



This is a repository copy of *Utilizing subjectivity level to mitigate identity term bias in toxic comments classification*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/186277/>

Version: Accepted Version

---

**Article:**

Zhao, Z. [orcid.org/0000-0002-3060-269X](https://orcid.org/0000-0002-3060-269X), Zhang, Z. [orcid.org/0000-0002-8587-8618](https://orcid.org/0000-0002-8587-8618) and Hopfgartner, F. [orcid.org/0000-0003-0380-6088](https://orcid.org/0000-0003-0380-6088) (2022) Utilizing subjectivity level to mitigate identity term bias in toxic comments classification. *Online Social Networks and Media*, 29. 100205.

<https://doi.org/10.1016/j.osnem.2022.100205>

---

© 2022 Elsevier B.V. This is an author produced version of a paper subsequently published in *Online Social Networks and Media*. Uploaded in accordance with the publisher's self-archiving policy. Article available under the terms of the CC-BY-NC-ND licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>).

**Reuse**

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.



[eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk)  
<https://eprints.whiterose.ac.uk/>

# Utilizing Subjectivity Level to Mitigate Identity Term Bias in Toxic Comments Classification

Zhixue Zhao<sup>a</sup>, Ziqi Zhang<sup>b</sup> and Frank Hopfgartner<sup>c</sup>

University of Sheffield, 211 Portobello, Sheffield, S1 4DP, United Kingdom

---

## ARTICLE INFO

*Keywords:*  
Language model  
Transfer learning  
Hate speech  
Classification

## ABSTRACT

Toxic comment classification models are often found biased towards identity terms, i.e., terms characterising a specific group of people such as “Muslim” and “black”. Such bias is commonly reflected in false positive predictions, i.e., non-toxic comments with identity terms. In this work, we propose a novel approach to debias the model in toxic comment classification, leveraging the notion of subjectivity level of a comment and the presence of identity terms. We hypothesize that toxic comments containing identity terms are more likely to be expressions of subjective feelings or opinions. Therefore, the subjectivity level of a comment containing identity terms can be helpful for classifying toxic comments and mitigating the identity term bias. To implement this idea, we propose a model based on BERT and study two different methods of measuring the subjectivity level. The first method uses a lexicon-based tool. The second method is based on the idea of calculating the embedding similarity between a comment and a relevant Wikipedia text of the identity term in the comment. We thoroughly evaluate our method on an extensive collection of four datasets collected from different social media platforms. Our results show that: 1) our models that incorporate both features of subjectivity and identity terms consistently outperform strong SOTA baselines, with our best performing model achieving an improvement in F1 of 4.75% over a Twitter dataset; 2) our idea of measuring subjectivity based on the similarity to the relevant Wikipedia text is very effective on toxic comment classification as our model using this has achieved the best performance on 3 out of 4 datasets while obtaining comparative performance on the remaining dataset. We further test our method on RoBERTa to evaluate the generality of our method and the results show the biggest improvement in F1 of up to 1.29% (on a dataset from a white supremacist online forum).


---

## 1. Introduction

Combating toxic comments online is an important area of research nowadays (Salawu, He and Lumsden, 2017; Fortuna and Nunes, 2018; Liu, Burnap, Alorainy and Williams, 2019a; Pamungkas, Basile and Patti, 2021a). Toxic Comment Classification (TCC) is commonly handled as a text classification task. TCC has taken years of research, moving from the earlier methods based on feature engineering with classic machine learning algorithms to Deep Neural Network (DNN)-based methods that automatically learn feature representations from data, to pre-trained language model-based methods that enrich feature representations by using large-scale unlabelled corpora. However, several studies have revealed bias in SOTA methods for TCC tasks, especially bias towards identity terms, known as identity term bias (Park, Shin and Fung, 2018; Dixon, Li, Sorensen, Thain and Vasserman, 2018; Kennedy, Jin, Davani, Dehghani and Ren, 2020; Pamungkas, Basile and Patti, 2021b). Identity terms are words or terms referring to specific groups of people, such as “Muslim”, “black”, “women” and “democrat”. The identity term bias is often associated with false positive bias (Dixon et al., 2018), because when an identity term appears in a non-toxic comment, the model tends to classify it as a toxic comment.

Limited studies have attempted to handle such bias and those methods follow a simple principle: ignoring or paying less attention to the identity terms. However, this overlooks the fact that identity terms can be essential and important features to make predictions. In this work, we explore a new approach: when an identity term appears in a comment, we ask the model to incorporate the level of “subjectivity” in the prediction. A comment with a low subjectivity level expresses more factual information and less personal feelings and opinions; while a comment with a high subjectivity level contains more personal opinions but less factual information. Intuitively, when a person discusses a specific group of people, we usually do not ignore the identity term referring to them. At the same time, we might pay more attention

---

 zhixue.zhao@sheffield.ac.uk (Z. Zhao); ziqi.zhang@sheffield.ac.uk (Z. Zhang); f.hopfgartner@sheffield.ac.uk (F. Hopfgartner)  
ORCID(s): 0000-0002-3060-269X (Z. Zhao)

to the overall subjectivity level of the comment to understand the speaker’s attitude towards them. We hypothesize that when a toxic comment is made about a group of people with an identity term, it is more likely to have a high level of subjectivity. Therefore, the likelihood of a comment being toxic can be associated with the two mutually dependent conditions: 1) whether the comment contains an identity term, and 2) the subjectivity level of that comment.

Building on the BERT model which is considered state of the art for many downstream NLP tasks, we propose a novel structure, Subidentity-Sensitive BERT (SS-BERT), where “Subidentity” denotes “subjectivity” and “identity”. SS-BERT makes use of the special embedding structure of BERT to “activate” the subjectivity features only when the comment contains identity terms. The idea behind it is to utilize the relationship among the comment toxicity, the subjectivity level of the comment and the presence of identity terms in the comment to debias the model and therefore to mitigate the identity term bias in the model prediction. In measuring the subjectivity level, we study two options. One is based on a SOTA tool for calculating subjectivity using a lexicon based approach. The other one is based on our idea of calculating the similarity between a comment containing an identity term and the Wikipedia summary text (to be defined later) regarding the identity term, and using the similarity as a proxy to “subjectivity”. Last but not least, we also adapt our method to RoBERTa to further validate the generalisation of our method.

We compare our proposed method against SOTA baselines (BERT and BERT+SOC (Devlin, Chang, Lee and Toutanova, 2019; Kennedy et al., 2020)) and a few alternative models designed for ablation analysis. We evaluate all models and methods on a wide range of TCC tasks with different dataset sizes, different text lengths and from different social media platforms. The results show that our method is able to mitigate identity term bias and improve toxic comment classification effectively. First, SS-BERT consistently outperforms BERT, regardless of how subjectivity is measured. The performance gain of SS-BERT is mainly attributed to predicting fewer false positives. This indicates that our idea of considering subjectivity and the presence of identity terms is helpful to mitigate the false positive bias, i.e., identity term bias. Second, SS-BERT consistently outperforms its alternative model SO-BERT (Subjectivity-Only BERT), which only uses subjectivity without considering the presence of identity terms. This indicates that simply learning the subjectivity level for all comments is not enough, and it is more informative to combine subjectivity with the presence of identity terms. Third, our Wikipedia-similarity based proxy to subjectivity is shown to be more effective than the SOTA lexicon-based tool in our TCC experiments as models based on this measure of subjectivity have outperformed their counterparts on three out of four tasks. This suggests that given a comment containing an identity term, comparing the meaning of that comment against a reference text describing that identity term can, to some extent, reflect the subjectivity level of the comment.

Our main contributions are as follows:

- We identify the bias towards identity terms found in BERT on toxic comment classification tasks, and the relationship between such bias and the subjectivity level of comments on a wide range of datasets.
- We propose and verify a hypothesis for identity term bias and subjectivity which is intuitively explainable.
- We introduce a novel BERT-based model which is able to incorporate the subjectivity level given the presence of identity terms to handle the identity term bias. Our model achieves consistent improvements across a range of datasets.
- We explore an idea to use a proxy for the subjectivity level of a comment in our method. The proxy is based on the similarity of the text and an “objective” reference text selected from Wikipedia.

The remaining part of the paper proceeds as follows: Section 2 gives an overview of related work; Section 3 analyses the identity term bias in the TCC tasks; Section 4 explains our proposed model SS-BERT and two different methods of measuring subjectivity; Section 5 presents the implementation details and results of our experiments; Section 6 summarises this work.

## 2. Related work

### 2.1. Toxic comment classification models

By “toxic comment”, we generally refer to different types of negative, unhealthy or disrespectful user-generated-content, which includes but is not limited to hate speech, abusive language, cyberbullying, etc (Kwok and Wang, 2013; Chavan and Shylaja, 2015; Burnap, Rana, Avis, Williams, Housley, Edwards, Morgan and Sloan, 2015). Toxic comment classification is commonly handled as a text classification task. In earlier studies, the majority of research

tackle TCC as a binary classification problem that distinguishes one particular type of toxic comment from all the others (Fortuna and Nunes, 2018). Some later studies frame toxic comment classification as multi-class classification tasks or multi-label classification tasks. In multi-class classification, a comment will be assigned into one of the multiple types of toxic comments. In multi-label classification tasks, a comment could be assigned into none or at least one (could be multiple) types of toxic comment.

Early research on TCC primarily makes use of traditional machine learning algorithms, such as logistic regression, Support Vector Machines and Naïve Bayes (Schmidt and Wiegand, 2017). These statistical machine learning methods work with manually engineered features such as N-gram, TF-IDF, lexical resources, linguistic features and sentiment analysis (Gitari, Zuping, Damien and Long, 2015; Nobata, Tetreault, Thomas, Mehdad and Chang, 2016; Schmidt and Wiegand, 2017). The last decade has seen much more improvement and exploration of deep neural networks (DNN)-based models in the task of text classification (Kumar, Dabas, Jain and Pawar, 2021; Kovács, Alonso and Saini, 2021). DNN-based models learn abstract features itself during training. In the context of TCC, popular DNN models include Convolutional Neural Networks, Recurrent Neural Networks, Bi-directional Long Short-Term Memory Networks, and hybrid neural networks which combines different DNN configurations (Kim, 2014; Del Vigna, Cimino, Dell’Orletta, Petrocchi and Tesconi, 2017; Yang, Yang, Dyer, He, Smola and Hovy, 2016; Schmidt and Wiegand, 2017; Zhang, Robinson and Tepper, 2018b).

As identified by Zhao, Zhang and Hopfgartner (2019), more training data is helpful for improving the performance of a TCC model. Unfortunately, task-specific training data is often scarce and expensive to create (Emmery, Verhoeven, De Pauw, Jacobs, Van Hee, Lefever, Desmet, Hoste and Daelemans, 2021). Transfer learning is a paradigm of methods that utilize very large scale unlabeled data to train a model on related tasks, then transfer knowledge acquired from this learning process to a new, different task (Lu, Behbood, Hao, Zuo, Xue and Zhang, 2015). The most common application of transfer learning is the use of pre-trained language models (LMs), especially transformer-based LMs, such as BERT, RoBERTa, XLM, etc. (Devlin et al., 2019; Conneau and Lample, 2019; Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer and Stoyanov, 2019b). The general idea behind is to first pre-train a LM in an unsupervised or self-supervised manner and then transfer the model to downstream tasks (Wu and Ong, 2021). Since the pre-training is unsupervised or self-supervised, the training can make use of extraordinarily large corpora, such as the corpus used for pre-training BERT that contains over 3 billion words (Devlin et al., 2019).

In the context of TCC, using a pre-trained LM involves adding task-specific layers (classifier layers) atop of the pre-trained LM for the downstream TCC task, and then train the new model where only the task-specific layers are trained from scratch (Devlin et al., 2019; Conneau and Lample, 2019; Liu et al., 2019b). In this way, the previously-acquired knowledge of the LM is transferred to the downstream TCC task. A few studies of TCC tasks explore these pre-trained LMs. For example, Mozafari, Farahbakhsh and Crespi (2019) and Zhao, Zhang and Hopfgartner (2021) transfer pre-trained LMs onto different downstream architectures. Zhao et al. (2021) test continued pre-training of a LM on an in-domain corpus, i.e., toxic comment datasets.

## 2.2. Unintended bias

The section below describes two aspects of unintended bias. First, we look at different types of unintended biases based on demographics. Second, we review existing methods for mitigating the unintended bias. These methods can be divided into two sub-groups: debiasing the dataset and debiasing the model.

### 2.2.1. Bias found in TCC models

Before proceeding to introduce the concept of unintended bias, it is important to acknowledge that machine learning models are designed to identify and use biased patterns in data to help the prediction tasks. For example, a model trained to identify toxic comments is intended to be biased towards features of toxic comments such that toxic comments receive higher scores than those which are not toxic (Dixon et al., 2018). Nonetheless, the model is not designed to discriminate people based on the groups, classes, or other categories to which they belong to or are perceived to belong to, such as gender, religion and race. If it does, we refer to this type of “discrimination” learned by the model as unintended bias (Dixon et al., 2018). That is, unintended bias in machine learning can appear as systemic differences in performance for different demographic groups (Hardt, Price and Srebro, 2016; Borkan, Dixon, Sorensen, Thain and Vasserman, 2019; Dixon et al., 2018). Unintended bias of machine learning has emerged as a new area of research in recent years and has raised many concerns among the NLP research community, including TCC research (Hovy and Spruit, 2016; Blodgett, Green and O’Connor, 2016; Tatman, 2017; Dixon et al., 2018; Blodgett, Barocas, Daumé III and Wallach, 2020).

A few recent works [distinguish](#) different unintended biases based on demographic features, such as gender bias, racial bias, and dialectal bias (Tan and Celis, 2019; Sap, Card, Gabriel, Choi and Smith, 2019; Davidson, Bhattacharya and Weber, 2019; Zhou, Sap, Swayamdipta, Choi and Smith, 2021; Park et al., 2018; Kennedy et al., 2020; Bolukbasi, Chang, Zou, Saligrama and Kalai, 2016; Zhang, Sneyd and Stevenson, 2020; Vaidya, Mai and Ning, 2020; Halevy, Harris, Bruckman, Yang and Howard, 2021). Studies by Davidson et al. (2019); Mozafari, Farahbakhsh and Crespi (2020); Xu, Pathak, Wallace, Gururangan, Sap and Klein (2021) focus on racial bias against users using African-American English. They find that tweets written in African-American English are predicted as toxic significantly more often than those written in standard American English. While this kind of bias is represented as dialectal bias, Mozafari et al. (2020) and Halevy et al. (2021) both refer it to as racial bias. Bolukbasi et al. (2016) and Zhang et al. (2020) study the gender bias found in the pre-trained Word2Vec word embeddings. Bolukbasi et al. (2016) show Word2Vec to contain female/male gender stereotypes. For example, the words like “receptionist” and “she” are strongly associated to each other, so are “maestro” and “he”.

While previous studies focus on one demographic group or one demographic feature, Park et al. (2018); Dixon et al. (2018); Kennedy et al. (2020) [introduce the expression “identity term” bias to investigate biases towards multiple demographic groups in TCC](#). “Identity terms” (also known as “group identifiers”) are words or terms referring to people with specific demographic characteristics, such as ethnic origin, religion, gender, or sexual orientation. [Representative identity terms include “Muslim”, “Black”, “women” and “democrat”](#). Park et al. (2018); Dixon et al. (2018); Kennedy et al. (2020) point out that TCC models tend to assign too much attention to such identity terms, resulting in incorrect predictions (Park et al., 2018; Dixon et al., 2018; Kennedy et al., 2020). Such bias towards identity terms often reflects the false positive predictions, known as false positive bias (Dixon et al., 2018; Halevy et al., 2021). For example, Park et al. (2018) give an example in their study that “You are a good woman” is predicted as “sexist”. One concept closely-related to “identity terms” is “bias sensitive words”, proposed by Badjatiya, Gupta and Varma (2019). They define that *a word  $w$  is defined as a bias sensitive word for a classifier if the classifier is unreasonably biased with respect to  $w$  to a very high degree*. For example, as discussed in their work, “dirty”, “shit”, “gotta”, “muslims”, “she” and “woman” are bias sensitive words (Badjatiya et al., 2019). The main difference between “identity terms” and “bias sensitive words” is that a bias sensitive word is unnecessarily an identity term referring to a group of people. Another point worth mentioning is that racial bias and gender bias cannot be taken as sub-concepts or sub-type of identity term bias directly as racial bias and gender bias is not necessarily shown up with identity terms. For example, racial bias can be reflected in the bias against African-American English and gender bias can be reflected in the gender stereotypes regarding professional positions existing in the model.

### 2.2.2. *Methods to mitigate bias*

TCC research has attempted to address the bias in two different perspectives: biases associated with datasets and biases associated with the model (Mozafari et al., 2020; Wiegand, Ruppenhofer and Kleinbauer, 2019). Biases associated with datasets mainly refer to the biases introduced by the data labelling or the collection process (Dixon et al., 2018; Davidson et al., 2019; Wiegand et al., 2019). Biases associated with the model refer to the biases learned by the model in the training process. **Debiasing the dataset** and **debiasing the model** are two different but non-exclusive approaches to handle the unintended bias on TCC tasks. The former essentially curates the data to remove the bias from the dataset, while the latter designs algorithms to mitigate the bias from the model. [It is hard to conclude which approach between these two is better based on current studies. This study aims to handle the identity term bias in TCC tasks by debiasing the model.](#)

Wiegand et al. (2019) investigate six TCC datasets to study why TCC datasets contain biases and the bias degree of different TCC datasets. For instance, they find that over 70% of sexist tweets from the dataset by Waseem and Hovy (2016) originate from two Twitter authors, which can contribute to the bias. Research on **debiasing the dataset** modifies the dataset, implicitly assuming there are bias-related features in the dataset that can be removed or reduced. Dixon et al. (2018) debias TCC datasets by adding non-toxic comments with identity terms to “balance” the training data, intuitively allowing the model to learn more features of non-toxic comments with identity terms. Vidgen, Thrush, Waseem and Kiela (2021) create “balanced” TCC datasets following an iterative procedure. The main idea is to let human annotators manually present data that are challenging for the model to predict correctly. This kind of data, also known as adversarial data and perturbations (to the model), manipulate the original text just enough to flip the label (e.g. from “Hate” to “Not Hate”) (Kaushik, Hovy and Lipton, 2019; Gardner, Artzi, Basmov, Berant, Bogin, Chen, Dasigi, Dua, Elazar, Gottumukkala et al., 2020). Adversarial data is believed to enrich current datasets that are created in potentially biased ways. For example, some hate speech datasets are created by collecting tweets containing



hate-related keywords, and some datasets are annotated by a limited group of annotators (Waseem, 2016; Wiegand et al., 2019; Fortuna and Nunes, 2018). In the method by ?, the model is trained on a new dataset that contains half adversarial data and half “originally entered content” (i.e., the collected real data). The prediction results of the current round are then used to guide human annotators to provide a new round of adversarial data. This process is repeated for four rounds. Therefore, more new adversarial data is provided by human annotators in each round to train a new model. ? have found that the model trained on the data from all four rounds together performs best and models trained on data from a single round generally perform better than models trained on non-adversarial data. However, they also acknowledge the potential annotator bias in their method.

Interestingly, there is an emerging debate between “curating data” and “studying the world as it is” in the NLP research community (Bender and Friedman, 2018; Bender, Gebru, McMillan-Major and Shmitchell, 2021; Buckman; Rogers, 2021). Those favouring the first approach argue that the real data reflects the real world which has discrimination and curating data is a method to combat discrimination against different social groups (Blodgett et al., 2020; Bender et al., 2021). Another supporting argument is that machine learning models may memorize specific facts which can expose personally identifiable information. Curating data is a way to remove these identifiable personal information for security reasons. What is more, deep learning models are vulnerable to basic perturbations and attacks, such as adversarial data designed by humans targeting on the specific bias of the model. This can be “solved” with shallow data curating (Ribeiro, Wu, Guestrin and Singh, 2020). However, those favouring “studying the world as it is” argue for algorithmic solutions to address similar issues mentioned above (Clark, Yatskar and Zettlemoyer, 2019; Rogers, 2021; Ribeiro et al., 2020; Garrido-Muñoz, Montejo-Ráez, Martínez-Santiago and Ureña-López, 2021). Also, they argue that curation means making conscious choices about what to include and what to exclude and this raises new questions: what is the standard for it, and what is the proper degree of curation (Rogers, 2021). Particularly, Sambasivan, Arnesen, Hutchinson, Doshi and Prabhakaran (2021) suggest that “conventional algorithmic fairness is west-centric”. Changing the data possibly inserts new values and the “west-centric” values which interpret the histories and cultures of non-western societies from a Western perspective can be controversial<sup>1</sup> (Amin, 1989; Blaut, 1993; Wallerstein, 1997). Last, Lissack (2021) argues that supports of curating data are “advocacy rather than research” (Rogers, 2021). We would like to direct interested readers to papers by Rogers, Baldwin and Leins (2021), Bender and Koller (2020), Garrido-Muñoz et al. (2021), Rogers (2021) and Sambasivan et al. (2021) for further details.

In contrast to debiasing the dataset, there are more studies on **debiasing the model**. Two popular approaches are introduced in the literature: using ensemble models and adding regularisation terms. In short, the idea of using ensemble models involves adding an additional classifier to learn the bias features and letting the main classifier learn bias-free features. Contrarily, the idea of adding regularisation terms is adding additional training objectives to penalise the bias-related features (Vaidya et al., 2020; Halevy et al., 2021; Kennedy et al., 2020; Clark et al., 2019). For an example of ensemble models, Clark et al. (2019) first train a “bias-only” model on a dataset whose data has been added with deliberate information, such as indicator features. They then train a second model in an ensemble with the pre-trained “bias-only” model on the original dataset<sup>2</sup>. The intuition behind this ensemble model is to allow the second model to learn less bias-related features as the “bias-only” model “absorbs” the bias-related features.

With respect to adding regularisation terms, more previous works include (Zhang, Lemoine and Mitchell, 2018a; Prost, Qian, Chen, Chi, Chen and Beutel, 2019; Xia, Field and Tsvetkov, 2020; Mozafari et al., 2020; Kennedy et al., 2020). Kennedy et al. (2020) find that BERT is over-attentive to identity terms and it neglects the context around the identity terms. This has led to many false positives. They thus propose BERT+SOC (Sampling and Occlusion), which is built atop BERT with an extra regularization term to predict an “importance score” of identity terms. The idea is to minimise the prediction differences between when an identity term is exposed to the model and when it is hidden from the model via the importance scores. Ideally, the “over-attended” identity terms will be assigned with low importance scores and thus they will become less indicative of whether the comment is hate speech or not. Therefore, their method encourages the model to pay less attention to identity terms, which can sometimes be actually useful features. Similarly, Prost et al. (2019) add a regularisation to penalise the dependence between the distribution of predicted probabilities and protected features, such as the dependence between toxic labels and identity terms. That is, the model attempts to

<sup>1</sup>Eurocentrism, also known as west-centric, refers to a discursive tendency to interpret the histories and cultures of non-European societies from a European (or Western) perspective (Amin, 1989; Blaut, 1993; Wallerstein, 1997). Common features of Eurocentric thought include: ignoring or undervaluing non-European societies as inferior to Western; ignoring or undervaluing what Asians or Africans do within their own society or seeing the histories of non-European societies simply in European terms, or as part of “the expansion of Europe” and its civilizing influence (Amin, 1989; Blaut, 1993; Wallerstein, 1997).

<sup>2</sup>Their method involves changing the dataset. However, we group it into debiasing the model as their final model is only trained on the original dataset.

minimise the prediction difference between protected features and other non-protected features. Another example is provided by Mozafari et al. (2020) who use a regularisation term to re-weight input samples to suppress the effect of highly correlative n-grams found in the training set. Studies that follow a similar direction include Zhang et al. (2018a) and Xia et al. (2020).

Zhou et al. (2021) compare the performances of debiasing the dataset and debiasing the model on one TCC task. In order to debias the dataset, they filter out “too easy” data that might contain spurious correlations or biases. They test different methods to find those “too easy” instances. For example, one method they use is AFLite proposed by Le Bras, Swayamdipta, Bhagavatula, Zellers, Peters, Sabharwal and Choi (2020) and its intuition is that examples predicted correctly by the simplest methods likely exhibit spurious biases. For debiasing the model, they insert additional training objectives to the model by adapting the method by Clark et al. (2019). They have found that models trained on debiased datasets (i.e., debiasing the dataset) achieve overall higher performance than models with additional training objectives (i.e., debiasing the model), while the latter performs better on lexical bias reduction. Here, lexical bias refers to the bias towards a list of “bad words”, also known as “Toxicity Triggers” in their paper. However, Zhou et al. (2021) have not compared their approach against other SOTA methods that debias the model, such as adding a regularisation proposed by Kennedy et al. (2020) and a multi-task-based method proposed by Vaidya et al. (2020). Another limitation is that Zhou et al. (2021) have experimented with only one task and this limits the generalisability of their findings.

### 2.3. Subjectivity for toxic comment classification

Previous studies show that toxic comments, such as offensive speech and hate speech, tend to be expressions of subjective feelings or opinions (Pang and Lee, 2004; Gitari et al., 2015; Burnap and Williams, 2016; Benito, Araque and Iglesias, 2019). Therefore, a few TCC studies have utilized subjectivity in classification. For example, Gitari et al. (2015) assume non-subjective sentences are not toxic comments and therefore they filter out non-subjective sentences with a rule-based approach prior to classification. These filtered sentences are considered automatically to contain non-hateful content. The intuition is to make the classification task “easier” by removing non-subjective sentences in advance. The study by Van Hee, Jacobs, Emmery, Desmet, Lefever, Verhoeven, De Pauw, Daelemans and Hoste (2018) use the positive and negative opinion word ratios and the polarity calculated with sentiment lexicons as the “subjectivity lexicon features” of a comment. They find that subjectivity lexicon features prove to be strong features for cyberbullying detection. In short, these studies have implied that the likelihood of a comment being toxic is associated with its subjectivity. However, none of them attempts to quantify the subjectivity level or utilize it in a non-binary fashion to improve the model. Furthermore, none of them has associated the subjectivity level with mitigating the identity term bias.

Although a few TCC studies explore utilizing subjectivity on TCC tasks, there is a lack of consensus on how subjectivity should be defined. For example, Gitari et al. (2015) briefly point out that “a subjective sentence expresses some feelings, views, or beliefs.” Other studies, such as Lin, He and Everson (2011) and Huo and Iwaihara (2020), directly focus on subjectivity detection but none of them gives a precise definition of subjectivity or an explanation of what kind of comments or texts are supposed to be labeled as “subjective”.

### 2.4. Summary

In summary, existing studies have found that SOTA models demonstrate bias towards identity terms on TCC tasks, also known as identity term bias. This has become a focus of study in recent years. Existing approaches to addressing this bias either debias the data or the model. Methods of debiasing the dataset usually modify the dataset to make it more balanced. However, it is unclear how such an approach can be generalised and transferred to a different task or domain. Methods of debiasing the model mostly follow a similar principle that encourages the model to ignore or pay less attention to identity terms. However, this overlooks the important fact that in particular situations, such terms are useful for prediction. As an example, given the sentence “women cannot drive” and the sentence “children cannot drive”, the identity terms “women” and “children” are crucial in correctly classifying the sentences. Ignoring these terms may lead to a false prediction. In this example, “children cannot drive” expresses a common sense rather than disrespect or hate towards children, while “women cannot drive” can be sexist.

In this work, we explore a new venue for debiasing TCC models utilizing the presence of identity terms in the comment and the subjectivity level of the comment. The novelty is that we consider identity terms together with the extent to which a message expresses subjective opinions (we refer to this as “subjectivity level” in the following). We demonstrate this with an in-depth analysis in the next section. Additionally, we propose that the semantic similarity between a comment and an “objective” reference text can be used as a proxy to measure subjectivity in our model.

**Table 1**

Summary of the four toxic comment classification tasks. “Toxic Proportion” refers to the proportion of “Toxic” comments after the conversion to binary classification.

Dataset	Source	Data Numbers	Original Labels	Toxic Proportion	Avg Text Length
WS(Kumar, Reganti, Bhatia and Maheshwari, 2018)	Stormfront	14,998	non-aggressive (42%), overtly aggressive (35%) covertly aggressive (23%)	11.17%	91
Tweet 18k(Waseem, 2016)	Twitter	18,625	racism (11%), sexism (20%), both (6.9%), neither	31.22%	96
Tweet 42k(Founta et al., 2018)	Twitter	42,314	abusive (9%), hateful (4%), normal (87%)	13.48%	123
Wiki(ConversationAI, 2017)	Wikipedia	159,571	toxic (10%), severe toxic (1%), obscene (5%), threat (0.3%), insult (5%), identity hate (1%)	10.17%	398

Specifically, given a comment containing an identity term, we compute the similarity between the comment and a Wikipedia description of the identity term and use the similarity as a proxy to the subjectivity level of the comment.

### 3. Analyzing bias towards identity terms

In most types of toxic comments, e.g., hate speech, aggressive language and abusive language, they are likely to express hate or encourage violence towards a person or group based on certain characteristics such as race, religion, sex, or sexual orientation (Gitari et al., 2015). Such expressions are intuitively more a reflection of personal feelings rather than fact-quoting. Therefore in the following, we analyse some commonly used toxic comment datasets to quantify this.

#### 3.1. The task and datasets

To provide a fair evaluation, we select four representative datasets with the aim to cover different social media platforms, dataset sizes, text lengths, toxicity classes (i.e., labels) and distributions over classes. One of the dataset is also used by one of our baseline model which will be introduced in Section 4 (Kennedy, Kogon, Coombs, Hoover, Park, Portillo-Wightman, Mostafazadeh, Atari and Deghani, 2018). Compared to previous studies that typically use 1 to 3 datasets, we have used a comparable collection of TCC datasets (Badjatiya, Gupta, Gupta and Varma, 2017; Kennedy et al., 2020; Caselli, Basile, Mitrović and Granitzer, 2021). Since the identity term bias is found in various types of toxic content and also to follow the practice by Kennedy et al. (2020) which study the identity term bias in the context of binary classification, we group different toxic comments into one group without distinguishing their specific types. Therefore, the task in this work is a binary toxic comment classification task where the model aims to predict if the comment is toxic or not.

The first dataset is collected from a white supremacist online forum (WS) (de Gibert, Perez, García-Pablos and Cuadros, 2018). We select this dataset as it is employed by Kennedy et al. (2020) to study identity term bias, thus allowing fair comparisons of results. It includes 10,703 posts in total, 1,196 of which are “toxic” and 9,507 are “non-toxic”<sup>3</sup>. The second and third datasets are both collected from Twitter. For the second dataset, denoted as *Twitter 18k* (Waseem and Hovy, 2016), we convert the labels “Racism”, “Sexism” and “Both” to “Toxic” and consider the label “Neither” as “non-Toxic”. The third dataset is denoted as *Twitter 42k* (Founta, Djouvas, Chatzakou, Leontiadis, Blackburn, Stringhini, Vakali, Sirivianos and Kourtellis, 2018). Similarly, labels of “Abusive” and “Hateful” are converted into “Toxic”, and “Normal” is treated as “non-Toxic”. We remove the examples of label “Spam” as they are not a type of toxic comments according to the typology of “hate-based rhetoric” proposed in Kennedy et al. (2018). After conversion, Twitter 18k contains 18,625 tweets in total with 5,814 “Toxic” tweets and 12,811 “non-Toxic” tweets. Twitter 42k has 5,705 tweets as toxic and 36,609 as normal<sup>4</sup>. The fourth dataset is collected from Wikipedia Talk page and annotated in a multi-label classification approach, denoted as *Wiki* (ConversationAI, 2017). There are six labels in total, namely “toxic”, “severe toxic”, “threat”, “obscene”, “insult” and “identity hate”. We convert the label of a post to “Toxic” if the post has at least one of the six labels, and convert the remaining posts to “non-Toxic”. This leads to 16,225 posts with “Toxic” labels and 143,346 with “non-Toxic” labels, respectively.

In short, the four selected datasets contain between 15,000 and 159,571 comments and cover three different social media platforms with different average comment lengths. Table 1 summarises the four datasets.

<sup>3</sup>The original binary labels are “hate” and “no hate”.

<sup>4</sup>The original version dataset includes over 80,000 tweets with their tweets IDs and labels published (Founta et al., 2018). We have retrieved 50,425 valid tweets using their tweets IDs. The remaining tweets have been deleted as by the point of this study. After removing “spam” tweets, 42,314 tweets are kept for this dataset.

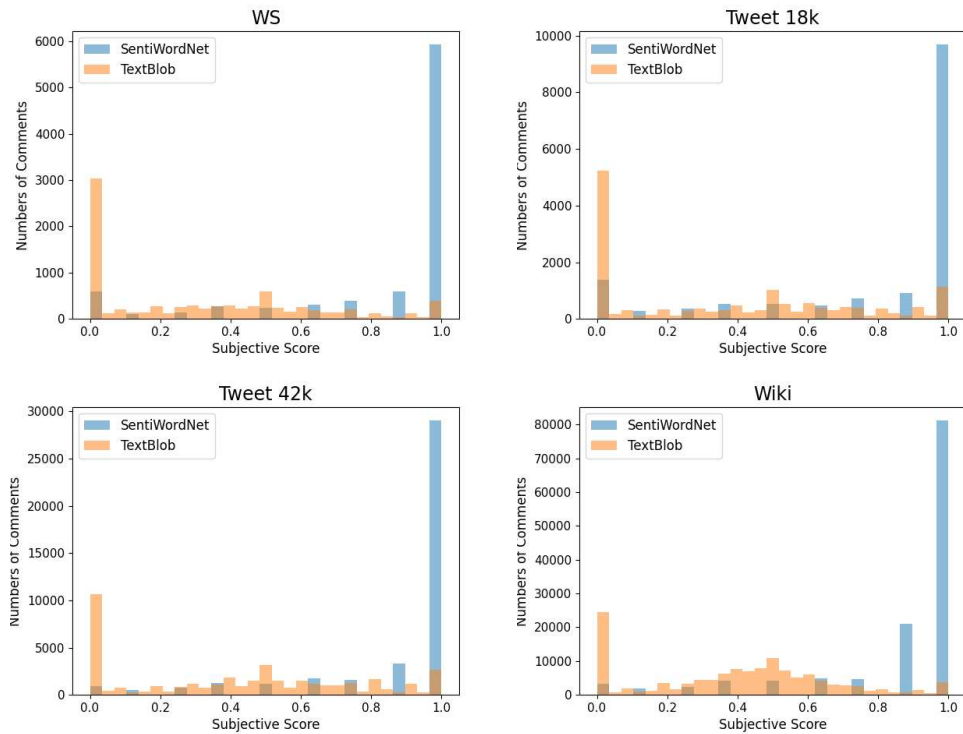


### 3.2. Defining “Subjectivity”

As discussed in Section 2.3, there is no consensus on the definition of subjectivity. In this work, we adopt the definition of “subjectivity level” by TextBlob (Loria, 2018) as mentioned before: the subjectivity level describes the extent to which the comment conveys personal opinion or factual information. A comment with a high level of subjectivity indicates that the comment contains more personal opinion and less factual information.

We use TextBlob<sup>5</sup> library (Loria, 2018) to generate a subjectivity score for each text in our datasets. Before choosing TextBlob, we also compared it against another tool, SentiWordNet (Sebastiani and Esuli, 2006). Given a text, both tools assign a subjectivity score within the range from 0.0 to 1.0 where 0.0 is very objective and 1.0 is very subjective.

However, no previous studies have compared the accuracy of these tools and we therefore conducted our own analysis, which led us to choose TextBlob over SentiWordNet. First, we study the distribution of the scores returned by the two tools. As shown in Figure 1, SentiWordNet assigns most texts an extreme score of “1” (very subjective). In other words, SentiWordNet has a distribution biased towards “1”, while TextBlob is less biased towards the two extremes.



**Figure 1:** The subjective scores distribution of TextBlob and SentiWordNet over the four datasets. The X-axis is the comment count of corresponding subjectivity scores indicated by the Y-axis. All four plots are based on the training dataset.

Second, we manually inspected a sample of 80 data instances from the four datasets (20 per dataset). We randomly selected those comments whose subjective scores by the two tools have a disagreement greater than 0.5. To give an example, the following comment:

*“my mothers father is full blonde Irish. my mothers mother is English Irish Scottish . my fathers father is a German with some Scottish and my fathers mother is danish and Norwegian. making me 6ft4 blonde hair blue eyes with a big barrel chest just like my german grandfather .”*

is scored differently by TextBlob (0.0682) and SentiWordNet (1). In this case, We take the score 0.0682 by TextBlob as the more accurate one. Our manual investigation shows that TextBlob is more accurate for 77.5% samples we validated.

<sup>5</sup><https://textblob.readthedocs.io/en/dev/#>

**Table 2**

Example comments that contain identity terms. Each pair of comments contain the same identity terms, while one is true positive and the other is false positive predicted by BERT. The subjectivity scores are assigned by TextBlob.

Identity Terms	Predictions	Comments	Subjectivity scores
muslim	True Positive (toxic)	i have been talking about it for years. the filthy <b>muslims</b> putting their sight on africa. in some ways they are worst then blacks.	1
	False Positive (non-toxic)	@ kelly ##turn ##er ##9 ##9 @ bu ##eller ##ish ##ere more <b>muslims</b> die in terrorist attacks than anyone else. for example, 80% of afghan civilians killed by taliban	0
gay	True Positive (toxic)	if this guy isn't gay he is one seriously good actor !!!!! i've seen hayes on will & grace and the way he acts so.... feminine he must be homosexual. no straight man can act so convincing ##ly <b>gay</b> . but then again sean acted pretty straight and normal in the movie where he played jerry lewis. my opinion is still in favour of him being gay though.	0.6444
	False Positive (non-toxic)	f ##yi i am a <b>gay</b> man , if anyone wants to chat just comment pl ##z	0.3417
whites	True Positive (toxic)	black people in sa do not want to unite with <b>whites</b> to remove corruption . we ' d rather endure rot & amp; suffer! so sad!	0.6444
	False Positive (non-toxic)	i ' m reporting here that i haven' t seen or heard anything on talk radio or tv that even hints at jewish su ##pre ##mac ##ist ethnic cleansing of <b>whites</b> from usa or europe .	0
women	True Positive (toxic)	rt @ mw##il##bu##r##6##2 i swear i am not sex ##ist , but <b>women</b> ' s tennis is the most boring sport to watch	0.75
	False Positive (non-toxic)	men and <b>women</b> are segregated in mosques .	0
jews	True Positive (toxic)	well i am sure that the <b>jews</b> have the cure for aids but they won't give it to the fa ##gs cause they already served their pu ##rp ##oise	0.8888
	False Positive (non-toxic)	if a third world savage attacks a <b>jew</b> , is it all jew ##ey ' s fault?	0

### 3.3. Predictive model: BERT

Influenced by Kennedy et al. (2020), we use BERT as the classifier to demonstrate the problem of identity term bias in TCC tasks. BERT would be a good baseline as it is widely used in various NLP tasks and many pre-trained language models are built upon it (Devlin et al., 2019; Liu et al., 2019b; Conneau and Lample, 2019).

In short, given a fixed length token sequence<sup>6</sup>, BERT assigns each token an initial embedding and an attention mask. The pre-trained embeddings are supposed to contain the semantic information of the token. The attention mask of a binary value (0 or 1) indicates if the token is a padded token or not to avoid the model performing attention on padded tokens. To be more specific, between the attention mask values, 0 is for tokens that are padded for making up to the fixed-length input so they are masked and the model does not need to attend to them during the training, 1 is for the actual tokens which the model attends to. Embeddings and masks are fed into the Transformer Encoder layers and the final logit output is used for the loss calculation and for the final classification. We train the initial BERT model on the four TCC tasks separately and investigate their prediction results. We follow the implementation details used in Kennedy et al. (2020), which are further explained in Section 5.

### 3.4. Bias in model predictions: qualitative analysis

Errors made by BERT are extracted for our analysis, i.e., false positive predictions and false negative predictions. As mentioned before, we use TextBlob to generate a subjectivity score for those comments (Loria, 2018).

We investigate those errors and select several representative examples as shown in Table 2. We observe that to correctly predict the toxic comments with identity terms, we need to account for the meaning of the whole sentence and the stance of the speakers. Subjectivity level is possibly one perspective from which this can be captured. For example, in the first pair which mentions “muslims” in Table 2, the toxic comment compares “blacks” with “muslims” and asserts that “they (Muslims) are worst than blacks” without any factual information, this indicates a high level of subjectivity. While the non-toxic comment with the term “muslims” attempts to describe an objective fact related to Muslims, although it contains a comparison, it attempts to include a specific figure to support the comparison<sup>7</sup>. This observation inspires us that the subjectivity score of the comments can be a helpful indicator when classifying a comment with identity terms.

### 3.5. Bias in model predictions: quantitative analysis

To further analyse the subjectivity level at scale, we conduct a quantitative analysis of the subjectivity score of false positive and true positive predictions. We separate those comments with identity terms and those without identity

<sup>6</sup>One word is converted to one or multiple tokens.

<sup>7</sup>The “fact” and figure the comment provides are unnecessarily true but it is not the topic we aim to study.

**Table 3**

25 identity terms used for bias analysis.

---

 muslim jew jews white islam blacks muslims women whites gay black democat islamic allah jewish lesbian transgender race  
 brown woman mexican religion homosexual homosexuality africans
 

---

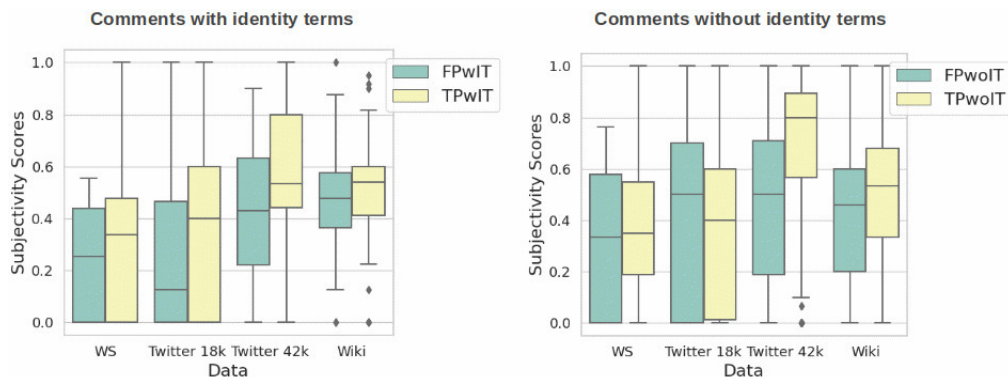
terms to examine the identity term bias. The identity term list is adopted from Kennedy et al. (2020) which includes 25 terms such as “muslim”, “jew”, and “women”, full list is shown below (in Table 3.5).

In this way, all comments are essentially split into four groups: true positive with identity terms (TPwIT), false positive with identity terms (FPwIT), true negative without identity terms (TNwoIT), false negative without identity terms (FNwoIT). We plot the subjectivity score distribution over false positives and true positives with box-plot diagrams. As shown on the left of Figure 2, for comments with identity terms, the true positives (i.e., toxic comments) show higher subjectivity levels than the false positives (i.e., non-toxic comments) across all tasks. First, false positives have a lower median of subjectivity scores than true positives across four datasets. Second, the false positive predictions have a generally smaller and lower interquartile range than the true positive predictions in the task. The lower subjectivity scores in false positive predictions may reflect the real-world scenario that when speakers talk about a demographic group such as female, Muslim or Asian in an objective way, e.g., describing the group neutrally, the speech is less likely to be disrespectful or offensive. On the other hand, toxic comments often involve subjective expressions.

Notably, the pattern of lower subjectivity level of false positives is consistent only among comments with identity terms. The comments without identity terms, as shown on the right diagram in Figure 2 do not indicate a consistent pattern between false positives and true positives. This shows that the feature of subjectivity level could be indicative only when considering the presence of identity terms at the same time. On the other hand, it also indicates that identity term bias can be addressed by considering the subjectivity level of a comment. Nonetheless, this is not to assert that a text mentioning identity terms in a subjective tone should be toxic. As shown in the left boxplot in Figure 2, there are indeed non-toxic comments with identity terms (i.e., FPwIT in green) that have been assigned with subjectivity scores over 0.5 on the dataset 42k and Wiki. We looked into these comments and present two examples:

*“mike ##pen ##ce not being able to have dinner alone with any woman other than his wife etc is just like being a strict muslim. ironic”* (score of 0.63).

*“ga##bs whites got nothing for free you won t it s the way life works stop w##hini##ng ab”* (score of 0.8).



**Figure 2:** The comparison of subjectivity scores between true positive (i.e., toxic comments, as coloured in yellow) and false positive predictions (i.e., non-toxic comments, as coloured in green) by an initial BERT classifier. This figure is better viewed in colour.

### 3.6. Summary

In summary, using the predictions by a SOTA BERT classifier, we observe that when identity terms are present, false positives tend to have lower subjectivity scores while true positives generally have higher subjectivity scores. Based on this observation, we hypothesize that the subjectivity score of a comment along with the presence of identity terms can be useful in classifying toxicity. We introduce our method to leverage this in the next section.

## 4. Subdentity-BERT (SS-BERT)

We propose a BERT-based model that make use of the structure of BERT embeddings to add the information of subjectivity and the presence of identity terms. The goal is to enable the model to pay attention to the subjectivity of a comment when the comment contains identity terms. When the identity terms are not present, the model should not consider the subjectivity. To do so, we append an additional “token” to the end of the token sequence for each comment to “notify” the model of the information of the subjectivity level and the presence of identity terms. The following sections explain this in detail.

### 4.1. SS-BERT model structure

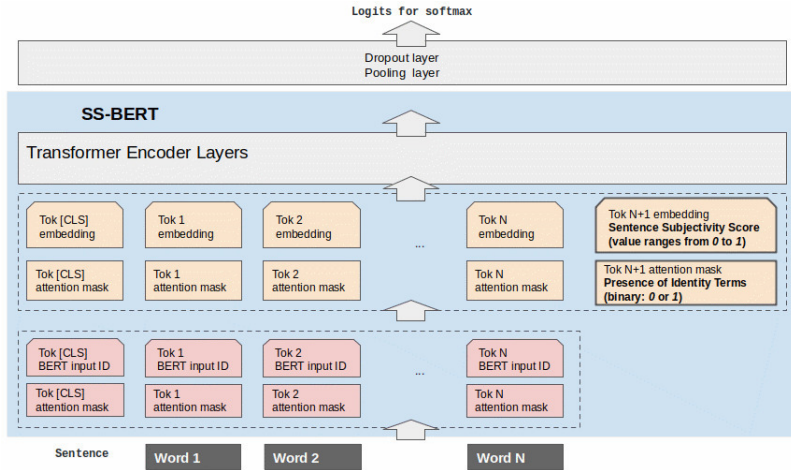


Figure 3: Illustration of a SS-BERT model on classification tasks. This figure is better viewed in colour.

As shown in Figure 3, building on BERT, we append an additional “token” to the end of the token sequence for each comment. We use the subjectivity score for the embedding of the token. To be more specific, we create a 3-D tensor with the same size of other token embeddings and each element’s value in the tensor is equal to the subjectivity score. In the scenario of BERT, the dimension size is 768 and thus the tensor for the added “token” is a 3D-tensor of size [batch size, 1, 768]. The tensor is denoted as “Sentence Subjectivity Score” in Figure 3.

For the corresponding attention mask (highlighted in yellow with bold borders in Figure 3), we set it to indicate the presence of identity terms so that if there is no identity term in the comment, the appended “token” will be masked. While if there is an identity term, the embedding of this “token”, i.e., the subjectivity score, will be attended by BERT.

### 4.2. Subjectivity score

We explore two different methods of measuring the subjectivity level of a comment. The first method is simply using the TextBlob library as mentioned before. The second is based on measuring the semantic similarity between the comment and a related Wikipedia description for the corresponding identity term found in the comment.

#### 4.2.1. TextBlob

We use TextBlob as described in Section 3.2 to assign a subjectivity score to each comment. The subjectivity scores range from 0.0 to 1.0 where 0.0 is very objective and 1.0 is very subjective. TextBlob uses a lexicon-based approach to compute subjectivity scores. It emphasises the impact of individual words (Sebastiani and Esuli, 2006; Loria, 2018). To be more specific, TextBlob uses a vocabulary and each word in the vocabulary is associated with a subjectivity score. In the case of polysemous words, it returns the average subjectivity scores over all the possible senses of that word. For out-of-vocabulary words, TextBlob assigns a subjectivity score of 0 to the word. A comment’s final subjectivity score is the mean subjectivity score of all its words.

#### 4.2.2. Wikipedia based similarity

Inspired by Zhang and Yu (2006); Zhang, Yu and Meng (2007); Kittur and Kraut (2008); Mesgari, Okoli, Mehdi, Nielsen and Lanamäki (2015), we derive objectivity from Wikipedia contents and use the derived objectivity as

[references to measure subjectivity](#). To be more specific, we make an assumption that the summary section of a Wikipedia article regarding an identity term is a relatively objective description of the identity term, and the similarity between the Wikipedia summary of a given identity term and a comment mentioning the same identity term can reflect the subjectivity level of the comment. The more similar they are, the more objective the given comment is. Zhang and Yu (2006); Zhang et al. (2007) have made similar assumptions that “all the contents of these pages (Wikipedia pages) are assumed to be objective”. In their work of opinion retrieval (i.e. finding relevant blog documents containing opinionated content for a given query topic), Zhang et al. (2007) compare the vectors’ similarity between given documents and related Wikipedia articles to find the opinionated content.

We propose to take the cosine similarity between a given comment and a Wikipedia summary text regarding an identity term as a proxy to the subjectivity of the comment. Specifically, given a comment with a certain identity term, e.g., “muslim”, “islam” and “lesbian”, we retrieve the first section (i.e., the “summary text”) of the corresponding Wikipedia article of the identity term. Then, the summary is fed to BERT which gives an embedding for this summary<sup>8</sup>. Second, we apply the same process to the given comment to create its corresponding BERT embedding. We pool<sup>9</sup> both embeddings (one for the Wikipedia summary and the other one is for the given comment) to one dimension of size 768 (the hidden size of BERT) and calculate their cosine similarity ( $\frac{u \cdot v}{\|u\| \|v\|}$ ).

The higher value of a cosine similarity, the more similar the comment is to the Wikipedia summary. Since the Wikipedia summary is generally an objective narrative of the topic (further discussed below), we hypothesize that this degree of “similarity” may capture the degree of “objectivity” to some extent. Therefore, the subjectivity of the comment, which is the opposite to objectivity, is calculated as:

$$1 - \frac{u \cdot v}{\|u\| \|v\|} \quad (1)$$

where  $u$  and  $v$  denote the two embedding vectors. For example, using this Wikipedia based subjectivity measure, the subjectivity score for “women cannot drive” is 0.4011 and for “children cannot drive” is 0.2937. These scores indicate the former has a higher subjectivity score, which is consistent with the discussion in Section 2.4 regarding these two sentences. This suggests that the Wikipedia summaries may have provided useful background information that helps the interpretation of the two messages that are only one-word different.

We acknowledge that this score is not a direct reflection of the subjectivity level, however, we argue that this to some extent, reflects the subjectivity level of a comment. Although Wikipedia articles are also user-generated content, their collaborative authoring nature [can potentially reduce](#) the risk of creating subjective content. Additionally, Wikipedia articles are scrutinized by a wide community which helps ensure the descriptions regarding identity terms are accurate and objective (Kittur and Kraut, 2008; Mesgari et al., 2015). In particular, articles about identity terms are edited many times by different contributors and reviewers from different backgrounds (Hu, Lim, Sun, Lauw and Vuong, 2007). Also, earlier studies showed that the summary section of a Wikipedia article usually defines the topic in question (Ye, Chua and Lu, 2009; Sankarasubramaniam, Ramanathan and Ghosh, 2014). To sum up, [it is reasonable to take the summary text from Wikipedia for a certain identity term as an objective description for the identity term](#). We show empirically later in the experiments that this is useful for our model. We call this method “Wikipedia based subjectivity” and denote it as (*Wikipedia*) when a model uses this score as subjectivity scores, i.e., SS-BERT (Wikipedia) and SS-BERT+SOC (Wikipedia).

Since SS-BERT does not attend to the subjectivity level of comments without identity terms and we cannot compute the Wikipedia based subjectivity for such comments, their subjectivity will be assigned as zero. For comments with multiple identity terms, we repeat the above process for each identity term and use the mean value as the final subjectivity score.

<sup>8</sup>We constrain the maximum length of the summary to be 500 tokens. The 25 embeddings (one embedding for one summary of its identity term) are saved as a lookup table to speed up the training process.

<sup>9</sup>Here, each embedding is a 2D tensor with a size of [token length, 768]. 768 is the hidden size of BERT. We average the first dimension which gives a 1D tensor with a size of [768].



## 5. Experiments

### 5.1. Comparative models and datasets

We design two baseline models and two variations of our SS-BERT, in order to fully evaluate the effects of our design of SS-BERT. For SS-BERT and the two variations, we experiment with the two different ways of calculating subjectivity as detailed in Section 4.2.

**Baselines** The first baseline is an initial BERT, as explained in Section 3.3. The second baseline, BERT+SOC, is an implementation of the SOTA method in Kennedy et al. (2020), as described in Section 2.2.2. To the best of our knowledge, Kennedy et al. (2020) is the only work that focuses on mitigating identity terms bias found in BERT.

**Subjectivity-Only BERT (SO-BERT)** To examine if the subjectivity level is generally helpful for TCC tasks regardless of the presence of identity terms, we create a variation of our method, SO-BERT, which only captures the information of subjectivity level but not the presence of identity terms. To do so, we adapt from the SS-BERT model with the attention mask always attending the added “token”. The rest of the model structure remains the same. We only test subjectivity scores by TextBlob for this comparative model, denoted as “SO-BERT (TextBlob)”, as our Wikipedia based subjectivity method cannot calculate the similarity without the presence of identity terms (which are used to look up the corresponding Wikipedia articles for comparison). Note that the SO-BERT is different from the previous TCC works incorporating subjectivity in their methods as SO-BERT embeds the subjectivity level in a pre-trained LM directly rather than using it prior to the model, such as the work by Gitari et al. (2015) as described in Section 2.3, or as feature in a traditional machine learning fashion.

**Subjectivity-BERT+SOC (SS-BERT+SOC)** combines the method of BERT+SOC by Kennedy et al. (2020) and our model (SS-BERT) to create a hybrid BERT-based model, SS-BERT+SOC. In short, it learns the subjectivity level and the presence of identity terms with the added “token” and also has an extra regularization term in its loss function which encourages the model to learn more from the context of the identity term and less from the identity term. We experiment with the two different subjectivity measures detailed in Section 4.2, i.e., SS-BERT+SOC (TextBlob) and SS-BERT+SOC (Wikipedia).

We evaluate each model on four different datasets as introduced above. Each dataset is split to training, validation and test datasets (80%, 10%, 10%). The results reported in this work are all on the testing dataset.

### 5.2. Implementation

We constrain the maximum length of each input instance to be 128 tokens for WS, Twitter 18k and Twitter 42k, and 400 tokens for Wiki. The maximum length is set up based on the average length of text in each dataset as detailed in Table 1 in the appendix.

The hyperparameter settings follow those in Kennedy et al. (2020). Accordingly, the batch size is set to 32. Adam optimization is implemented with a starting learning rate of  $2 \times 10^{-5}$ . The validation is performed every 200 steps and the learning rate is halved every time the validation F1 decreases. The model stops training after the learning rate halved five times. We also re-weight the training loss to handle the imbalance labels as Kennedy et al. (2020) does.

We used a single NVIDIA Tesla V100 GPU for all experiments. We run the two baseline models (BERT and BERT+SOC) with the code provided by Kennedy et al. (2020). Our code for SS-BERT, SS-BERT+SOC and SO-BERT is also built on Kennedy et al. (2020), such that our implementation uses the same software packages<sup>10</sup>. For each task, a model runs 10 times independently to give a mean F1 score, following the implementation of Kennedy et al. (2020).

### 5.3. Results: F1 comparison

**SS-BERT** We first analyse the F1 performance of SS-BERT. Table 4 shows F1 scores of the two baseline models and the two SS-BERT models with different subjectivity measures on four datasets. Overall, both SS-BERT models outperform the two baseline models consistently across four datasets.

Between the two baselines<sup>11</sup>, BERT+SOC is able to improve BERT on WS which reflects the results reported in Kennedy et al. (2020). However, the F1 score is only marginally higher on Twitter 42k and remains the same on Twitter 18k, while it decreases on Wiki compared to BERT. The under-performance of BERT+SOC on Wiki may indicate that BERT could have benefited from training on the significantly larger dataset (compared to WS, Twitter 18k and

<sup>10</sup>Therefore, we use “bert-base-uncased” BERT as Kennedy et al. (2020).

<sup>11</sup>We notice that our results of BERT and BERT+SOC on the WS dataset are different from that reported in Kennedy et al. (2020), as our F1 are higher. While our results are obtained by re-running their code as-is, a possible reason for this difference is that the only version of the data we can download has been modified from that used in the authors’ original study.

**Table 4**

The comparison of F1 between SS-BERT and baseline models on the four TCC tasks. The mean F1 score and its standard deviation are from ten independently runs for each model presented.

Data	Data Size	Baseline Models				SS-BERT			
		BERT		BERT+SOC		SS-BERT(TextBlob)		SS-BERT(Wikipedia)	
		F1	std	F1	std	F1	std	F1	std
WS	10,703	0.5811	0.0204	0.5885	0.0209	0.5952	0.0203	<b>0.5970</b>	0.0175
Twitter 18k	18,625	0.7780	0.0204	0.7780	0.0055	<b>0.7804</b>	0.0080	0.7803	0.0052
Twitter 42k	42,314	0.7637	0.0071	0.7643	0.0101	0.7683	0.0059	<b>0.8000</b>	0.0081
Wiki	159,571	0.7680	0.0175	0.7548	0.0135	0.7693	0.0086	<b>0.7735</b>	0.0086

**Table 5**

The comparison of F1 of different methods utilizing subjectivity scores on the 4 TCC tasks. The mean F1 score and its standard deviation are from 10 independently runs for each model presented. The model under-performing one of the baselines are enclosed in parentheses “[ ]” and models under-performing both baselines are enclosed in parentheses “[ [ ] ]”.

Data	Data Size	SS-BERT		Comparative Models with Subjectivity			
		SS-BERT(TextBlob)		SS-BERT+SOC(TextBlob)		SO-BERT(TextBlob)	
		F1	std	F1	std	F1	std
WS	10,703	<b>0.5952</b>	0.0203	0.5912	0.0216	0.5909	0.0247
Twitter 18k	18,625	<b>0.7804</b>	0.0080	0.7785	0.0050	[[0.7774]]	0.0055
Twitter 42k	42,314	<b>0.7683</b>	0.0059	0.7660	0.0056	[0.7636]	0.0061
Wiki	159,571	<b>0.7693</b>	0.0133	[0.7568]	0.0112	[0.7654]	0.0151

Data	Data Size	SS-BERT		Comparative Models with Subjectivity	
		SS-BERT(Wikipedia)		SS-BERT+SOC(Wikipedia)	
		F1	std	F1	std
WS	10,703	0.5970	0.0175	<b>0.5980</b>	0.0272
Twitter 18k	18,625	0.7803	0.0052	<b>0.7812</b>	0.0036
Twitter 42k	42,314	<b>0.8000</b>	0.0081	0.7687	0.0068
Wiki	159,571	<b>0.7735</b>	0.0086	[[0.7539]]	0.0149

Twitter 42k) such that the extra learning objective enhancing the contextual information around identity terms may have had negligible influence on the model. In contrast, our two SS-BERT models outperform BERT on all datasets. This suggests the mechanism of attending to subjectivity based on the presence of identity terms cannot be compensated by dataset size. Therefore, the results show that our model brings unique benefits and that is the reason SS-BERT still outperforms BERT on Wiki.

Between the two different subjectivity measures, SS-BERT (TextBlob) and SS-BERT (Wikipedia) achieve comparable performance on the two smaller datasets, WS and Twitter 18k, while SS-BERT (Wikipedia) works better on the two bigger datasets, Twitter 42k and Wiki. A possible explanation for this might be that the average comment length of Twitter 42k (123) and Wiki (398) are longer than WS (91) and Twitter (96) 18k. They are in fact, of more similar lengths to the Wikipedia summary text. A short comment intuitively may contain less semantic information and may not provide sufficient features for the similarity computation. Another possible explanation for this is that the Wikipedia summary text has provided background information regarding identity terms in addition to the target datasets.

**Variation models using subjectivity** We compare our method SS-BERT against the other two variation models which also utilize subjectivity scores, namely SO-BERT and SS-BERT+SOC. Table 5 shows models using subjectivity

**Table 6**

Summary of false positives and false negatives of BERT and SS-BERT on the 4 datasets. Mean values of the performance across 10 runs are shown. The lowest False Positive on each task is in bold.

DATA	BERT		SS-BERT (TextBlob)		SS-BERT (Wikipedia)	
	False Positive	False Negative	False Positive	False Negative	False Positive	False Negative
WS	<b>34</b>	57	<b>34</b>	55	36	54
Twitter 18k	149	117	<b>146</b>	117	149	114
Twitter 42k	227	78	214	81	<b>134</b>	101
Wiki	750	168	738	170	<b>608</b>	176

scores by TextBlob on the top and those by models using Wikipedia based subjectivity at the bottom. The model under-performing one of the baselines shown in Table 4 are enclosed in parentheses “[ ]” and models under-performing both baselines are enclosed in parentheses “[ [ ] ]”.

Comparing SS-BERT against SS-BERT+SOC, we notice the following patterns. When using TextBlob for measuring subjectivity, SS-BERT consistently achieves the best performance on all tasks. When using Wikipedia based subjectivity, SS-BERT noticeably outperforms SS-BERT+SOC on the two bigger datasets and achieve comparable results to SS-BERT+SOC on the two smaller datasets. It is worth noting that the F1 by SS-BERT-SOC on WS and Twitter 18k are only marginal higher than SS-BERT whereas SS-BERT+SOC obtains F1 that are lower than the two baselines on the Wiki dataset. Overall, SS-BERT works better than SS-BERT+SOC. A possible reason can be that the extra regulation from SOC might dilute the impact that SS-BERT brings to the model.

Comparing SS-BERT against SO-BERT, we notice adding subjectivity information to the model alone does not lead to improvement to BERT on most tasks. In other words, the model that considers subjectivity regardless of the presence of identity terms, does not consistently benefit TCC tasks. Only adding the information of the subjectivity and the identity terms presence together consistently improves over the baselines. This reflects the intuition we mentioned previously and the pattern we identified in Section 3.5 that the subjectivity level of a comment is an indicative feature for toxicity only if identity terms are present in the comment.

Another important finding is that although BERT+SOC is designed to mitigate the identity term bias, it is not able to learn the subjectivity level of comments with identity terms. Therefore, SS-BERT+SOC outperforms BERT+SOC consistently, suggesting that adding the subjectivity level and the presence of identity terms can improve BERT+SOC performance.

#### 5.4. Results: identity term bias

We compare the erroneous predictions between SS-BERT and the baseline BERT to further investigate SS-BERT’s performance, especially regarding the model’s ability to handle false positive bias and identity term bias. First, as shown in Table 6, [SS-BERT \(TextBlob\)](#) and [SS-BERT \(Wikipedia\)](#) can both reduce the false positive predictions on Twitter 42k and Wiki, while their false positives predictions on WS and Twitter 18k are comparable to the baseline BERT. The dataset Twitter 42k benefits from SS-BERT the most, considering the number of reduced false positives with respect to the dataset size. This is consistent with Figure 2 which shows a noticeable gap of subjectivity levels between false positive and true positive predictions from BERT on Twitter 42k.

Reflecting on the improved performances on the two bigger datasets, Twitter 42k and Wiki, SS-BERT (Wikipedia) effectively reduces the false positives on these two datasets. However, SS-BERT (Wikipedia) cannot effectively reduce false positives on WS or Twitter 18k. [Nevertheless, SS-BERT \(Wikipedia\) marginally reduces false negatives on these two datasets. This overall leads to improvement in F1 \(shown in Table 4\) as SS-BERT \(Wikipedia\) maintains comparable levels of false positives on the two smaller datasets.](#)

These results may be explained by the fact that the baseline model BERT has more false positives than false negatives over the two datasets. To be more specific, 74.43% and 81.70% erroneous predictions are from false positive cases for Twitter 42k and Wiki, while the numbers are only 37.36% and 56.02% for WS and Twitter 18k. The higher false positive percentage from the baseline model indicates more room for improvements. Last, as we mentioned in Section 2.2 and identified by [Wiegand et al. \(2019\)](#), the dataset Twitter 18k can be problematic as over 70% of the tweets labelled as sextists were written by just two Twitter users. [We have also identified some arguable labels in the](#)

**Table 7**

Examples of false positives and false negatives with identity terms that are not included in the list by Kennedy et al. (2020).

	Comments	Group identity terms	SS-BERT Predictions
1.	liberal is just the pc word for rap ##ist .	liberal	False Negative
2.	i get what your saying and there are thousands of non irish here who will never contribute anything to this land .	irish	False Negative
3.	i suppose the concept of a male cow and a female cow producing more cows , and the repeating process is too much for them to understand .	female	False Negative
4.	traditionally , there have always been a lot of them despite the fact that ne asian ##s are the extreme version of our breeding strategy .	asian	False Negative
5.	there is no doubt the chinese and ja ##ps have giving the world culture .	chinese	False Positive
6.	the reason they are killing themselves is because our weak liberal gov ##er ##ment are too afraid to make them do any sort of training . .	liberal	False Positive
7.	she came to hate the african po ##ache ##rs and showed no mercy towards them .	african	False Positive
8.	these girls are not there for the cooking # mk ##r what a joke .	girls	False Positive

false positives from the two smaller datasets in our investigation. The predictions for some comments are potentially correct even though these comments are annotated differently by humans.

### 5.5. Error Analysis

We perform an error analysis to shed light on the limitations of the best performing model (SS-BERT) with a focus on the identity term bias. First, the identity term list we use may not “fit” the datasets well. An updated identity term list may bring further improvement. Specifically, there is only a small percentage of comments that contain the identity terms used in the model. For WS, Twitter 18k, Twitter 42k and Wiki, the ratio of comments with identity terms in the list are 21.20%, 19.90%, 4.22%, and 6.02% respectively. Therefore, only a small percentage of the datasets could have benefited from our method. For example, we observe that some false negatives and false positives contain identity terms like “liberal”, “irish”, “asian”, and “chinese”, which are not included in the identity term list used in our model. Examples are listed in Table 7 as shown below. Kennedy et al. (2020) uses a supervised approach combined with manual screening to select identity terms. [However, in practical applications, it is hard to generalise this method to new tasks with unseen data. We believe it would be a promising direction that explores different methods to compile an identity term list that helps different TCC tasks with our method.](#)

Second, the subjectivity scores given by TextBlob are not always accurate for each comment with identity terms. We have investigated the code of TextBlob and found several limitations of its method. For example, they simply take the mean subjectivity scores over all words in a comment, which may be insufficient to capture the subjectivity level accurately if the comment contains a large percentage of objective words that reduce the overall subjectivity level of the comment. Also, TextBlob assigns out-of-vocabulary words a subjectivity score of 0 and this will penalise comments with many misspellings. We believe accurate subjectivity scores can potentially help SS-BERT predict them more accurately. For example, the false negative with the identity term “white”:

*“why white people used to say that sex was a sin used to be a mystery to me until i saw the children of browns and mixed race children popping up all around me”*

has been assigned a low subjectivity score of 0.125. The presence of identity terms and the low subjective score might mislead the model to make the incorrect prediction. More examples of this kind are shown in Table 8.

### 5.6. Generalization

To examine the generalization of our method, we also test our method on RoBERTa, which is another popular pre-trained LM for TCC tasks (Liu et al., 2019b; Baruah, Das, Barbhuiya and Dey, 2020). RoBERTa has a similar structure with BERT where each token is given an embedding and an attention mask (Liu et al., 2019b). We add the information of subjectivity scores to the embeddings and the presence of identity terms to attention masks, as exactly how we do with SS-BERT. To provide a fair comparison, we use an initial RoBERTa as the baseline. The implementations are similar to SS-BERT as detailed in Section 4. The RoBERTa model and the SS-RoBERTa model are built on our code of BERT and SS-BERT. We use the pre-trained “roberta-base” version of RoBERTa and adopt the subjectivity scores by TextBlob. The hyperparameter settings and dataset processing are the same as the experiments of BERT-based

**Table 8**

Examples of false positives assigned with high subjectivity scores and false negatives assigned with low subjectivity scores.

Comments	Subjectivity Scores	SS-BERT Prediction
1. @ mad ##dr ##11 there is not and never has been a country called palestine, micro ##bra ##in. and that land has been the home of jews for 3000 years.	1	False Positive
2. buying black jelly beans for my kids... in case they re cutting up and need a chill pill. they hate them.	0.7444	False Positive
3. @ love egypt @ washington ##post start by reading the historian will duran ##t. his book " our oriental heritage " describes the. muslim invasion.	0.6000	False Positive
4. It lesbian chronicle late at the office part 1 she ran her hands over maria ##m s breasts	0.6000	False Positive
5. homosexuality scientist discuss, if eugene was homosexual. see therefore: bern ##d - ulrich her ##ge ##mo ##ller , mann fur mann , page 210 magnus hirsch ##feld , die homosexual ##ita ##t des mann ##es und des wei ##bes, 1914 / 1984 , page 66 ##1 albert mo ##ll, be ##ru ##hm ##te homo ##se ##x ##uel ##le, wi ##es ##bad ##en, 1910, page 36	0.6667	False Positive
6. lo ##l no i meant its a shame its not in common . i have a lot of gay friends so you don ' t need to worry that i am against gay ##s : ) in terms of the afc article i can promise i won ##t repeat myself but if i have an advancing argument or need to respond to someone who is clearly wrong then i will do so but in a polite manner only . how about i ask you before i post something to it and if you agree i will ? the problem with this though is that i am up at stupid hours and so you won ' t be up . ok i ' ll look for	0.7238	False Positive
7. why white people used to say that sex was a sin used to be a mystery to me until i saw the children of browns and mixed race children popping up all around me .	0.1250	False Negative
8. saudi cleric believes earth does not rev ##ol ##ve around sun . there must be 50 ways to return to the 7th century with # islam as your expert guide	0	False Negative
9. this ata ##sca ##der ##o woman is accusing of physically abu ##sing two elderly relatives faces felony charges	0.1429	False Negative
10. rt @ ashes ##cho ##w : women ' s studies : learn how to tell other women they should study something other than women ' s studies rt @ eli ##ah r : http : / / t . . . .	0.3750	False Negative
11. no wonder this races achieve nothing .	0	False Negative
12. lack of compassion seems to go hand and hand with lack the of white blood.	0	False Negative

**Table 9**

Performance comparison between SS-RoBERTa and RoBERTa. The mean F1 score and its standard deviation, the mean false negative and the mean false positive are from 10 independently runs for each model presented. The better F1 result for each task is highlighted in bold.

DATA	RoBERTa				SS-RoBERTa			
	F1	std	False Negative	False Positive	F1	std	False Negative	False Positive
WS	0.5879	0.0233	48	52	<b>0.5955</b>	0.0253	46	55
Twitter 18k	0.7975	0.0052	96	151	<b>0.7992</b>	0.0072	93	152
Twitter 42k	0.7512	0.0095	70	262	<b>0.7543</b>	0.0081	70	256
Wiki	0.7429	0.0077	143	876	<b>0.7521</b>	0.0123	126	858

models. The results (Table 9) show that our method consistently improves the baseline RoBERTa across 4 datasets with the maximum improvement up to 1.29%.

It is somewhat surprising that BERT outperforms RoBERTa in our experiments as literature generally shows that RoBERTa performs better than BERT on many other NLP tasks (Liu et al., 2019b). While our results may appear inconsistent, some earlier studies reported similar results. For example, Baruah et al. (2020) have shown BERT (macro F1 0.6501) outperforms RoBERTa (0.6130) on the task of aggression identification. Mutanga, Naicker and Olugbara (2020) have shown BERT (0.73) outperforms RoBERTa (0.69) in terms of F1 on the task of hate speech detection. One possible explanation can be that RoBERTa is trained on a much bigger training corpus than BERT. The bigger training corpus of RoBERTa (16GB BERT corpus + 144GB corpus from CC-NEWS, OPENWEBTEXT and story-like style text from CommonCrawl) enhances the knowledge of the formal language. However, content in TCC datasets is written in much more informal language. Another possible reason is that RoBERTa removes the training objective of next sentence prediction from BERT. However, this could have helped the model consider the overall context when interpreting a text.



## 6. Conclusion

Identity term bias is commonly found as a limitation of the recent SOTA TCC methods. It affects TCC performance as it often leads to false positive predictions. However, only a few studies have investigated the identity term bias and they tackle the issue based on the same principle of paying less attention to the identity term. In this paper, we proposed a novel approach to tackle the identity term bias. This is achieved by training a model to pay additional attention to the subjectivity level of comments only when an identity term appears. Our approach utilizes the BERT embeddings structure to embed the information of both subjectivity levels and identity terms presence. We also proposed a new method to measure the subjectivity of a comment when it contains an identity term. The novelty here is that instead of lexicon-based methods in previous works, our method uses the semantic similarity to the relevant Wikipedia summary text of that identity term as a proxy to subjectivity.

Our extensive evaluation showed that our model SS-BERT outperforms SOTA methods on a wide range of TCC tasks. The results reveal that our method can mitigate the bias towards identity terms and reduce the false positive predictions effectively. Also, the results indicate that semantic similarity calculated by our method potentially reflects the subjectivity level of a comment. Our future work will look to address the limitations discussed before, i.e., developing an extensive identity term list and addressing inaccuracies in computing subjectivity. We are also interested in applying the Wikipedia based subjectivity measure to comments without identity terms. Another issue not addressed in this study and will be studied in the future is generalising our method to other pre-trained models which have different structures from BERT, such as Transformer-XL that does not include attention masks (Dai, Yang, Yang, Carbonell, Le and Salakhutdinov, 2019).

## References

- Amin, S., 1989. Eurocentrism, trans. russell moore.
- Badjatiya, P., Gupta, M., Varma, V., 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations, in: The World Wide Web Conference, pp. 49–59.
- Badjatiya, P., Gupta, S., Gupta, M., Varma, V., 2017. Deep learning for hate speech detection in tweets, in: Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee. pp. 759–760.
- Baruah, A., Das, K., Barbhuiya, F., Dey, K., 2020. Aggression identification in english, hindi and bangla text using bert, roberta and svm, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pp. 76–82.
- Bender, E.M., Friedman, B., 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6, 587–604.
- Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623.
- Bender, E.M., Koller, A., 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5185–5198.
- Benito, D., Araque, O., Iglesias, C.A., 2019. Gsi-upm at semeval-2019 task 5: Semantic similarity and word embeddings for multilingual detection of hate speech against immigrants and women on twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 396–403.
- Blaut, J.M., 1993. The colonizer's model of the world: Geographical diffusionism and Eurocentric history. volume 1. Guilford Press.
- Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H., 2020. Language (technology) is power: A critical survey of “bias” in nlp, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5454–5476.
- Blodgett, S.L., Green, L., O'Connor, B., 2016. Demographic dialectal variation in social media: A case study of african-american english, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1119–1130.
- Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T., 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29, 4349–4357.
- Borkan, D., Dixon, L., Sorensen, J., Thain, N., Vasserman, L., 2019. Nuanced metrics for measuring unintended bias with real data for text classification, in: Companion proceedings of the 2019 world wide web conference, pp. 491–500.
- Buckman, J., . Fair ml tools require problematic ml models. URL: [shorturl.at/frGY4](https://shorturl.at/frGY4).
- Burnap, P., Rana, O.F., Avis, N., Williams, M., Housley, W., Edwards, A., Morgan, J., Sloan, L., 2015. Detecting tension in online communities with computational twitter analysis. *Technological Forecasting and Social Change* 95, 96–108.
- Burnap, P., Williams, M.L., 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science* 5, 11.
- Caselli, T., Basile, V., Mitrović, J., Granitzer, M., 2021. HateBERT: Retraining BERT for abusive language detection in English, in: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), Association for Computational Linguistics, Online. pp. 17–25. URL: <https://aclanthology.org/2021.woah-1.3>, doi:10.18653/v1/2021.woah-1.3.
- Chavan, V.S., Shylaja, S., 2015. Machine learning approach for detection of cyber-aggressive comments by peers on social media network, in: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE. pp. 2354–2358.
- Clark, C., Yatskar, M., Zettlemoyer, L., 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4069–4082.

- Conneau, A., Lample, G., 2019. Cross-lingual language model pretraining, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 7059–7069.
- ConversationAI, 2017. Toxic comment classification challenge: Identify and classify toxic online comments. URL: <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge>.
- Dai, Z., Yang, Z., Yang, Y., Carbonell, J.G., Le, Q., Salakhutdinov, R., 2019. Transformer-xl: Attentive language models beyond a fixed-length context, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2978–2988.
- Davidson, T., Bhattacharya, D., Weber, I., 2019. Racial bias in hate speech and abusive language detection datasets, in: Proceedings of the Third Workshop on Abusive Language Online, pp. 25–35.
- Del Vigna, F., Cimino, A., Dell’Orletta, F., Petrocchi, M., Tesconi, M., 2017. Hate me, hate me not: Hate speech detection on facebook, in: Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), pp. 86–95.
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186.
- Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L., 2018. Measuring and mitigating unintended bias in text classification, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 67–73.
- Emmery, C., Verhoeven, B., De Pauw, G., Jacobs, G., Van Hee, C., Lefever, E., Desmet, B., Hoste, V., Daelemans, W., 2021. Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. *Language Resources and Evaluation* 55, 597–633.
- Fortuna, P., Nunes, S., 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51, 85.
- Founta, A.M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., Kourtellis, N., 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *ICWSM*.
- Gardner, M., Artzi, Y., Basmov, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., et al., 2020. Evaluating models’ local decision boundaries via contrast sets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pp. 1307–1323.
- Garrido-Muñoz, I., Montejó-Ráez, A., Martínez-Santiago, F., Ureña-López, L.A., 2021. A survey on bias in deep nlp. *Applied Sciences* 11, 3184.
- de Gibert, O., Perez, N., García-Pablos, A., Cuadros, M., 2018. Hate speech dataset from a white supremacy forum, in: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pp. 11–20.
- Gitari, N.D., Zuping, Z., Damien, H., Long, J., 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering* 10, 215–230.
- Halevy, M., Harris, C., Bruckman, A., Yang, D., Howard, A., 2021. Mitigating racial biases in toxic language detection with an equity-based ensemble framework, in: *Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–11.
- Hardt, M., Price, E., Srebro, N., 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29.
- Hovy, D., Spruit, S.L., 2016. The social impact of natural language processing, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 591–598.
- Hu, M., Lim, E.P., Sun, A., Lauw, H.W., Vuong, B.Q., 2007. Measuring article quality in wikipedia: models and evaluation, in: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 243–252.
- Huo, H., Iwaihara, M., 2020. Utilizing bert pretrained models with various fine-tune methods for subjectivity detection, in: *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, Springer, pp. 270–284.
- Kaushik, D., Hovy, E., Lipton, Z.C., 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.
- Kennedy, B., Jin, X., Davani, A.M., Dehghani, M., Ren, X., 2020. Contextualizing hate speech classifiers with post-hoc explanation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5435–5442.
- Kennedy, B., Kogon, D., Coombs, K., Hoover, J., Park, C., Portillo-Wightman, G., Mostafazadeh, A., Atari, M., Dehghani, M., 2018. A typology and coding manual for the study of hate-based rhetoric.
- Kim, Y., 2014. Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, pp. 1746–1751. URL: <https://aclanthology.org/D14-1181>, doi:10.3115/v1/D14-1181.
- Kittur, A., Kraut, R.E., 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination, in: Proceedings of the 2008 ACM conference on Computer supported cooperative work, pp. 37–46.
- Kovács, G., Alonso, P., Saini, R., 2021. Challenges of hate speech detection in social media. *SN Computer Science* 2, 1–15.
- Kumar, R., Reganti, A.N., Bhatia, A., Maheshwari, T., 2018. Aggression-annotated corpus of hindi-english code-mixed data, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).
- Kumar, V., Dabas, A., Jain, I., Pawar, K., 2021. A comprehensive analysis of deep learning techniques for documentation classification, in: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), IEEE, pp. 228–235.
- Kwok, I., Wang, Y., 2013. Locate the hate: detecting tweets against blacks, in: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, pp. 1621–1622.
- Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M., Sabharwal, A., Choi, Y., 2020. Adversarial filters of dataset biases, in: International Conference on Machine Learning, PMLR, pp. 1078–1088.
- Lin, C., He, Y., Everson, R., 2011. Sentence subjectivity detection with weakly-supervised learning, in: Proceedings of 5th International Joint Conference on Natural Language Processing, pp. 1153–1161.
- Lissack, M., 2021. The slodderwetenschap (sloppy science) of stochastic parrots — a plea for science to not take the route advocated by gebru and bender. URL: <https://michael-lissack.medium.com/the-slodderwetenschap-sloppy-science-of-stochastic-parrots-a-plea-for-science-to-not-take-the-b74fcf50dcce>. last accessed 01 Jan 2022.

- Liu, H., Burnap, P., Alorainy, W., Williams, M.L., 2019a. Fuzzy multi-task learning for hate speech type identification, in: The World Wide Web Conference, ACM. pp. 3006–3012.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019b. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 .
- Loria, S., 2018. textblob documentation. Release 0.15 2.
- Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., Zhang, G., 2015. Transfer learning using computational intelligence: a survey. Knowledge-Based Systems 80, 14–23.
- Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F.Å., Lanamäki, A., 2015. “the sum of all human knowledge”: A systematic review of scholarly research on the content of wikipedia. Journal of the Association for Information Science and Technology 66, 219–245.
- Mozafari, M., Farahbakhsh, R., Crespi, N., 2019. A bert-based transfer learning approach for hate speech detection in online social media, in: International Conference on Complex Networks and Their Applications, Springer. pp. 928–940.
- Mozafari, M., Farahbakhsh, R., Crespi, N., 2020. Hate speech detection and racial bias mitigation in social media based on bert model. PloS one 15, e0237861.
- Mutanga, R., Naicker, N., Olugbara, O.O., 2020. Hate speech detection in twitter using transformer methods. International Journal of Advanced Computer Science and Applications 11.
- Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y., 2016. Abusive language detection in online user content, in: Proceedings of the 25th international conference on world wide web, International World Wide Web Conferences Steering Committee. pp. 145–153.
- Pamungkas, E.W., Basile, V., Patti, V., 2021a. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. Information Processing & Management 58, 102544.
- Pamungkas, E.W., Basile, V., Patti, V., 2021b. Towards multidomain and multilingual abusive language detection: a survey. Personal and Ubiquitous Computing , 1–27.
- Pang, B., Lee, L., 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), pp. 271–278.
- Park, J.H., Shin, J., Fung, P., 2018. Reducing gender bias in abusive language detection, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2799–2804.
- Prost, F., Qian, H., Chen, Q., Chi, E.H., Chen, J., Beutel, A., 2019. Toward a better trade-off between performance and fairness with kernel-based distribution matching. arXiv preprint arXiv:1910.11779 .
- Ribeiro, M.T., Wu, T., Guestrin, C., Singh, S., 2020. Beyond accuracy: Behavioral testing of nlp models with checklist, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4902–4912.
- Rogers, A., 2021. Changing the world by changing the data, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online. pp. 2182–2194. URL: <https://aclanthology.org/2021.acl-long.170>, doi:10.18653/v1/2021.acl-long.170.
- Rogers, A., Baldwin, T., Leins, K., 2021. ‘just what do you think you’re doing, dave?’ a checklist for responsible data use in nlp, in: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 4821–4833.
- Salawu, S., He, Y., Lumsden, J., 2017. Approaches to automated detection of cyberbullying: A survey. IEEE Transactions on Affective Computing .
- Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., Prabhakaran, V., 2021. Re-imagining algorithmic fairness in india and beyond, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 315–328.
- Sankarasubramanian, Y., Ramanathan, K., Ghosh, S., 2014. Text summarization using wikipedia. Information Processing & Management 50, 443–461.
- Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N.A., 2019. The risk of racial bias in hate speech detection, in: Proceedings of the 57th annual meeting of the association for computational linguistics, pp. 1668–1678.
- Schmidt, A., Wiegand, M., 2017. A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pp. 1–10.
- Sebastiani, F., Esuli, A., 2006. Sentiwordnet: A publicly available lexical resource for opinion mining, in: Proceedings of the 5th International Conference on Language Resources and Evaluation, pp. 417–422.
- Tan, Y.C., Celis, L.E., 2019. Assessing social and intersectional biases in contextualized word representations, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 13230–13241.
- Tatman, R., 2017. Gender and dialect bias in youtube’s automatic captions, in: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, pp. 53–59.
- Vaidya, A., Mai, F., Ning, Y., 2020. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection, in: Proceedings of the International AAAI Conference on Web and Social Media, pp. 683–693.
- Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., Hoste, V., 2018. Automatic detection of cyberbullying in social media text. PloS one 13, e0203794.
- Vidgen, B., Thrush, T., Waseem, Z., Kiela, D., 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1667–1682.
- Wallerstein, I., 1997. Eurocentrism and its avatars: The dilemmas of social science. Sociological bulletin 46, 21–39.
- Waseem, Z., 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter, in: Proceedings of the First Workshop on NLP and Computational Social Science, Association for Computational Linguistics, Austin, Texas. pp. 138–142. URL: <http://aclweb.org/anthology/W16-5618>.
- Waseem, Z., Hovy, D., 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, pp. 88–93.

- Wiegand, M., Ruppenhofer, J., Kleinbauer, T., 2019. Detection of abusive language: the problem of biased datasets, in: Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers), pp. 602–608.
- Wu, Z., Ong, D.C., 2021. On explaining your explanations of bert: An empirical study with sequence classification. arXiv preprint arXiv:2101.00196 .
- Xia, M., Field, A., Tsvetkov, Y., 2020. Demoting racial bias in hate speech detection, in: Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, pp. 7–14.
- Xu, A., Pathak, E., Wallace, E., Gururangan, S., Sap, M., Klein, D., 2021. Detoxifying language models risks marginalizing minority voices, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2390–2397.
- Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E., 2016. Hierarchical attention networks for document classification, in: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp. 1480–1489.
- Ye, S., Chua, T.S., Lu, J., 2009. Summarizing definition from wikipedia, in: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP, pp. 199–207.
- Zhang, B.H., Lemoine, B., Mitchell, M., 2018a. Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 335–340.
- Zhang, H., Sneyd, A., Stevenson, M., 2020. Robustness and reliability of gender bias assessment in word embeddings: The role of base pairs, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pp. 759–769.
- Zhang, W., Yu, C., 2006. Uic at trec 2006 blog track: a notebook paper. In Proc. of TREC 2006 .
- Zhang, W., Yu, C., Meng, W., 2007. Opinion retrieval from blogs, in: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 831–840.
- Zhang, Z., Robinson, D., Tepper, J., 2018b. Detecting hate speech on twitter using a convolution-gru based deep neural network, in: European semantic web conference, Springer. pp. 745–760.
- Zhao, Z., Zhang, Z., Hopfgartner, F., 2019. Detecting Toxic Content Online and the Effect of Training Data on Classification Performance. Technical Report. EasyChair.
- Zhao, Z., Zhang, Z., Hopfgartner, F., 2021. A comparative study of using pre-trained language models for toxic comment classification, in: Companion Proceedings of the Web Conference 2021, pp. 500–507.
- Zhou, X., Sap, M., Swayamdipta, S., Choi, Y., Smith, N.A., 2021. Challenges in automated debiasing for toxic language detection, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 3143–3155.

Zhixue Zhao is a final year PhD student at the Information School of the University of Sheffield. Her PhD research is focusing on transfer learning for toxic comment classification tasks. She starts working as a post-doc researcher in the Department of Computer Science, University of Sheffield, in January 2022. Her post-doc project focus on explainable artificial intelligence and disinformation discussion modelling.

Dr Ziqi Zhang is a lecturer at the Information School of University of Sheffield. His research focus on text mining on the Web, particularly methods for automatically creating and linking structured knowledge bases, and information extraction from the social media resources. Between late 2016 and 2017, he worked as a lecturer in Computer Science at the Computing and Technology Department, Nottingham Trent University.

Dr Frank Hopfgartner is a Senior Lecturer in Data Science at the Information School of University of Sheffield. His research to date can be placed in the intersection of information systems (e.g., information retrieval and recommender systems), content analysis and data science. He has (co-)authored over 150 publications in above mentioned research fields, including a book on smart information systems, various book chapters and papers in peer-reviewed journals, conferences and workshops. To date, he has successfully acquired over £1 Million in research funding from national and international sources to support his research.