language and translation technology team

Nearest neighbour approaches for Emotion Detection in Tweets

FACULTY OF SCIENCES

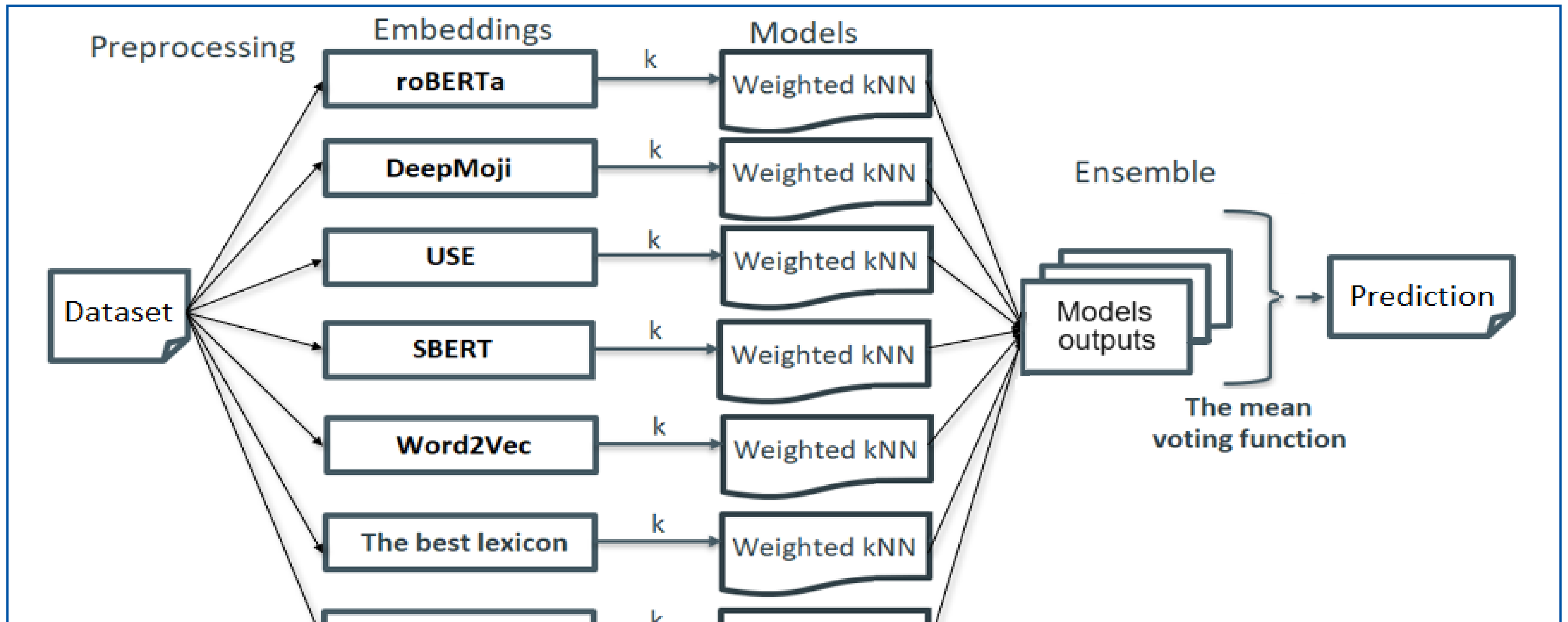
By O. Kaminska, Ch. Cornelis and V. Hoste

Introduction

- This is a classification task **SemEval-2018 Task 1** *"Affect in Tweets"* for English tweets labeled with *emotions' intensity* (anger, sadness, joy, fear).
- We aimed to use the *interpretable* Machine Learning (ML) method (the **weighted k Nearest Neighbor** (wkNN)) to compete with *Neural Networks* (NN) based solutions.

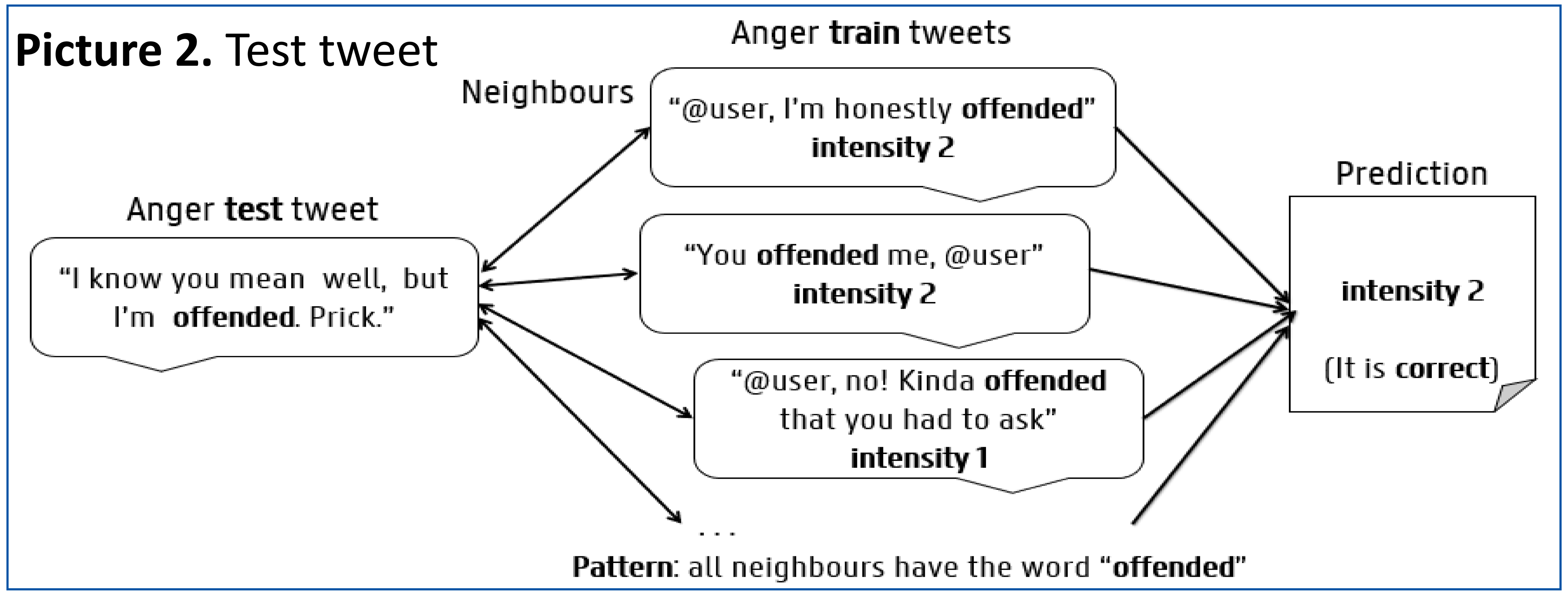
Experiments

- We examined five *text embedding* approaches' performances by tuning the best *text preprocessing* and the number of neighbours **k** for the wkNN model for each. As a similarity function for the wkNN, we used **cosine**. Moreover, we used Pearson Correlation Coefficient (**PCC**) for evaluation.
- To improve results, we appended scores from **emotional lexicon** vocabularies to the embedding vectors.
- Finally, we made the **ensemble** of *the best* embedding models with the tuned voting function applied to the models' outputs and selected embeddings to use. Our final solution is illustrated in *Picture 1*.



Picture 1. Ensemble

Picture 2. Test tweet



Results for the test data

- We checked our solution on the test data (PCC decreased compared with train) and submitted predicted labels to the competition page - we received **fourth place** with **PCC = 0.635** (*Table 1*), where top-3 solutions were NN-based.
- We evaluated several test tweets manually to examine the solution's explainability and found some patterns (*Picture 2*).

Table 1. Winners

Team	Average PCC	Anger	Fear	Joy	Sadness
#1	0.695	0.706	0.637	0.720	0.717
#2	0.653	0.670	0.588	0.686	0.667
#3	0.646	0.667	0.536	0.705	0.673
#4 (we)	**0.635**	**0.638**	**0.601**	**0.631**	**0.670**

Conclusion and the future work

- We evaluated the *explainable ML* method for the emotion detection and showed - optimizations and ensemble method *can compete* with NN solutions.
- In the future, we plan to incorporate another method for an **unbalanced** fear dataset and consider the inherently **fuzzy nature** of emotion data.

GHENT UNIVERSITY

This work was supported by the Odysseus programme of the Research Foundation – Flanders (FWO)