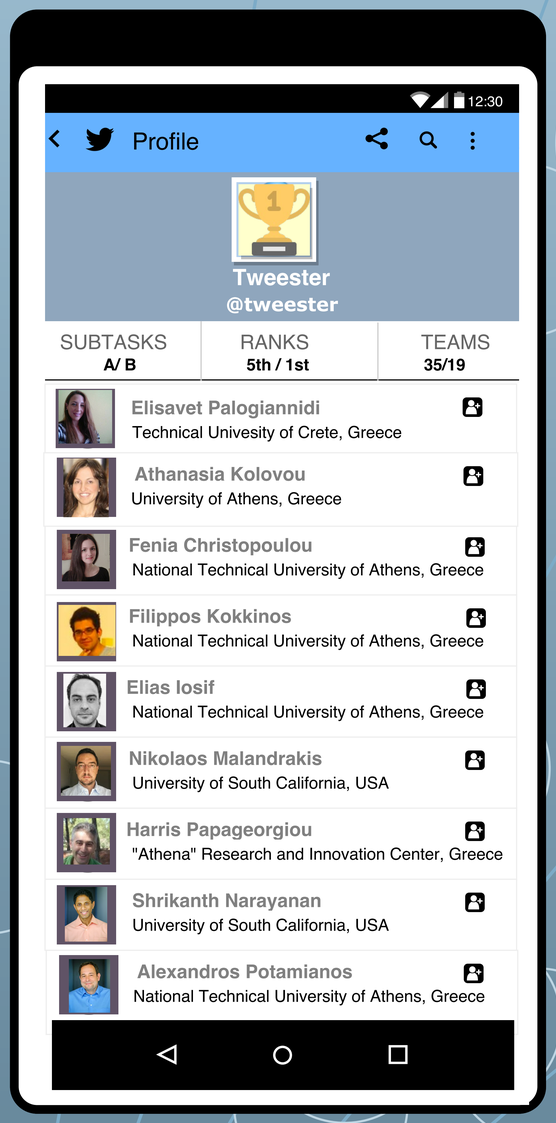
tweester

Subtask B

at SemEval 2016 Task 4

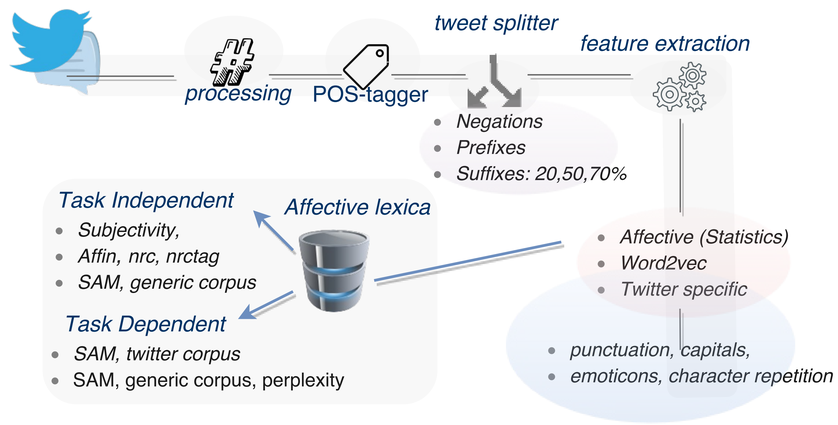
Sentiment Analysis in Twitter using Semantic-Affective Model Adaptation



Data

- General purpose data: a Web Snippet corpus (116M), ANEW
- Twitter dataset: 115M tweets (created for tweester)
- SemEval provided:
  - Substask. A: train 2013/2016
  - Substask. B: train 2016

Semantic Affective system (Baseline)



- Tools: POS-tagging, multiword expression, hashtag expansion
- *Semantic similarity implies affective similarity*: SAM *"Distributional Semantic Models for Affective Text Analysis, Malandrakis et al. 2013"*

$$\hat{v}(t_j) = a_0 + \sum_{i=1}^{N} a_i v(w_i) S(t_j, w_i)$$

($\hat{v}(t_j)$: affective rating of the unknown token; N: number of seeds; $v(w_i)$: affective ratings of seeds; a_0: bias; a_i: weights assigned to seeds; $S(t_j, w_i)$: Semantic similarity between tokens)

- Two step feature selection, Naive Bayes (NB) tree classifier

Topic Modeling - based System (TM)

- ***Adapt semantic space on each tweet***
- LDA → detect topics (16) → split corpus → a semantic model (SM) for each subcorpus → tweet-adapted semantic model $S(\cdot)$ (weighted mixture of SMs) → affective ratings

In **Subtask A** TM is used as features in Baseline and in **Subtask B** as independent system (NB tree)

Word2Vec-based System

- Relies on tweet's semantic representation
- Represent each word as vectors, and average to represent tweet
- Random Forest classifier, with tweet embedding features

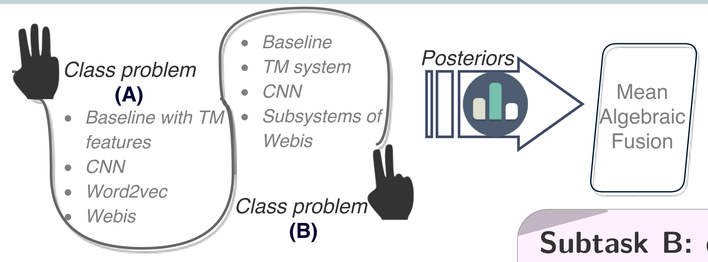
Convolutional Neural Network (CNN)

Kim et al., "*Convolutional Neural Networks for Sentence Classification*", EMNP 2014

- Tweets are represented as sentence matrices M
- M: Concatenation of tweet's word vectors $\in \mathbb{R}^{300}$
- Word vectors derive from combination of Google News dataset and a twitter corpus (115M)
- $M \rightarrow$ CN $\rightarrow$ features $\rightarrow$ max-pooling layer $\rightarrow$ soft max layer

Additional system: **Webis (W)**, SemEval 2015 (Ensemble of four classifiers)

Experimental: Subtasks A(5/35), B(1/19), D(1/15)



Ranks

SubTask	Winner	Tweester	Tweester best
A	0.633	0.608	0.624
B	0.797	0.797	0.827
D	0.034	0.027	0.027

Tweester best: using subset of the systems

Subtask B: combinations

Baseline (B) : **0.821**
TM : 0.753
CNN :0.752
B+CNN: **0.827**
B+CNN+TM:**0.818**
Without B: 0.765
Without TM: 0.78
Without CNN: **0.798**

#EndOfStory

- Creating domain relevant polarity lexica boosts performance
- New idea: Topic modeling tweet adaptation
- Focus effort on 3-class problem and systems diversity