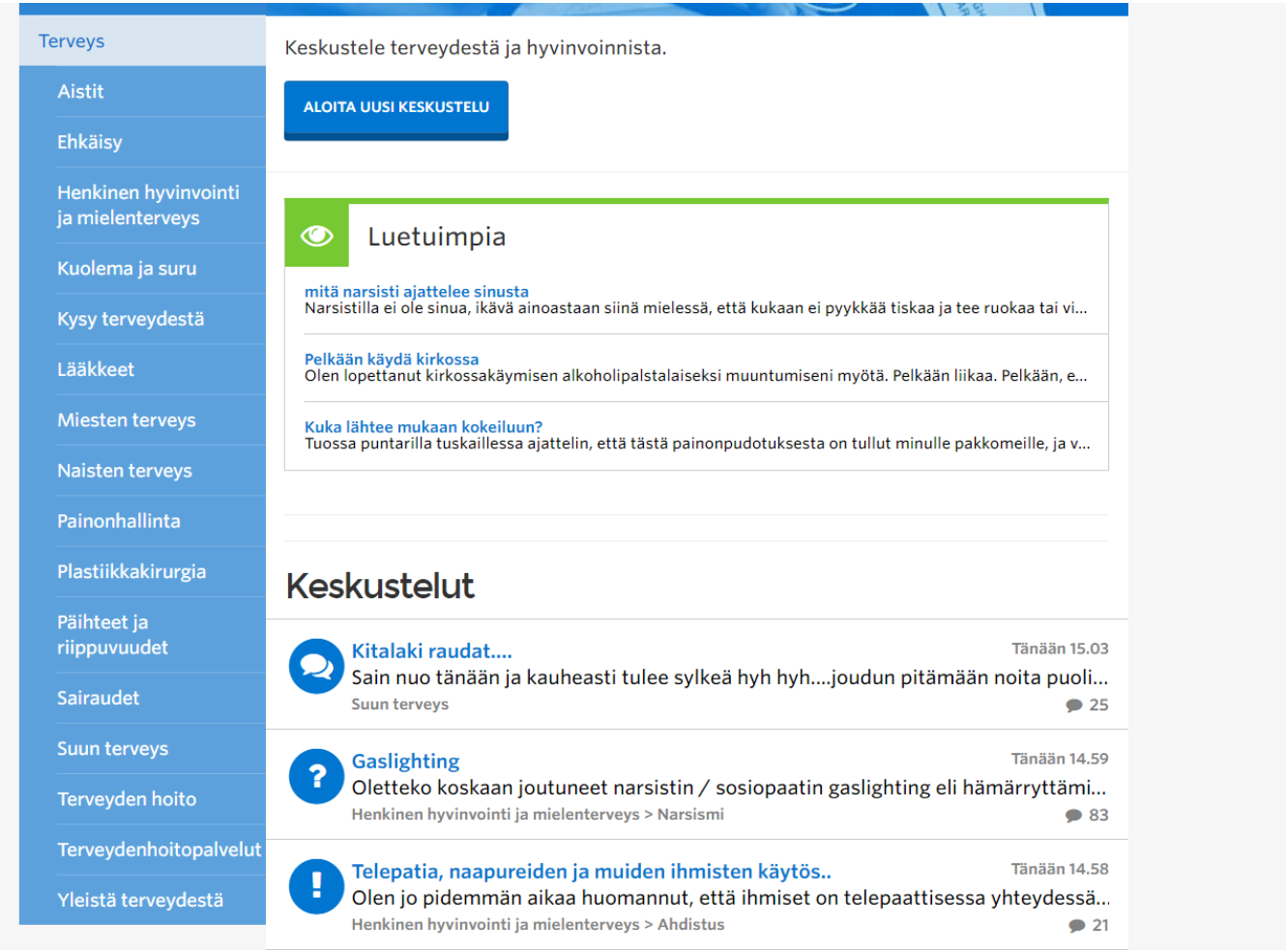
What Ifs

- Ensemble learning

Comments

- Multi-label classification

Suomi24 Discussion Forum

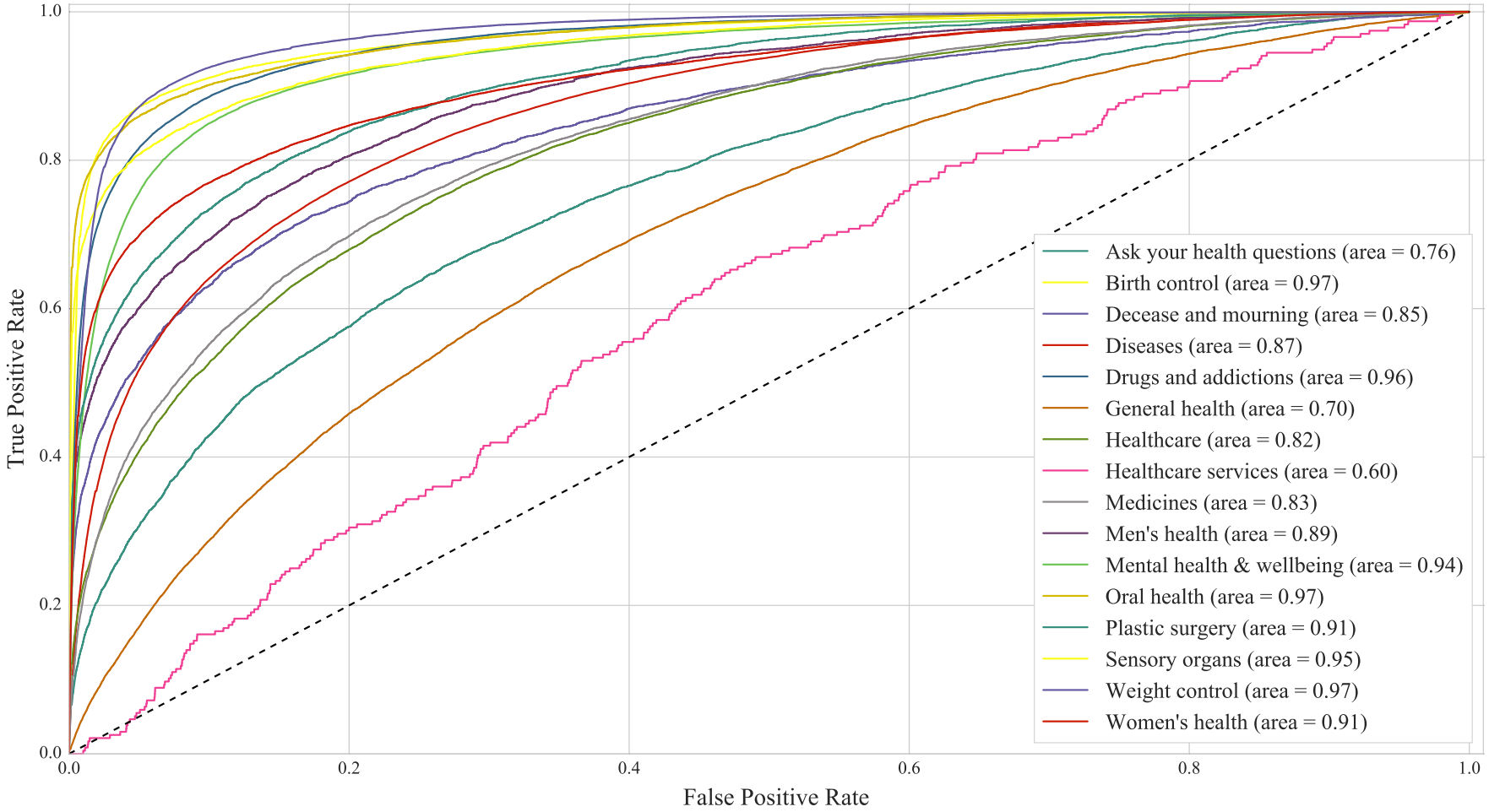


Suomi24 Discussion Forum

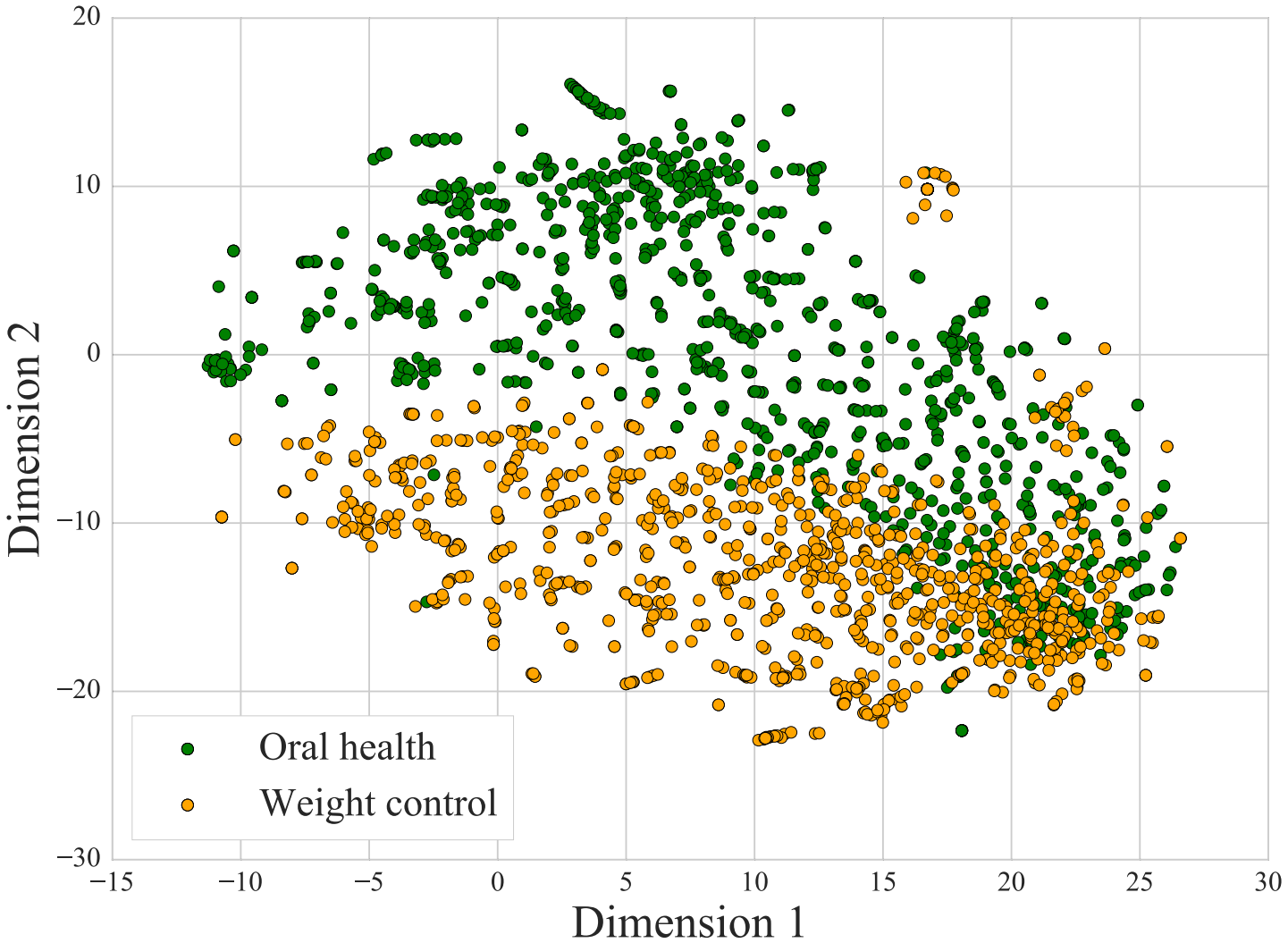
		Word Count	
Category	Number of Posts	Mean	Median
Ask your health questions	6,012	74.5	57
Birth control	13,213	58.6	47
Decease and mourning	4,108	94.3	62
Diseases	86,035	76.4	52
Drugs and addictions	34,346	68.6	35
General health	16,415	69.2	49
Healthcare	15,461	61.0	34
Healthcare services	236	61.5	35
Medicines	9,743	52.5	36
Men's health	4,067	55.8	36
Mental health & wellbeing	70,017	103.0	66
Oral health	10,959	59.3	43
Plastic surgery	6,123	55.6	39
Sensory organs	8,164	64.5	46
Weight control	49,257	69.1	47
Women's health	18,569	69.5	55

Suomi24 Discussion Forum

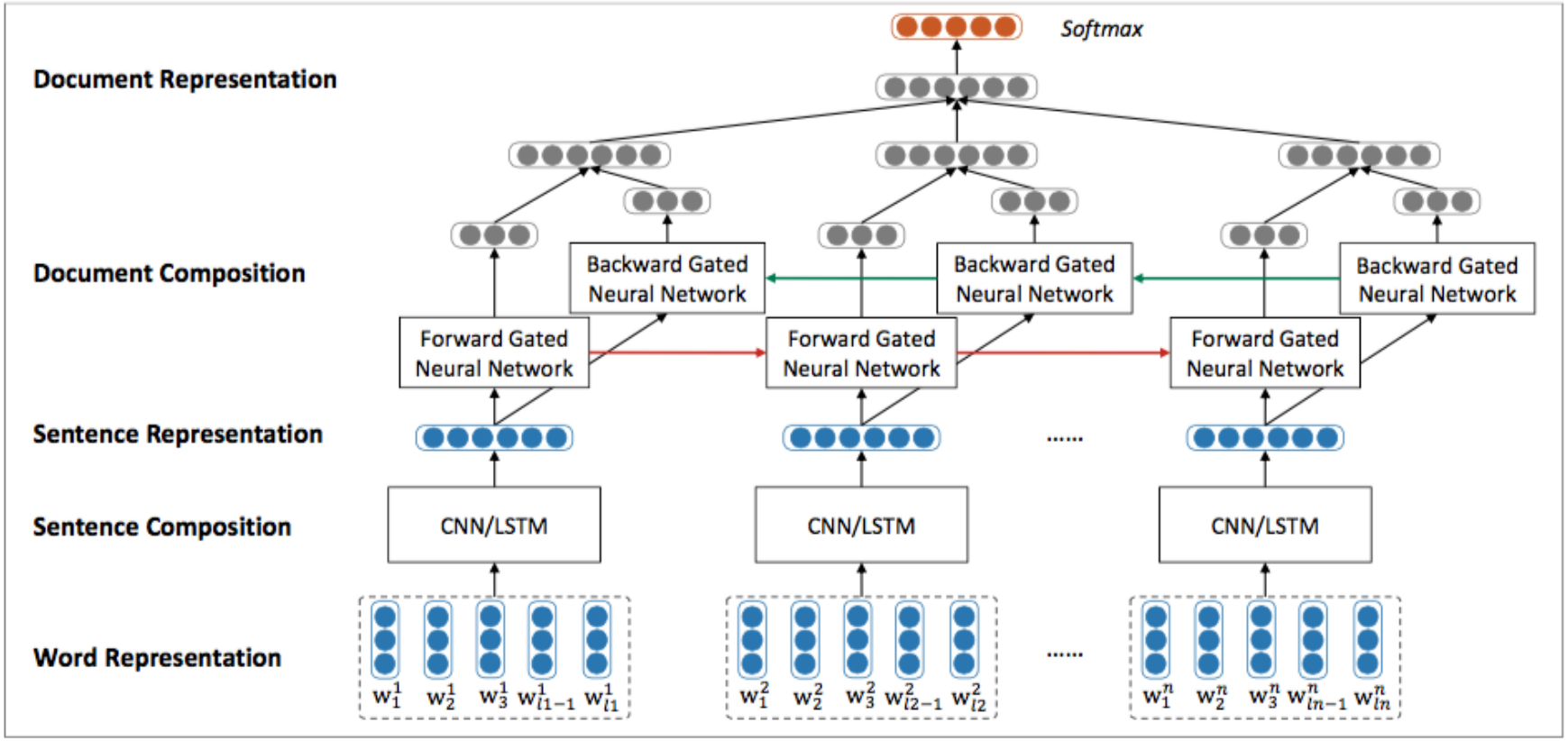
Suomi24 Discussion Forum



Suomi24 Discussion Forum



State of the Art : Deep Learning



Presentation

Oguzhan Gencoglu
ogencoglu

Co-founder & Chief Data Scientist @ Top Data Science \\ Researcher @ Tampere University of Technology

Top Data Science
Finland
oguzhan.gencoglu@topdatascie...
https://fi.linkedin.com/in/ogenc...

Overview | Repositories 27 | Stars 3 | Followers 18 | Following 0

Popular repositories — Customize your pinned repositories

ArtsyNetworks
Deep Learning + Arts
Python ★ 16 ⑂ 7

WhatsCooking
Kaggle - What's Cooking
Python ★ 5 ⑂ 3

OpenFinland
AI for Society Good
Python ★ 3

SGN41006_Spring2016
SGN 41006 - Signal Interpretation (Spring 2016)
Python ★ 3

AudioClassification
Audio Classification in Python
Python ★ 1 ⑂ 1

ladder
Forked from CuriousAI/ladder
Ladder network is a deep learning algorithm that combines supervised and unsupervised learning
Python ★ 1

THANK YOU