RuG@GermEval: Offensive Speech in German Social Media

XIAOYU BAI	FLAVIO MERENDA	CLAUDIA ZAGHI	TOMMASO CASELLI	MALVINA NISSIM
x.bai.5@student.rug.nl	f.merenda@rug.nl	c.zaghi@student.rug.nl	t.caselli@rug.nl	m.nissim@rug.nl

TASK & DATA

• Automatic identification of *offensive language* in German tweets.

   offensive language = "hurtful, derogatory or ob- scene comments made by one person to another person".

• **Task 1 - coarse**: classify each tweet either as OFFENSIVE or as OTHER

• **Task 2 - fine**: classify each tweet as INSULT, ABUSE, PROFANITY, or OTHER

SHARED TASK DATA

Class	Samples
Task 1: Binary task	
OFFENSE	1,688
OTHER	3,321
Task 2: Multi-class task	
ABUSE	1,022
INSULT	595
PROFANITY	71
OTHER	3,321

ADDITIONAL TRAINING DATA

• *Political Speech Project (PSP)*: journalistic initiative on the quality of on-line political discourse in the EU.

• ***549 manually annotated offensive messages from Germany.***

• Used in Task 1 (coarse classification).

• https://rania.shinyapps.io/PoliticalSpeechProject/
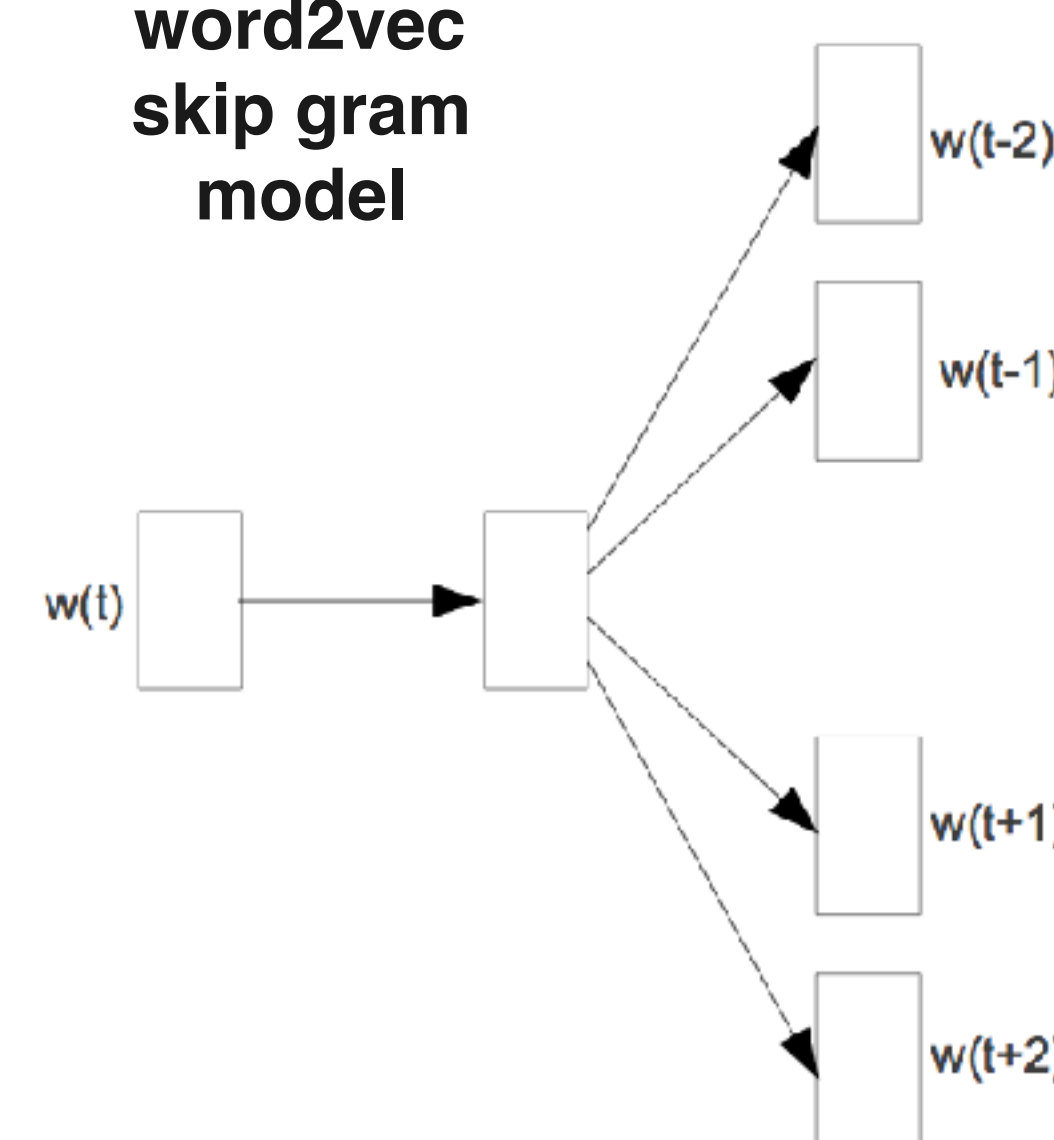
SOURCE-DRIVEN EMBEDDINGS: THE MAKING OF

WORKING HYPOTHESIS:

on-line communities represent different sources of data, and their user-generated contents can be used as proxies for specialized information (Pariser, 2011; Bozdag & van den Hoven, 2015; Seargeant & Tagg, 2018).

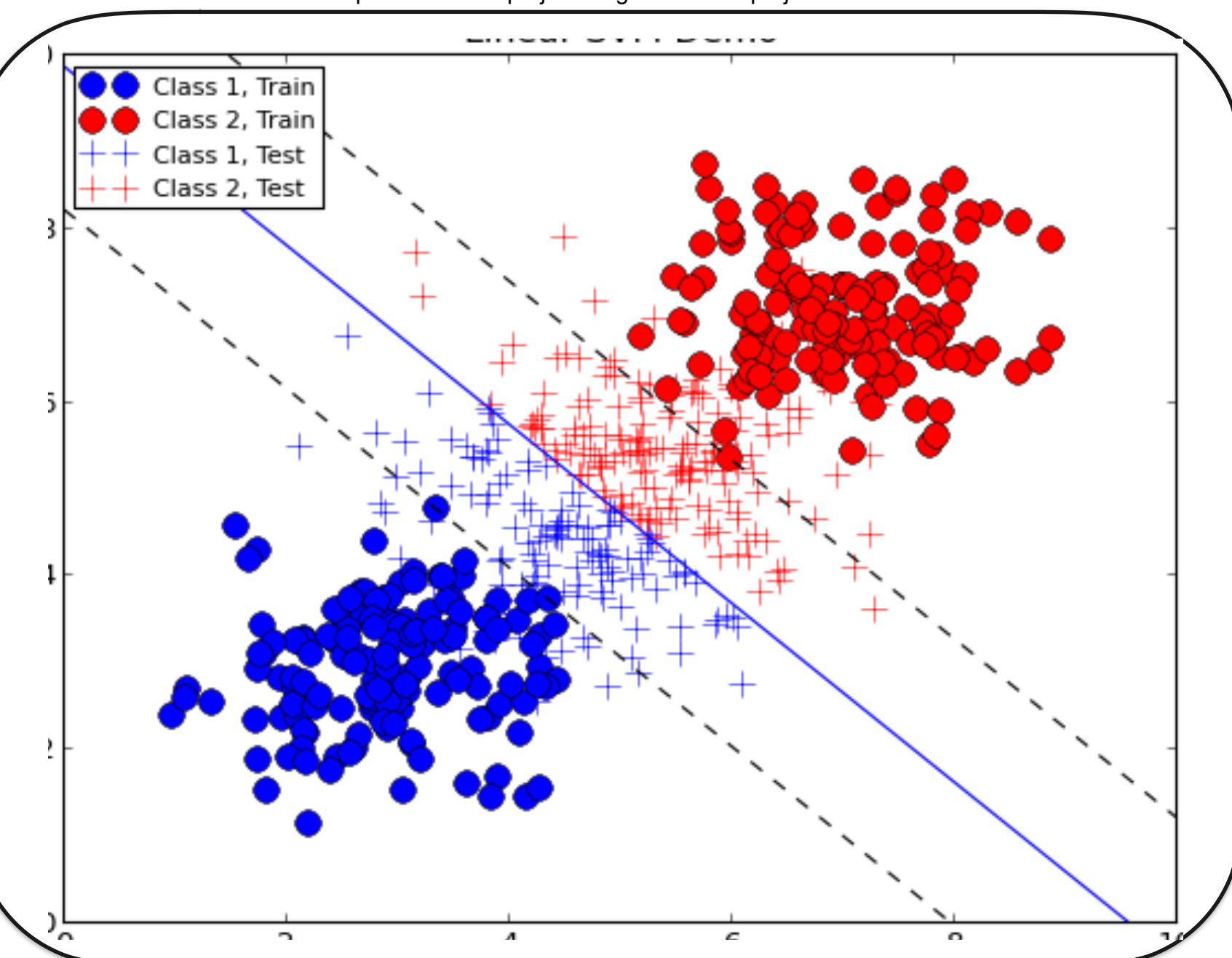| Page Name	Comments
AfD-Fraktion AGH	6,933
Alice Weidel	279,435
Asylflut stoppen	3,461
NPD - Die soziale heimatpartei	138,611
Total	**428,440**

word2vec skip gram model

OUTPUT

• *300 dimension hate embeddings* based on Facebook comments;

• *52 dimension hate embeddings* based on Facebook comments;

• 52 *dimension hate-oriented embeddings*, using information from the hate embeddings plus the shared task training data.

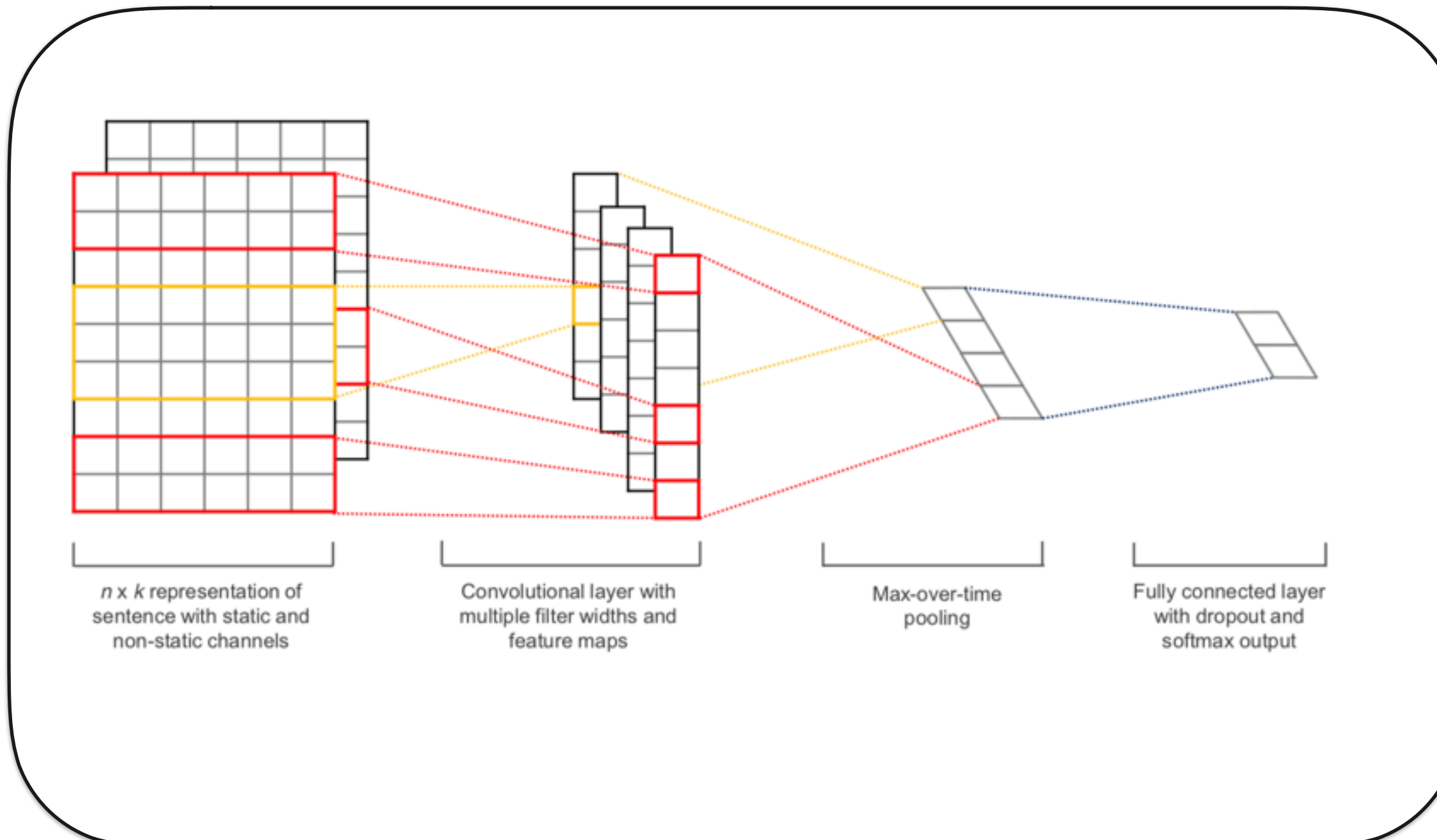## THE SYSTEMS

LINEAR SVM MODEL
http://www.mtechprojects.org/linear-svm-projects.html

• Basic pre-processing
• Word unigrams and bigrams
• Character n-grams range (3, 7)
• 52 dimension hate embeddings
• Weighted classes (both in coarse and fine-grained classification)
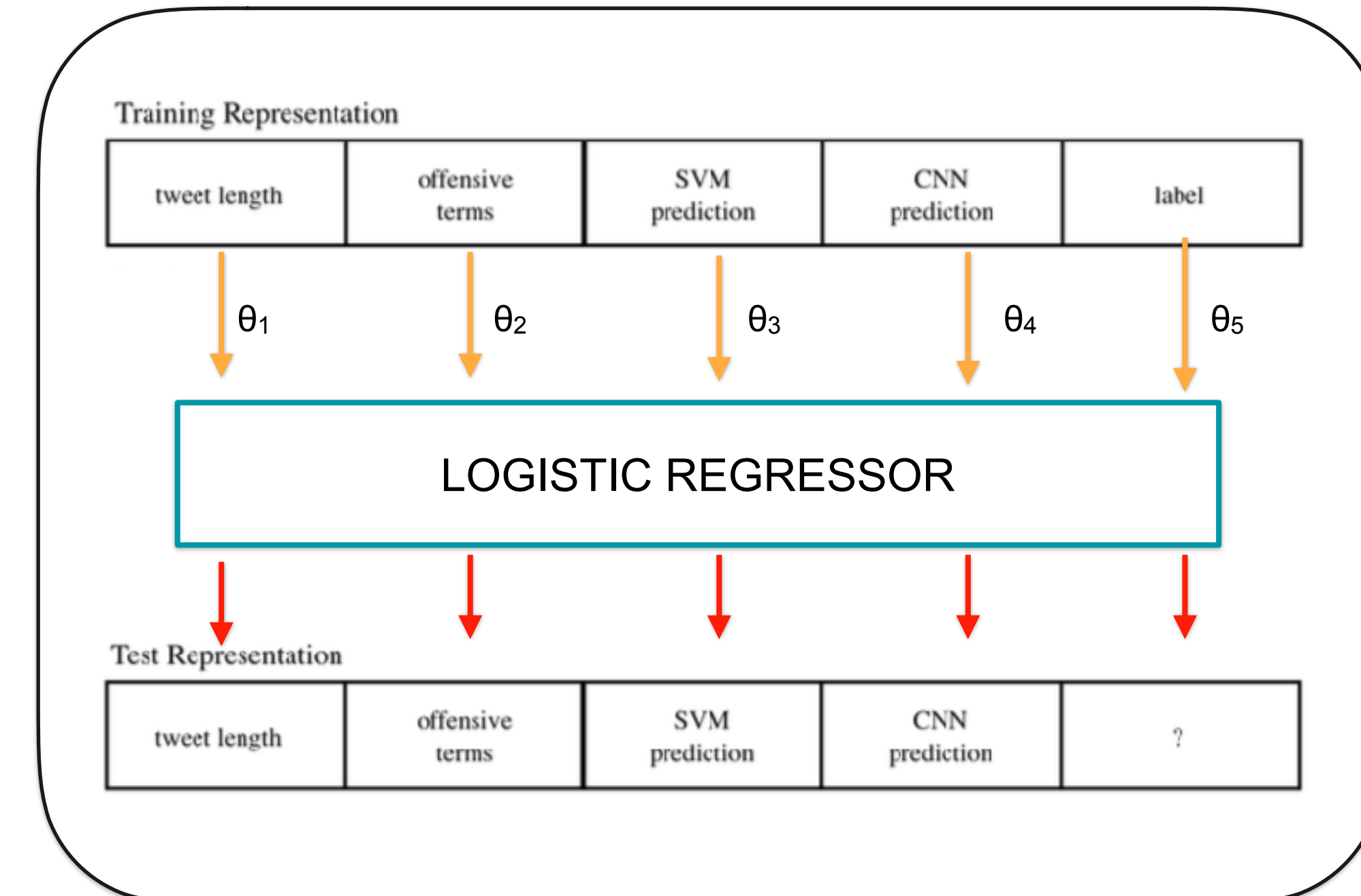
MULTI-LAYERED CNN MODEL
https://arxiv.org/pdf/1408.5882.pdf

• Basic pre-processing
• 300 dimension hate embeddings
• 1 conv layer, 1 max-pooling layer, 1 dense layer
• Number of filters: 6; Filter sizes: 3, 5, 8; Strides: 1; Activation function: ReLU; Padding: valid; Optimizer: Adam

ENSEMBLE MODEL

• Basic pre-processing
• Length of the tweet
• Presence/absence of German offensive terms, (http://www.hyperhero.com/de/insults.htm)
• SVM predictions (coarse-grained classification)
• CNN prediction (coarse-grained classification)

RESULTS AND ERRORS

	F1 (macro) - 20% Training	F1 (macro) - Test
Task 1: Binary task		
Baseline	*39.49*	*25.00*
SVM binary	71.90	68.29
CNN binary	73.05	67.80
Ensemble binary	74.45	70.38
Task 2: Multi-class task		
Baseline	*19.75*	*19.75*
SVM multi-class	40.75	39.82

• Lower performance on the OFFENSE class
• Task 1:
  - low Recall for SVM (41.76) and Ensemble (42.01)
  - low Precision for CNN (53.39)

✽ Lack of knowledge on major real-world historical and political events in Germany.
  • *Ein besetztes und belogenes Volk und Land! !LBRI Der Krieg ist nicht zu Ende!*
  • *@NetReaper Es gibt kein Palästina. Glücklicherweise.*
  • *@MartinSchulz Dürfen Jungen auch werden, was sie wollen?*
  • *NrW Minister Jäger! : Die Grosse Pfeife spricht* 😂😂😂😂😂

✽ Lack of adequate representations of emojis
  • *@yxyzyxy Schon wieder Nazi als Unwort!*😡
  • *@_tsukino_usako @allesevolution @Marenleinchen66 Deine Argumentation ist grandios. Typisch feministisch* 😁

✽ Lack of proper treatment of nouns - lowercased in pre-processing phase
  • *@GreenClaudia Grün ist Abfall und Abfall muss auf den Kompost!!!*