Comparison of Traditional Machine Learning Approach and Deep Learning Models in Automatic Cyberbullying Detection for Polish Language

Maciej Biesek

Junior NLP Engineer @ Samsung Electronics Poland R&D Center

maciejbiesek@gmail.com
https://www.linkedin.com/in/maciejbiesek/

May 31, 2019

Overview

1. Introduction
2. Dataset
   - Task 6-1
   - Task 6-2
3. Systems
   - model1-svm
   - model2-gru
   - model3-flair
4. Evaluation
5. Conclusions and future work

Introduction

A Pew Research Center survey finds that 59% of U.S. teens have personally experienced at least one of six types of online harassment [Anderson 2018]:

- offensive name-calling,
- spreading of false rumors,
- receiving explicit images they did not ask for,
- having explicit images of them shared without their consent,
- psychical threats,
- constant asking where they are and what are they doing by someone other than a parent.

Dataset

Dataset used to train and evaluate presented systems contains tweets collected from openly available Twitter discussions with applied anonymization of posts.
Preprocessing:

- removed retweets (duplicated tweets with *RT* tag at the begging) from the training set, in case of testing one only *RT* tag was deleted, but duplicated tweets were preserved,
- each post was cleaned from whitespaces (\r, \n, \t), digits, punctuation marks, emojis and *@anonymized_account* tags,
- lowercasing and tokenization using spaCy tool,
- deleted from the training set posts with zero or only one token,
- for **model1-svm** stopwords were removed, also using spaCy tool

Task 6-1

- **goal:** distinguish between neutral tweets (class 0) and those which contain any kind of harmfulness (class 1), including cyberbullying, hate speech and so on,
- binary classification,
- highly imbalanced.

Type of tweets	**Training dataset**	**Testing dataset**
Neutral	8607	866
Harmful	755	134

Task 6-2

- **goal:** classify tweets into three classes: neutral (0), cyberbullying (1) and hate speech (2),
- cyberbulling →harmful action is addressed towards private persons, hate speech →to a public person, entity or large group,
- imbalanced

Type of tweets	**Training dataset**	**Testing dataset**
Neutral	8607	866
Cyberbullying	243	25
Hate speech	512	109

Systems

- In both tasks posts are the same (in Task 6-2 those previously annotated as harmful were further divided into cyberbullying and hate speech), so architectures of three presented models are the same in both cases except the number of output classes,
- The whole code of models is publicly available `https://github.com/maciejbiesek/poleval-cyberbullying`.

model1-svm

- **input:** tokenized dataset with stopwords removed,
- **pipeline:**
  - documents are converted to a matrix of token counts,
  - transformed to the TF-IDF representation,
  - on these features linear SVM classifier is trained,
- built using scikit-learn tool.

model2-gru

- **input:** sequences of tokens with maximum length of 20 (short sentences are padded and longer ones trimmed) mapped to the FastText 300-dimensional embedding matrix,
- **architecture:**
  - bidirectional GRU (128 units in both directions),
  - dense network (2 layers with ReLU activation function, intermediate of size 50 with 0.5 dropout and 2 or 3 neurons as an output) on the top of concatenated final states from forward and backward passes,
  - to deal with class imbalancement weighted softmax cross entropy is used as a loss function, optimized using Adam (with default value of learning rate),
- implemented using Tensorflow library.

model3-flair

- **input:** FastText word embeddings stacked with forward and backward character-level language models (Contextual String Embeddings) trained on 1B words corpus of Polish [Borchmann 2018],
- **architecture:**
  - bidirectional GRU network with hidden size of 128,
  - linear classifier.
- implemented using Flair framework.

Evaluation

- baseline1 →vector with random values sampled from the set $\{0, 1\}$ in Task 6-1 and $\{0, 1, 2\}$ in Task 6-2,
- baseline2 →vector filled with zeros, as *neutral* is the most frequent class in both tasks.

Task 6-1

Model	**Prec**	**Rec**	**F-score**	**Acc**
baseline1	13.65	**47.76**	21.23	52.50
baseline2	0.00	0.00	0.00	86.60
model1-svm	60.49	36.57	**45.58**	**88.30**
model2-gru	63.83	22.39	33.15	87.90
model3-flair	**81.82**	13.43	23.08	88.00

Task 6-2

Model	**Micro F-score**	**Macro F-score**
baseline1	31.60	33.28
baseline2	86.60	30.94
model1-svm	**87.60**	**51.75**
model2-gru	78.80	49.15
model3-flair	86.80	45.05

Conclusions and future work

- SVM outperforms deep learning approaches in both tasks. Probably the reason of that is relatively small imbalanced dataset – more complicated networks can easily overfit in such case. Proposed solution for it would be acquiring more data with harmful examples,
- the assumption behind Flair model was that rich representation of input data would lead to high results in classification – probably the weakest link there was an easy linear classifier, using more sophisticated one (eg. dense multilayer network) could be beneficial,
- it has been proved that convolutional neural networks perform remarkably well in sentence classification and even in cyberbullying detection [Ptaszynski 2017],
- [Ptaszynski 2017] shows that incorporating other features except words/tokens (NER, POS-tagging) is usefull in cyberbullying detection.

References

Biesek M. (2019)
Comparison of Traditional Machine Learning Approach and Deep Learning Models in Automatic Cyberbullying Detection for Polish Language
Proceedings of the PolEval 2019 Workshop

Anderson M. (2018)
A Majority of Teens Have Experienced Some Form of Cyberbullying
https://www.pewinternet.org/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/

Borchmann Ł., Gretkowski A. and Graliński F. (2018)
Approaching Nested Named Entity Recognition with Parallel LSTM-CRFs
Proceedings of the PolEval 2018 Workshop

Ptaszynski M., Eronen J. K. K. and Masui F (2017)
Learning Deep on Cyberbullying is Always Better Than Brute Force

Questions