

# Hate Speech Detection with Comment Embeddings

Nemanja Djuric, Jing Zhou, Robin Morris,  
Mihajlo Grbovic, Vladan Radosavljevic, Narayan Bhamidipati  
Yahoo Labs, 701 First Ave, Sunnyvale CA, USA  
{nemanja, jingzh, rdm, mihajlo, vladan, narayanb}@yahoo-inc.com

## ABSTRACT

We address the problem of hate speech detection in online user comments. Hate speech, defined as an “abusive speech targeting specific group characteristics, such as ethnicity, religion, or gender”, is an important problem plaguing websites that allow users to leave feedback, having a negative impact on their online business and overall user experience. We propose to learn distributed low-dimensional representations of comments using recently proposed neural language models, that can then be fed as inputs to a classification algorithm. Our approach addresses issues of high-dimensionality and sparsity that impact the current state-of-the-art, resulting in highly efficient and effective hate speech detectors.

## 1. INTRODUCTION

In the age of ever-increasing volume and complexity of the internet, millions of users have unrestricted access to vast amounts of content that allows for privileges unimaginable several decades ago, such as access to knowledge bases or latest news within just a few clicks. However, due to internet’s non-restrictive nature and, in certain countries, legal protection of free speech which also includes hate speech [4], some users misuse the medium to promote offensive and hateful language, which mars experience of regular users, affects business of online companies, and may even have severe real-life consequences [1]. To mitigate these detrimental effects, many companies (including Yahoo, Facebook, and YouTube) strictly prohibit hate speech on websites they own and operate, and implement algorithmic solutions to discern hateful content. However, scale and multifacetedness of the task renders it a difficult endeavour, and hate speech still remains a problem in online user comments.

Curiously, despite prevalence and large impact of online hate speech, to the best of our knowledge there exist only a few published works addressing this problem. In [1] (see also references therein) authors extract linguistic and bag-of-words (BOW) features and explore several classifiers to detect hateful tweets following the 2013 incident in Wool-

wich, UK. In [6] authors use BOW representation of user comments and train Support Vector Machine to filter anti-semitic content. Motivated by [6], authors of [2] use BOW and Naïve Bayes to flag racist comments. Interestingly, in all these works authors comment on limitations of BOW-based representation of text. This especially holds in the context of hate speech where offenders often use simple yet effective tricks to obfuscate their comments and make them more difficult for automatic detection (such as replacing or removing characters of offensive words), while still keeping the intent clear to a human reader. This results in high-dimensionality and large sparsity of the problem, making models susceptible to overfitting [6]. To address these issues, in this work we propose an approach that learns low-dimensional, distributed representations of user comments, allowing for efficient training of effective hate speech detectors.

We note that the task is different from, albeit related to, sentiment analysis [5] as there are no shades of hate speech and, unlike hate speech, even negative sentiment provides useful and actionable insights. Related work also includes attempts to remove offensive words without modifying the underlying meaning of comments [7]. This approach is however not applicable to hate speech detection as the conveyed message itself is considered harmful and should be removed.

## 2. PROPOSED APPROACH

We propose a two-step method for hate speech detection. First, we use paragraph2vec [3] for joint modeling of comments and words, where we learn their distributed representations in a joint space using the continuous BOW (CBOW) neural language model. This results in low-dimensional text embedding, where semantically similar comments and words reside in the same part of the space. Then, we use the embeddings to train a binary classifier to distinguish between hateful and clean comments. During inference, for newly observed comment, we infer representation by “folding in” using already learned word embeddings, as detailed in [3].

### 2.1 Neural language model

Neural language models take advantage of word order, and state the same assumption of  $n$ -gram language models that words that are close in a sentence are also statistically more dependent. In this work, we use the CBOW model as a component of paragraph2vec [3], which, based on the surrounding words, tries to predict the central word, as well as the user comment the words belong to.

More formally, let us assume we are given a set  $\mathcal{D}$  of  $M$  documents,  $\mathcal{D} = \{d_1, d_2, \dots, d_M\}$ , where each document  $d_m$

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). Copyright is held by the author/owner(s).  
WWW 2015 Companion, May 18–22, 2015, Florence, Italy.  
ACM 978-1-4503-3473-0/15/05.  
<http://dx.doi.org/10.1145/2740908.2742760>.

