Using Knowledge Graphs to improve Hate Speech Detection

Poojitha Maheshappa
IIT Kharagpur
poojitha.maheshappa@iitkgp.ac.in

Binny Mathew
IIT Kharagpur
binnymathew@iitkgp.ac.in

Punyajoy Saha
IIT Kharagpur
punyajoys@iitkgp.ac.in

ABSTRACT

With the increasing cases of online hate speech, there is an urgent demand for better hate speech detection systems. In this paper, we utilize Knowledge Graphs (KGs) to improve hate speech detection. Our initial results shows that incorporating information from KG helps the classifier to improve the performance.

CCS CONCEPTS

• **Social and professional topics → Hate speech**; • **Computing methodologies → Knowledge representation and reasoning**.

KEYWORDS

Knowledge Graphs, Hate Speech, Text Classification

ACM Reference Format:

Poojitha Maheshappa, Binny Mathew, and Punyajoy Saha. 2021. Using Knowledge Graphs to improve Hate Speech Detection. In *8th ACM IKDD CODS and 26th COMAD (CODS COMAD 2021), January 2–4, 2021, Bangalore, India.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3430984.3431072

1 INTRODUCTION

Online hate speech has resulted in several gruesome scenarios. This has led to an increasing interest among reserachers to develop better systems. In recent studies, knowledge integration has proven to improve performance on NLP tasks due it's ability to enhance the model with semantic information [3]. We attempt to demonstrate how knowledge integration, through KGs, would enhance the performance of text based models for hate seech classification.

2 APPROACH

For our work, we have chosen to work with the English subset of the hatEval dataset [1]. It is a binary classification problem where we predict whether a tweet in English is hateful or not hateful. 42% of the examples are hateful in the training, development and test sets. We use, like in the competition, the macro-averaged F1-score as the evaluation metric. As part of the preprocessing, the URLs, Hashtags and Mentions were normalized.

In our approach, we first create a Knowledge Graph from the training data. In order to do this, we use a spot-entity annotator TagMe [2] to extract the entities from the preprocessed training data. The annotator identifies meaningful substrings, called "spots" and links them to the relevant DBpedia entities; corresponding subject-predicate-object triplets are queried using SPARQL. These triplets are used to create our one-hop KG, initial assumption being the nodes are connected to each other by only one type of relation. We use two models as part of our baseline. First, an LSTM model using pre-trained fastText word vectors Second, an LSTM Model with node2vec embeddings of the entities in the text as input.

model.pdf

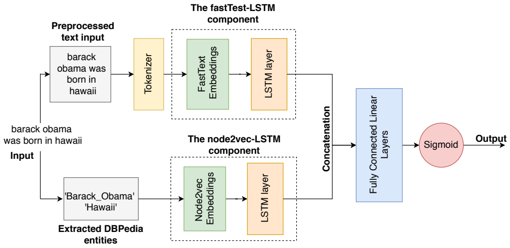
Preprocessed text input
barack obama was born in hawaii
Tokenizer
The fastTest-LSTM component
FastText Embeddings
LSTM layer
barack obama was born in hawaii
Input
The node2vec-LSTM component
'Barack_Obama' 'Hawaii'
Extracted DBPedia entities
Node2vec Embeddings
LSTM layer
Concatenation
Fully Connected Linear Layers
Sigmoid
Output

Figure 1: **Overall Architecture of our Hate-KG model.**

Method	Macro F1-Score
fastText + LSTM	0.591
node2vec + LSTM	0.548
Hate-KG	**0.618**

Table 1: Results in terms of F1-scores.

The overall architecture of our *Hate-KG* model is shown in figure 1. The first component comprises an LSTM layer preceded by the fastText embedding layer. The second component comprises an LSTM layer preceded by the node2vec embedding layer. The output of these two components are concatenated and passed through two fully connected linear layers. The input to the fastText-LSTM component is the tokenized sentences where as the input to the node2vec-LSTM component is the corresponding sequence of extracted entities.

3 RESULTS

Table 1 shows the performance of the models. The node2vec+LSTM model performed poorly when compared to other models. our **Hate-KG** had the best performance among all the models. As compared to the text based model, Hate-KG had a relative improvement of 4.5% in F1-Score. We plan to use knowledge graph enhanced transformers [3] (ERNIE) and relation graphs to improve the performance in the future.

REFERENCES

[1] Valerio Basile, C. Bosco, E. Fersini, Debora Nozza, V. Patti, F. Pardo, P. Rosso, and M. Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *SemEval@NAACL-HLT*.

[2] Paolo Ferragina and Ugo Scaiella. 2010. TAGME: On-the-Fly Annotation of Short Text Fragments (by Wikipedia Entities) *(CIKM '10)*.

[3] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. arXiv:1905.07129 [cs.CL]

CODS COMAD 2021, January 2–4, 2021, Bangalore, India

ACM ISBN 978-1-4503-8817-7/21/01.
https://doi.org/10.1145/3430984.3431072