KONVENS 2021
Konferenz zur Verarbeitung natürlicher Sprache
Conference on Natural Language Processing

Proceedings of the Multilingual and Multimodal Hate Speech Workshop 2021

6–9 September, 2021
Heinrich Heine University Düsseldorf
Düsseldorf, Germany

Foreword

As the number and availability of social media platforms grow, the spread of hate speech among online communities (such as Twitter, Facebook, Reddit, Youtube, and so on) is also dramatically increasing. While there is no internationally agreed-upon definition, hate speech is broadly defined as a speech targeted to a given community or group with the potential of inciting violence towards them (Jacobs et al. 2000, Walker 1994). Davidson et al. (2017) discriminate between hate speech (languages used to express hatred towards a target group and incites violence) and offensive speech (usage of rude, hurtful, derogatory, obscene, or insulting language to upset or embarrass people). In this workshop, we broadly define hate speech as "inappropriate language" that is used in online communities, which can be expressed via text, image, or video and could be ultimately handled using an automatic approach.

One of the particular aspects that we touched upon is the multilinguality. The current developments on multilingual transformer models attract NLP researchers in multiple domains. The recent work by Ghosh Roy et al. (2020) showcases how to build hate speech detection systems with a pre-trained multilingual Transformer-based text encoder. In this workshop, we would like to discuss the challenges, approaches, frameworks, and technologies that could facilitate hate speech detection in multilingual environments and the consequences and implications of hate speech detection approaches for multilingual setup.

Moreover, multimodal aspects, day by day, become an integral part of those above-mentioned communication mediums (Twitter, Reddit, Facebook, etc.). If the message producer provides two sources of information together to deliver his/her message, then it suffices to assume that the meaning is distributed into both modalities to some degree. Particularly, in determining whether a multimodal tweet or memes accompanied with an image or video carries hateful content, uni-modal approaches can easily fail in case of cynicism: e.g. a tweet/meme with a very innocent-looking text accompanied by a very targeted and offensive image content or the other way round.

The goal of this workshop is to bring researchers with experience in different domains and languages together to

- Discuss the latest development towards the detection and counter-speech research on hate speech
- Bring the multilingual and multimodal aspects into the foreground
- Facilitate networking and encourage collaboration
- Create a future avenue for multimodal, multilingual, and cross-lingual hate speech research

The proceedings and the program are available from the workshop website (https://sites.google.com/view/mmhs2021/home).

Düsseldorf, September 2021

Özge Alaçam

Seid Muhie Yimam

Organizing Committee:

Özge Alaçam (Universität Hamburg)
Seid Muhie Yimam (Universität Hamburg)

Program Committee:

Abhik Jana (Universität Hamburg)
Abinew Ali Ayele (BiT, Ethiopia)
Binny Mathew (IIT Kharagpur, India)
Chris Biemann (Universität Hamburg)
Darina Gold (Universität Duisburg-Essen)
Punyajoy Saha (IIT Kharagpur, India)
Torsten Zesch (Universität Duisburg-Essen)
Xintong Wang (Universität Hamburg)

Invited Speakers:

Torsten Zesch (Universität Duisburg-Essen) on Multimodal Hate Speech
Animesh Mukherjee (IIT Kharagpur, India) on Multilingual Hate Speech
Sahana Udupa (LMU Munich) on Social Aspects of Hate [Extreme] Speech

Table of Contents

GerMemeHate: A Parallel Dataset of German Hateful Memes Translated from English

Darina Gold, Piush Aggarwal, and Torsten Zesch
Language Technology Lab
University of Duisburg-Essen
firstname.lastname@uni-due.de

Abstract

Hateful memes constitute an important part of online harassment, but are challenging due to their visio-linguistic nature and because annotated datasets are rare for languages other than English. We present GerMemeHate, a parallel dataset of memes translated from English. We describe the annotation process, analyze the resulting dataset, and evaluate how well English detection models can be transferred to German.

1 Introduction

Detecting hateful memes (which combine visual and textual information) is a relatively new task (Kiela et al., 2020) within the wider area of hate speech detection (Basile et al., 2019; Burnap and Williams, 2014). While there are annotated datasets of hateful memes for English (Kiela et al., 2020), Italian (Miliani et al., 2020), and Tamil (Suryawanshi and Chakravarthi, 2021), we are not aware of any German dataset. Instead of collecting German memes, we take English memes and manually translate the textual part into German. In this way, we are constructing a parallel meme corpus.

While translating memes is challenging due to potential cultural mismatches and the resulting selection effects, the resulting parallel corpus of memes might also shed some light on the interaction between textual and visual features, as only the text is changing while the picture (and possibly the detection part) stays the same. Translation also forces us to be more explicit about the underlying principles of memes and to make first steps towards a richer annotation scheme that goes beyond a binary hatefulness label.

German hate speech detection as well as corpora to evaluate its performance has been of interest in the recent years (Ross et al., 2016; Benikova et al., 2017; Wojatzki et al., 2018; Struß et al., 2019; Wiegand et al., 2018). Furthermore, there have been several translations of English sentiment datasets to German (Waltinger, 2010a,b; Remus et al., 2010). However, to our knowledge, there has been no German meme data set and no translation of a meme dataset overall.

Our contribution in this paper is threefold: 1) a methodology to create a translated meme data set, 2) GerMemeHate, a German hate meme data set, and 3) the analysis of the dataset and transfer learning experiments on the newly annotated dataset. We explore the possibility of translating a hate meme dataset as it on the one hand is a potential way to create datasets in languages other than English, without the initially very costly prefiltering process as described by Kiela et al. (2020). On the other hand, through the richer annotation process, we are able to show the culture-specific nature of hate memes themselves as well as their perception.

Our labels and translations for 230 memes, together with the corresponding meme ID are publicly available under CC-BY license.[1]

2 Parallel Corpus Construction

In this section, we describe the process of constructing the parallel English-German meme dataset.

2.1 Source Corpus

We base our study on the English meme data by Kiela et al. (2020), which is often called *Facebook Meme Corpus*. The Facebook Meme corpus started with a set of 162k memes, which was filtered, reconstructed, and annotated in several steps. In the final set there are exactly 10k memes, all of which have been synthetically re-created (for copyright reasons) or newly created in the case of *confounders*. The hatefulness rating on a ternary

[1] https://github.com/MeDarina/GerMemeHate
We cannot publish the full memes, as the license of the original dataset allows no redistribution.

Proceedings of the Multilingual and Multimodal Hate Speech Workshop 2021, pages 1–6,
Düsseldorf, Germany, September 6, 2021.

scale was performed by 3 annotators had a Cohens κ of .68 and was transformed to a binary rating.

A special property of the source corpus is that it contains so-called *confounders*, where the image and/or text of the original meme was changed so that it is not hateful anymore. While they might be useful for the training process and make the development and test sets 'harder', confounders pose a challenge to the manual creation of a parallel corpus. Some of the confounders are unnatural, as they do not seem to convey a clear communicative intent (see Figure 1b).[2]

As *all translation is interpretation*, confounders are hard to work with, we argue that confounders should be treated as a kind of artificial data augmentation strategy, and not be translated. Thus, we remove confounders by looking for memes with exactly the same image or text. Of those near-duplicates, we keep only the ones labeled as hateful.

Another challenge with the dataset (or actually any meme dataset), is that many of the memes are very culture-specific and hard to make sense of without near-native language capabilities and a deep understanding of current social and political issues in the US. We are arguing that for hate speech datasets (and meme datasets in particular), it is not sufficient to say it is an 'English' dataset, while it is actually a US-centric dataset that happens to be written in English.

As one of our research questions is to do an evaluation of transfer learning settings, we only use a part of the development set for constructing the parallel corpus, but leave the training set as is. From the originally 300 randomly chosen memes chosen from the development dataset after an automatic filtering for confounders, 70 (23%) were filtered out due to different reasons making them unsuitable for translation.

2.2 Annotation & Translation

The annotation and translation is carried out by two native German speakers, who are fluent in English. The **translation** is intended to be as close to the original as possible, while keeping ambiguities, as well as historical and cultural references. Additionally to the manual translation, we add an automatic translations layer using Google Translate.[3]

[2]Examples are for illustration and not from the dataset. None of the examples in any way represent the opinion of the authors.

[3]https://cloud.google.com/translate

We additionally annotate the **image-text relation** found in a meme, as we want to understand how many memes can only be understood through processing image and text together in a truly multimodal fashion.

Similar to Marsh and White (2003); Chen et al. (2013); Vempala and Preoţiuc-Pietro (2019), we distinguish the image-text relation types according to the closeness of the conceptual relationship between image and text. More specifically, we distinguish between three image-text relation types that are used similarly in other multi-modal research (Chen et al., 2013; Taib and Ruiz, 2007; Ruiz et al., 2006): REDUNDANT if the image does not relate nor contribute to the meme meaning (e.g. in Figure 1a the text says "obama voters" and the image shows a smiley); SUPPORTING if the image confirms, and enhances the texts message, but the message would stay the same without it at large (e.g. in Figure 1b, the text says "obama voters" and the image shows a crowd of people cheering for Obama) ; and CONTRIBUTIVE if the meaning of the meme can only be understood when interpreting text and image together (e.g. in Figure 1c, the ambiguity is a result of both the image and the text – the text says "obama voters" and the image shows monkeys, indicating that people voting for Obama have properties of monkeys i.e. do not have human intelligence).

Ambiguity, whether syntactic, lexical, semantic or of another type, is one of the linguistic devices used in humor (Bucaria, 2004; Bekinschtein et al., 2011; Kagan, 2020) and thus also in memes. It poses two difficulties for us: 1) we have to understand it, and 2) it has to be translated. In our annotation, we mark whether the original meme contains ambiguity and whether it could be translated.

In the original dataset, **hatefulness** is annotated as a binary label. We re-annotate the label on the basis of the translation using the same definition as the original Facebook Meme Corpus. The difference of our re-annotation to the original annotation might be indicative of the amount of culture-specific memes in the original dataset.

2.3 Final Corpus

After the annotation, we filter memes (i) that annotators did not understand (e.g. because a confounder –that we missed in the initial filtering– lacks any coherence between image and text), or (ii)

(a) NON-HATE, REDUNDANT

(b) NON-HATE, SUPPORTING

(c) HATE, CONTRIBUTIVE

Figure 1: Annotation Scheme for generating parallel hate meme corpus. a) REDUNDANT denote no linkage between image and text; SUPPORTING denote similar semantic inference from image and text; and CONTRIBUTIVE denote additional inference from the image to the text.

that annotators marked as non-translatable (mainly due to making use of an ambiguity that does not translate well to German). For example, one of the memes used the (near) homophones *aunts* and *ants*, but annotators found it impossible to carry that over to German. The final parallel corpus contains 230 memes.

3 Corpus Analysis

Translation We treat our manual translations as a real gold standard. Translations of both human annotators are generally quite similar. We find that the BLEU-3 score between the annotators is .39 (see Figure 2). The BLEU score between each human and the automatic machine translation is a bit lower (.35 and .32). The difference in BLEU between manual and automatic translation is not big, so that we expect a model based on translated memes to work reasonably well.

However, we find that a qualitative analysis shows that manual translation are often better in subtle but crucial ways. For example, in a meme that makes a reference to the Holocaust by talking of "concentrating ... in a camp", 'camp' should be translated as 'Lager' instead of 'Camp' to retain the reference.

Ambiguity After filtering 4% of ambiguous memes from the original dataset, as their ambiguity could not be translated according to both annotators, our annotators agree on 4% of the memes containing textual ambiguity in our final corpus. As we are not aware of other meme datasets making a similar analysis, we cannot make any direct comparisons, but the numbers seems small compared to personal experience with hateful memes.

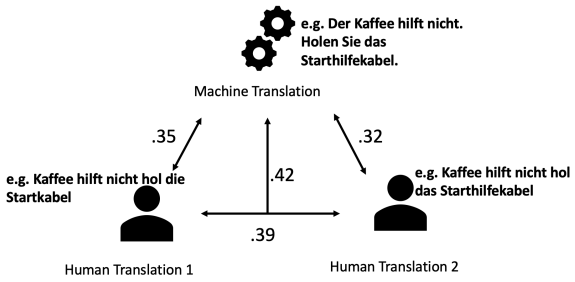

Figure 2: Average BLEU-3 scores between manual and machine translators.

Hatefulness We obtain a κ of .67 on HATEFULNESS, which is comparable to the .68 of the original dataset. The agreement between the original English HATE labels and the ones annotated on the filtered German translated version is a bit lower with a κ of .54.

After adjudication, 38% of German memes are marked as hateful, compared to 52% in the parallel English set. So our German annotators were considerably less likely to label a meme as hateful, even if working with the same annotation manual.

If we take a closer look at the disagreements, we find that the 45 false negatives (labeled as hateful in the English version, but non-hateful in the German version), were mainly related to 1) cultural knowledge, some of which are not known (In meme 50261: PoC not being able to swim) or not considered hateful by our German annotators (4 memes implied Asians eating dogs)[4] or 2) references to religion (7 of the 45 memes targeted religion or religious groups).[5] There were only 9

[4] memes 24098, 48296, 58672, 65832

[5] memes 4857, 26453, 34018, 41890, 46920, 70953, 83954

Approach	re-annotated F_1	original F_1
Majority Class	.39	.34
English (Extended VL-Bert)	.62	.66
German (Zero-shot)	.54	.54
German (AMT)	.55	.56

Table 1: Transfer learning results (macro F_1). First column are our re-annotated hatefulness label, second column is the translated meme with the original hatefulness annotation.

false positives (which were labeled as hateful in the German version, but not in the English one). 4 of those 9 were on the topic of the second world war, where German law is known to take a strict stance.[6]

Image-Text-Relation We obtain a κ of .55 for the IMAGE-TEXT relation. Almost all disagreement is between the classes CONTRIBUTIVE and SUPPORTING. This shows that there is further need to refine the annotation scheme.

We adjudicate disagreements by always taking the 'more informed' decision, i.e. if one annotator said CONTRIBUTIVE this takes precedence over REDUNDANT, as we assume that the other annotator probably did not understand the meme properly. Using this scheme, 69% of memes are truly multimodal (CONTRIBUTIVE), 28% have an image-text relation where the image supports the textual content (SUPPORTING), and only 3% of memes are REDUNDANT.

4 Transfer-learning Experiments

In our experiments, we envision a situation where we want to transfer the English detection model to another language for which no annotated training data is available. We cannot directly use the best-performing system (Zhu, 2020), because it is built over the pre-trained model where English specific settings are used. Therefore, we train a German model based on translated meme texts. We use the Google translation API to translate all memes from the training set of the Facebook dataset (Kiela et al., 2020). This training set does not include any memes used in the parallel corpus used for evaluation.

[6] memes 14865, 27543, 38195, 79042

4.1 Models and Features

To validate the appropriateness of our parallel subset, we compute also results on the parallel English memes. For this purpose, we experiment with the Extended VL-Bert classification model (Aggarwal et al., 2021).

For German, we use existing pre-trained models for both modalities to extract the features that are then concatenated and provided as input to a fully connected feed forward network for final classification.

For text features, we experiment with BERT-base multilingual cased model (Devlin et al., 2018), which is pretrained on 104 languages having 110M parameters, and with German BERT model (Chan et al., 2020) having the same number of parameters pretrained on the German Wikipedia dump (6 GB of raw text files), the OpenLegalData dump (2.4 GB) and news articles (3.6 GB).

For image features, we use a deep convolutional based neural network (22 layers) called Googlenet (Szegedy et al., 2015) pretrained on the ImageNet dataset[7]. We also use two variants of ResNet (He et al., 2016) which are also pretrained on ImageNet. The first variant is a 152 layered residual network having 60M parameters called ResNet-152. The second variant (Wide ResNet-101-2) has 127M parameters with an architecture similar to ResNet with additional hidden layers.

We also apply a **Zero-Shot** approach, where we use a BERT-base multi-lingual cased pretrained language model to extract English meme text features, Wide ResNet-101-2, and Googlenet for image feature extraction (Choi et al., 2021).

Furthermore, we report the **majority class baseline**.

4.2 Results & Analysis

First, to ensure that our parallel corpus is not a biased selection from the original dev set, we compare the performance of the English classification model on the full dataset and the English part of our parallel dataset. The model yields a macro $F1$ score of .66 in both cases. This shows that our subset is roughly equivalent in classification difficulty.

Table 1 presents the full set of results. We show performance on two different sets of hatefulness labels, where 're-annotated' is from our German annotators and 'original' is the labels provided in the

[7] https://image-net.org/

Facebook dataset, but copied over to the translated memes.

The best German Zero-shot result (BERT-base multi-lingual cased & GoogleNet) is on par with the best German AMT result (GermanBert & GoogleNet). If choosing between those two choices, zero-shot should be preferred as it is computationally cheaper.

Results on the German part of the parallel corpus are consistently lower than on the English part, even if we use the original hate labels. This indicates that the availability of large pre-trained visual-linguistic models such as VL-Bert for English is an advantage that German is lacking at the moment.

Although, all the aforementioned models beat the majority class baselines, they still have huge room of improvement.

5 Conclusions & Future Work

Our experiments show that translating a meme dataset is challenging. Probably, it would be better to collect original examples from the target language. As memes are often international, a subset could also be truly parallel.

Nevertheless, we have shown that it is possible to translate memes and created the first German hate meme corpus (with parallel English memes). The huge differences between the original hate label and our re-annotation are further evidence of the well-known problem of subjectivity in hate speech annotation.

Future work should concentrate on richer annotation schemes for memes and collecting a truly multi-lingual (and potentially parallel) meme dataset.

References

Piush Aggarwal, Michelle Espranita Liman, Darina Gold, and Torsten Zesch. 2021. VL-BERT+: Detecting protected groups in hateful multimodal memes. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 207–214, Online. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, MN, USA.

Tristan A. Bekinschtein, Matthew H. Davis, Jennifer M. Rodd, and Adrian M. Owen. 2011. Why clowns taste funny: the relationship between humor and semantic ambiguity. *Journal of Neuroscience*, 31(26):9665–9671.

Darina Benikova, Michael Wojatzki, and Torsten Zesch. 2017. What does this imply? examining the impact of implicitness on the perception of hate speech. In *International Conference of the German Society for Computational Linguistics and Language Technology*, pages 171–179, Berlin, Germany. Springer.

Chiara Bucaria. 2004. Lexical and syntactic ambiguity as a source of humor: The case of newspaper headlines. *Humor*, 17(3):279–309.

Pete Burnap and Matthew L. Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: interpretation and communication for policy decision making. In *Internet, Policy and Politics Conference*, pages 1–18, Oxford, UK.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tao Chen, Dongyuan Lu, Min-Yen Kan, and Peng Cui. 2013. Understanding and classifying image tweets. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 781–784, Barcelona, Spain.

Hyunjin Choi, Judong Kim, Seongho Joe, Seungjai Min, and Youngjune Gwon. 2021. Analyzing zero-shot cross-lingual transfer in supervised nlp tasks. In *25th International Conference on Pattern Recognition (ICPR)*, pages 9608–9613, Milan, Italy.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA.

Olga Kagan. 2020. Humor creation and the ambiguity of morpho-syntactic phenomena. *Russian Linguistics*, 44(1):59–78.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624, Virtual. Curran Associates, Inc.

Emily E. Marsh and Marilyn Domas White. 2003. A taxonomy of relationships between images and text. *Journal of Documentation*, 59(6):647–672.

Martina Miliani, Giulia Giorgi, Ilir Rama, Guido Anselmi, and Gianluca E. Lebani. 2020. DANKMEMES@ EVALITA2020: The meme-ing of life: memes, multimodality and politics. In *Proceedings of EVALITA*, Online. CEUR.org.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS-A Publicly Available German-language Resource for Sentiment Analysis. In *Proceedings of LREC*, pages 1168–1171, Valletta, Malta.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. pages 6–9, Bochum, Germany.

Natalie Ruiz, Ronnie Taib, and Fang Chen. 2006. Examining the Redundancy of Multimodal Input. In *Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments*, pages 389–392, Sydney, Australia.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, Manfred Klenner, et al. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language. pages 354–365, Erlangen-Nürnberg, Germany.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on Troll Meme Classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 126–132.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper With Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, Boston, MA, USA.

Ronnie Taib and Natalie Ruiz. 2007. Integrating semantics into multimodal interaction patterns. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 96–107, Brno, Czech Republic. Springer.

Alakananda Vempala and Daniel Preoţiuc-Pietro. 2019. Categorizing and Inferring the Relationship between the Text and Image of Twitter Posts. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 2830–2840, Florence, Italy.

Ulli Waltinger. 2010a. Germanpolarityclues: A lexical resource for german sentiment analysis. In *Proceedings of LREC*, pages 1638–1642, Valletta, Malta.

Ulli Waltinger. 2010b. Sentiment analysis reloaded: A comparative study on sentiment polarity identification combining machine learning and subjectivity features. In *Proceedings of the 6th International Conference on Web Information Systems and Technologies*, pages 203–210, Valencia, Spain.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of KONVENS*, pages 1–10, Vienna, Austria.

Michael Wojatzki, Tobias Horsmann, Darina Gold, and Torsten Zesch. 2018. Do Women Perceive Hate Differently: Examining the Relationship Between Hate Speech, Gender, and Agreement Judgments. In *Proceedings of KONVENS*, pages 110–120, Vienna, Austria.

Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution.

Offensive Language Detection in Semitic Languages

Marina Litvak and **Natalia Vanetik** and **Yaser Nimer** and **Abdulrhman Skout**
Sami Shamoon College of Engineering,
Beer-Sheba, Israel
{marinal,natalyav,yaserni,abdalsk}@ac.sce.ac.il

Abstract

Unfortunately, offensive language in social media is a common phenomenon nowadays. It harms many people and vulnerable groups. Therefore, automated detection of offensive language is in high demand and it is a serious challenge in multilingual domains. Various machine learning approaches combined with natural language techniques have been applied for this task lately. This paper contributes to this area from several aspects: (1) it introduces two new annotated datasets collected from Facebook and from Twitter in Hebrew and Arabic languages[1], respectively; (2) it proposes a publicly available system for offensive text detection based on supervised learning; (3) it reports evaluation results on these datasets using multiple machine learning algorithms. The website is publicly available at https://off-lang-2.herokuapp.com/. Both datasets can be downloaded from https://github.com/AbdulrhamnSkout/Dataset.

1 Introduction

Most of the recent papers on offensive language detection report about the application of deep neural networks (DNNs), such as LSTMs, RNNs, CNNS, GRUs combined with word embeddings, and transformers, for separating offensive language from legitimate texts (Zampieri et al., 2019). In the last couple of years, transformer models like ELMO (Peters et al., 2018) and BERT (Devlin et al., 2018) have been most popular and successful for offensive language identification (Liu et al., 2019; Ranasinghe et al., 2019).

The clear majority of the offensive detection studies deal with English partially because most available annotated datasets contain English data (Zampieri et al., 2019).

The attention of international communities to this task emphasizes its "multilingual challenge"—many researchers contributed to this area by developing multilingual methodologies and annotated corpora in multiple languages. For example, such languages as Arabic (Mohaouchane et al., 2019), Dutch (Tulkens et al., 2016), French (Chiril et al., 2019), Turkish (Çöltekin, 2020), Danish (Sigurbergsson and Derczynski, 2020), Greek (Pitenis et al., 2020), Italian (Poletto et al., 2017), Portuguese (Fortuna et al., 2019), Slovene (Fišer et al., 2017), and Dravidian (Yasaswini et al., 2021) languages were explored for this task. Also, the multilingual methods and datasets for offensive language detection were proposed. Hate Speech and Offensive Content Identification (HASOC) 2019 (Mandl et al., 2019) and 2020 (Mandl et al., 2020) were dedicated to evaluate technology for finding Offensive Language and Hate Speech in multiple low-resource languages. HASOC 2019 provided Twitter posts for Hindi, German and English. HASOC 2020 has created test resources for Tamil and Malayalam in native and Latin script. Posts were extracted mainly from YouTube and Twitter. Both tracks have attracted much interest from over 40 research groups. In Ranasinghe and Zampieri (2020), authors addressed the multilinguality challenge by applying cross-lingual contextual word embeddings and transfer learning. They made predictions in low-resource languages, such as Bengali, Hindi, and Spanish.

Despite the great international effort, many low-resource languages got much less attention than others. For example, we found only one work on Hebrew language (Liebeskind and Liebeskind, 2018), but the authors do not share their dataset

[1]Using different social platforms for these languages is motivated by their popularity in various countries/communities. For example, Twitter is not very popular in Israel, while Facebook being the most popular.

Proceedings of the Multilingual and Multimodal Hate Speech Workshop 2021, pages 7–12,
Düsseldorf, Germany, September 6, 2021.

publicly[2]. Motivated by the lack of relevant resources and research for Hebrew, we decided to fill this gap by (1) creating a new annotated dataset for this language and (2) performing experiments on this dataset using multiple machine learning algorithms. Also, given a fact that Arabic and Hebrew belong to the same language family while Arabic has much more resources, we wanted to compare between their analyses with long-term aim of exploring cross-lingual transfer learning for the offensive language detection in Semitic languages.

Our contribution is multi-fold: (1) we introduce and share with the public two new annotated datasets of Facebook comments in Hebrew and Twitter comments in Arabic; (2) we introduce and share for unrestricted use a Web application that provides a real-time solution for offensive language detection; (3) we report the results of evaluation on two Semitic language-datasets using multiple supervised models.

2 New datasets

Here we present two datasets, that we collected in Semitic languages. We titled our datasets OLaH (Offensive Language in Hebrew) and OLaA (Offensive Language in Arabic), respectively. Table 1 shows the data statistics for the two datasets, including their partition to a train and a test sets.

2.1 OLaH Dataset

The data is a collection of comments from particular groups [3] in Facebook, where offensive language is frequently observed. We collected 2000 comments, written in Hebrew, using Graph API[4]. For retrieving relevant texts, we used a list of keywords, which are usually part of a typical offensive vocabulary, or describe domains usually containing offensive language in Hebrew. Figure 1 contains an example (but not final) list of such words.

2.2 OLaA Dataset

The data is a collection of comments from Twitter. We used 6000 of annotated comments from the existing dataset[5] and extended it by another 3000 comments, collected using Twitter API (Makice, 2009), using the same (binary) annotation approach. In total, we collected 9000 comments, written in Arabic. For retrieving relevant texts, we used a list of keywords, which are usually part of a typical offensive vocabulary or describe domains usually containing offensive language in Arabic.[6] The example of such words can be seen in Figure 1, right.

2.3 Data Annotation and Cleaning

The OLaH/OLaA datasets were annotated by three Hebrew/Arabic native speakers. Each comment was assigned two labels. In a case of disagreement between two annotators, the third one–controller–assigned the final label. Finally, the Kappa agreement between annotators was 0.82 and 0.75 for OLaH and OLaA, respectively.

To keep clean texts, we filtered out words in a foreign language (other than Hebrew or Arabic for OLaH and OLaA, respectively), numbers, URLs, punctuation marks. We also normalized words by removing repeating letters, transforming words like לדדדך (gooooo) to לך (go).

3 The method

Our method is based on a purely supervised approach, where every text is classified into one of two classes, based on a trained model. Models are trained on texts (training data) written in Hebrew or Arabic, collected and annotated as described in Section 2, and then applied on Hebrew or Arabic texts, respectively.

Our approach also tries to verify the intuitive assumption that offensive content usually carries some negative sentiment. Therefore, we use this information—sentiment polarity—to enrich text representation before classification. We hypothesize that, given a precise tool for the sentiment analysis (SA) and avoiding noise introduction, this additional particle of information may improve the classification performance.

3.1 Text representation and classification

Our approach to text representation and classification (depicted in Figure 2) consists

[2]The dataset was shared with us on GitHub upon our request from the authors and it will be used in further research.

[3]ynet, the shadow, 0404 , תנועת רגבים, ביתר ירושלים, וואלה , חמל, דיווח ראשוני, ביביסטים

[4]https://developers.facebook.com/docs/graph-api/

[5]https://raw.githubusercontent.com/AbeerAbuZayed/Hate-Speech-Detection_OSACT4-Workshop/master/Dataset/train_data.csv

[6]Please, note that some words are not a part of the typical offensive vocabulary, but, based on our observations, they are usually good indicators of a domain/topic containing offensive content.

Table 1: Dataset statistics

Name	Source	Lang	Len (min-max, avg)	Train	Test	Pos	Neg
OLaH	Facebook	He	(1-198, 11)	80%	20%	40.5%	59.5%
OLaA	Twitter	Ar	(1-84, 17)	80%	20%	28%	72%

Word in Hebrew	Translation
בושה	shame
אפס	zero
זו**נה	f***ing
זבל	trash
מחבל	terrorist
חמור	donkey (idiot)
הומו	gay
ביבי	Bibi (Netanyahu)
לפיד	Lapid (Yair)

Word in Arabic	Translation
يهود	Jewish
سني	Sunni
شيعي	Shiite
عربي	Arab
لقيط	bastard
ارهابي	terrorist
حمار	donkey (idiot)
دين	religions
كلب	dog

Figure 1: Examples of the search keywords for offensive content in Hebrew (left) and Arabic (right).

of the following steps:
(1) Representing comments with one of the following: (a) tf*idf vectors, where every comment is treated as a separate document; (b) extended tf*idf vectors with sentiment information—we used the pretrained HeBERT (Chriqui and Yahav, 2021) and AraBERT (Antoun et al., 2020) models for Hebrew and Arabic, respectively.
(2) Training and application of four ML supervised models (see Section 4.1) on both types of vectors.

3.2 System demonstration

Our system follows the standard web server-client architecture. The backend server is responsible for the input data preprocessing and classification, given pre-trained model. The server does not aim at training new models and datasets storage.

The front end provides a very basic and user-friendly interface for submitting text data and displaying the classification results. The displayed output depends on the use case. The system provides the end-user with four options: (1) selecting a language, (2) classifying one piece of text, (3) classifying a Twitter account based on the last 50 posts of a user, and (4) classifying each text in a dataset by uploading it to the system and downloading back the classification results. The classification label is displayed to the user on the front-end page in use case (2), while the percentage of comments that were recognized as offensive is displayed to the user in use cases (3) and (4).

4 Experiments

Our experiments aim at three goals: (1) to analyze two new datasets and show that their quality is sufficient for the training systems for offensive language detection; (2) to test our hypothesis and evaluate the gain (if any) of the SA to the offensive language detection by comparing performance with and without SA; (3) to select the best model for each language for integrating them within our system.

4.1 Models and baselines

We used RandomForest (RF) (Ho, 1995; Breiman, 2001), Support Vector Machine (SVM) (Cortes and Vapnik, 1995), Logistic Regression (LR) (Walker and Duncan, 1967), and XGBoost (XGB) (Chen and Guestrin, 2016), applied on BOW (tf*idf vectors) as baselines. The same models were applied on extended with SA labels vectors, denoted as RF_SA, SVM_SA, LR_SA, and XGB_SA, respectively. Also, we applied fine-tuned BERT models–HeBERT on OLaH and AraBERT (Antoun et al., 2020) on OLaA. Both models were trained on big datasets of Hebrew/Arabic texts but different tasks. Therefore, we fine-tuned them on our training data for both languages.[7]

[7]Fine-tuning of BERT models is not a part of our classification pipeline from Figure 2.

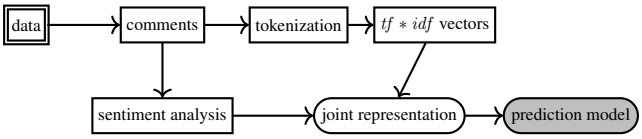

Figure 2: Text classification pipeline.

Table 2: The evaluation results

	OLaH		OLaA	
Model	**Acc**	**F1**	**Acc**	**F1**
RF	**0.783**	0.701	0.932	0.860
RF_SA	0.759	0.655	0.929	0.852
SVM	0.766	0.713	0.932	0.862
SVM_SA	0.771↑	0.724↑	0.931	0.862
LR	0.761	0.645	0.928	0.848
LR_SA	0.776↑	0.681↑	0.928	0.848
XGB	0.433	0.605	0.923	0.839
XGB_SA	0.724↑	0.639↑	0.929↑	0.854↑
BERT	0.775	**0.729**	**0.981**	**0.964**

4.2 Software

All baselines are implemented in sklearn (Pedregosa et al., 2011) python package. Our neural model is implemented with Keras (Chollet et al., 2015) with the TensorFlow backend (Abadi et al., 2015). Experiments were performed using Google Colaboratory service (Bisong, 2019). NumPy and Pandas libraries were used for data manipulation.

4.3 Evaluation Results

Table 2 shows the evaluation results (accuracy and F1 scores) for four baselines and BERT models on two datasets, with and without SA. Arrow (↑) denotes improvement, if detected, between each method scores with and without sentiment analysis. The best scores per dataset are marked in bold.

As it can be seen, the BERT models, despite being trained on different amount of training samples,[8] are superior for both datasets in terms of F1, meaning that RF obtained better accuracy on account of the "negative" (not offensive) class.

Only three models for Hebrew and one for Arabic (out of five[9]) demonstrate an improvement given sentiment labels. As such, there is a weak evidence approving a possible gain of sentiment information involved in the text representation vectors. It is also worth noting that the difference in SA gain in two languages can be explained by different outputs of the SA tools—while HeBERT produces polarity vectors with real values for three dimensions (positive, neutral, negative), AraBERT produces two-dimensional binary vectors, where each dimension stand for positive or negative sentiment. Obviously, binary values do not carry much information.

We can see that results on Arabic for all models are much better than on Hebrew. This outcome can be probably explained by the dataset quality – in contrast to the OLaH, OLaA was not collected from scratch and the existing dataset was extended by additional samples. There is also possible that preprocessing tools for Arabic are of a higher quality. As it is well known, text preprocessing affects very much the quality of further text analysis. In general, we can conclude that both datasets are of good quality because all supervised models provided reasonable results.[10]

Our error analysis have an evidence of incapability of tf-idf representation to capture sarcasm and irony presented in misclassified texts.

5 Conclusions and Future Work

This paper introduces two new annotated datasets for offensive language detection in Hebrew and Arabic languages and analyzes different supervised learning methods for offensive text detection in social media. We evaluate these methods on both datasets and conclude that classification accuracy is affected significantly by different training sizes of BERT for the two languages whereas adding sentiment information does not improve the results in most of the models. We also implemented a publicly available system for offensive text detection based on our best classification models.

In the future, we plan to explore more multilingual text representations and language models and apply transfer learning techniques for multilingual and cross-lingual analysis.

[8] AraBERT was pre-trained on 8.6 billion words, while HeBERT on 1 billion only.

[9] Both BERT models did not demonstrate improvement with sentiment labels (except precision for Hebrew), therefore we excluded their results from Table 2.

[10] We can make this conclusion because we avoided creating bias in our datasets, where all comments from particular groups belong to one class.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Ekaba Bisong. 2019. *Building machine learning and deep learning models on Google cloud platform*. Springer.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Patricia Chiril, Farah Benamara, Véronique Moriceau, Marlène Coulomb-Gully, and Abhishek Kumar. 2019. Multilingual and multitarget hate speech detection in tweets. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN-PFIA 2019)*, pages 351–360. ATALA.

François Chollet et al. 2015. Keras. https://github.com/fchollet/keras.

Avihay Chriqui and Inbal Yahav. 2021. Hebert & hebemo: a hebrew bert model and a tool for polarity analysis and emotion recognition. *arXiv preprint arXiv:2102.01909*.

Çağrı Çöltekin. 2020. A corpus of turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in slovene. In *Proceedings of the first workshop on abusive language online*, pages 46–51.

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Chaya Liebeskind and Shmuel Liebeskind. 2018. Identifying abusive comments in hebrew facebook. In *2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)*, pages 1–5. IEEE.

Ping Liu, Wen Li, and Liang Zou. 2019. Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 87–91.

Kevin Makice. 2009. *Twitter API: Up and running: Learn how to build applications with the Twitter API*. " O'Reilly Media, Inc.".

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for Information Retrieval Evaluation*, pages 29–32.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

Hanane Mohaouchane, Asmaa Mourhir, and Nikola S Nikolov. 2019. Detecting offensive language on arabic social media using deep learning. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 466–471. IEEE.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word

representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in greek. *arXiv preprint arXiv:2003.07459*.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate speech annotation: Analysis of an italian twitter corpus. In *4th Italian Conference on Computational Linguistics, CLiC-it 2017*, volume 2006, pages 1–6. CEUR-WS.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844.

Tharindu Ranasinghe, Marcos Zampieri, and Hansi Hettiarachchi. 2019. Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification. In *FIRE (Working Notes)*, pages 199–207.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for danish. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3498–3508.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in dutch social media. *arXiv preprint arXiv:1608.08738*.

Strother H Walker and David B Duncan. 1967. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. Iiitt@ dravidianlangtech-eacl2021: Transfer learning for offensive language detection in dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 187–194.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Multi-task Learning to Analyze the Influence of Offensive Language in Hate Speech Detection

Gretel Liz De la Peña Sarracén
Universitat Politècnica de València
gredela@posgrado.upv.es

Paolo Rosso
Universitat Politècnica de València
prosso@dsic.upv.es

Abstract

Hate speech detection has become the focus of lots of research in order to mitigate the negative effects the hatred can cause on the Internet. One of the problem is the unclear boundary with other related phenomena, such that datasets from different phenomena are often used interchangeably. In this work we propose a multi-task deep learning strategy that combines offensive language detection and hate speech detection. Offensive language is a closely related phenomenon which is widely studied as well. Our main idea is to investigate how hate speech detection is affected when datasets for this related task are consider. In this sense, we train a Transformer-based model on the two tasks, rather than considering hate speech as the unique optimization objective. Our experimental results suggest that this multi-task learning can slightly decrease the false negatives in the detection of hate speech detection, but can affect the general performance compared to a single-task learning strategy.

1 Introduction

Nowadays, hate speech is a major problem in social media. Therefore, the automatic detection of this phenomenon has become a trending topic (Schmidt and Wiegand, 2017; Naseem et al., 2020; Cao et al., 2020; Ousidhoum et al., 2019). According to (Fortuna and Nunes, 2018), hate speech can be defined as a language that attacks or incites hate against groups, based on specific characteristics such as religion, sexual orientation, or other. Besides, there are related phenomena such as offensive language and aggressive language, among which there is an unclear delimitation. In fact, in some works the same label is used to refer to the different types indistinguishably, whereas in others, different labels are used to refer to the same type of language. In this sense, (Poletto et al., 2020) compiled a glossary of terms and their definitions. The authors analyze the relation among different types of language, the way they have been addressed in the evaluation campaigns, and provide a wide view of resources centered on hate speech, including datasets and lexicons. This problem is analyzed also in (Davidson et al., 2017), where the separation of hate speech from offensive language is indicated as a key challenge. The authors performed experiments and the results suggest that conflating hate speech and offensive language can lead to mislabel offensive language as hateful due to overly broad definitions.

In this paper we study the relation between hate speech and offensive language by training a Transformer-based model with the multi-task learning (MTL) paradigm. In this sense, we consider losses for both tasks in the objective function. We share the same model parameters across the tasks instead of separately fine-tuning models.

- We propose a multi-task learning (MTL) model that uses the pre-trained BERT-base model to combine datasets from hate speech and offensive language detection tasks.
- We compare the performance of the MTL model with a single-task learning (STL).
- We use a transfer learning technique as alternative to the MTL paradigm in order to analyze how the use of datasets built for offensive language detection can influence hate speech detection.

The rest of this paper is organized as follows. Section 2 summarizes the related work and Section 3 presents our MTL model as well as the methodology we used to train and evaluate the model. Section 4 introduces the STL model and the transfer learning technique that we use. Then, Section 5 describes the experiments and presents a discussion of the results. Finally, Section 6 presents an error analysis and Section 7 concludes the paper.

2 Related Work

The techniques used in the detection of hate speech and related concepts range from traditional methods based on the use of lexical features to methods based on deep learning. In general, the approaches can be divided into those based on lexicon and those based on machine learning techniques.

Proceedings of the Multilingual and Multimodal Hate Speech Workshop 2021, pages 13–18,
Düsseldorf, Germany, September 6, 2021.

Several systems have used traditional machine learning models such as support vector machines (SVM) and logistic regression (LR) (Fortuna and Nunes, 2018; Basile et al., 2019; Zampieri et al., 2020). The work (MacAvaney et al., 2019) proposed an approach based on SVM that achieved state-of-the-art performance. Other systems were based on deep learning models such as convolutional neural network (CNN) and recurrent neural networks (RNN), including also attention mechanisms (Gröndahl et al., 2018; Magalhaes, 2019; Zhang et al., 2018). Moreover, some works proposed ensembles of neural networks, as (Zimmerman et al., 2018) which combined ten CNNs with different initialization weights in an ensemble to improve the detection of hate speech. In last years, the bidirectional encoder representations from Transformer (BERT) (Devlin et al., 2018) and other recent models have been used widely (Basile et al., 2019; Mandl et al., 2019; Zampieri et al., 2020; Mozafari et al., 2020; Samghabadi et al., 2020). In (Mozafari et al., 2019), authors used BERT in the case of hateful content within social media by transfer learning with fine-tuned methods. The results show that this strategy obtains a considerable performance in comparison to other existing approaches.

Multi-task Learning. Multi-task learning has been applied on a multitude of machine learning tasks (Zhang and Yang, 2021). (Liu et al., 2019) showed the success of applying multi-task learning for training a model to obtain outstanding performance on a number of natural language processing tasks. However, few works have employed multi-task learning strategies for hate speech detection. (Kapil and Ekbal, 2020) proposed a first idea of multi-task model based on CNN and RNN architectures to leverage useful information from multiple datasets related to hate speech. (Farha and Magdy, 2020) and (Rajamanickam et al., 2020) used a multi-task learning approach with polarity and emotion information to improve hate speech detection. The results showed that polarity and emotion detection can be beneficial for hate speech speech detection. Considering the relation between hate speech and offensive language, as well as the success of BERT and multi-task learning, our approach relies on a multi-task learning approach based on BERT.

3 The Multi-task Model

In order to design the model, we followed the definition of multi-task learning presented in (Zhang and Yang, 2021) as follows:

Given m learning tasks $\{T_i\}_{i=1}^m$, where all the tasks or a subset of them are related, multi-task learning aims to learn the m tasks together to improve the learning of a model for each task T_i by using the knowledge contained in all or some of other tasks.

Let $D_i = \{x_j^i, y_j^i\}_{j=1}^{N_i}$ be the dataset for the task T_i with N_i samples, where x_j^i is the jth sample and y_j^i is its label. Then, with the multi-task learning paradigm, a model learns from the sets $\{D_i\}_{i=1}^m$ at the same time. In our case, $m = 2$, since we have the tasks hate speech detection (T_1) and offensive language detection (T_2). Each $x_j^1 \in D_1, j = \overline{1, N_1}$ or $x_j^2 \in D_2, j = \overline{1, N_2}$ is a text, and each $y_j^1 \in D_1$ or $y_j^2 \in D_2$ is one of the values 0 or 1.

The first point to determine is the form through which knowledge is shared among the tasks. Usually, multi-task learning can be implemented by either soft or hard parameters sharing. In the first case, there is a model for each task with their parameters. Then, any technique such as regularization is used to encourage similarity between the parameters. On the other hand, hard parameters sharing is applied by sharing hidden layers between the tasks, while keeping several task-specific output layers (Caruana, 1993). This second strategy is widely used since it is seen to greatly reduce the risk of overfitting (Baxter, 1997). Bearing in mind our purpose of studying how the tasks can influence each other, we designed a model with shared parameters as a means to fit a system with datasets from both hate speech and offensive language. Hence, we focus on the hard parameter sharing.

3.1 Our Model

Our MTL model has one input and two outputs. Each output corresponds to one of the tasks: hate speech and offensive language detection. Thus, in this model we use the data from both tasks. All the texts are processed by the tokenizer of BERT and feed a shared encoder. This allows each task benefits from the others as they share features. Then, a task-specific output head is attached on the top of the output of the encoder for each task. The shared encoder allows updates to occur on the same weights in the training. Therefore, the layers are fine-tuned according to the two downstream tasks.

3.2 Shared Encoder

We base on the BERT architecture to build the shared encoder, such that we tokenize the texts in the same way as in BERT. That is, we represent each text t into a sequence of N+1 tokens $\{Tok0, Tok1, ..., TokN\}$, with Tok0 = [CLS], the special token in BERT for classification. This token sequence is then used as input to a BERT model to extract a sequence of textual hidden states h_t of size $(N+1) \times dim$, where dim is the BERT hidden size. Thus, we obtain a vector sequence from the encoder as follows:

$$\begin{aligned} h_t &= BERT(\{Tok0, Tok1, ..., TokN\}) \\ &= \{h_t^0, h_t^1, ..., h_t^N\} \end{aligned} \quad (1)$$

In order to save computation we only take the output of the first token (h_t^0 from CLS) as the representation of the whole sequence.

3.3 Task Outputs

A task-specific classifier is applied over the output of the shared encoder for each task. In both hate speech

and offensive language detection, we use a fully connected layer with the softmax function (σ) to obtain a classification output. Then, we used cross entropy to calculate the loss in each output T as Equation 2. Where y_i^T is the true class of a text t from the task T, and $\hat{y_i}^T$ its predicted value, calculated as Equation 3, W and b are the classifier's parameters.

$$\mathcal{L}^T(y_t^T, \hat{y_t}^T) = -\Sigma_t \; y_t^T * log(\hat{y_t}^T) \tag{2}$$

$$\hat{y_t}^T = \sigma(y_i^T W + b) \tag{3}$$

3.4 Losses and Backpropagation

In order to iterate over the tasks, we separate them to generate batches and calculate the loss for only the specific task head of the model. We use the sum of the two dataset to determine the number of batches to generate in each epoch. We do backpropagation once, i.e. we run the batch of one dataset through the model and calculate the loss for the corresponding task, ignoring loss calculations for the other task. Then, we run the batch for the second dataset through the model and calculate the loss for that second task. Once those two batches are processed we sum their losses and perform backpropagation. Therefore, the objective function can be formulated as follows:

$$\begin{aligned} & min \sum_{T \in \{HS, OL\}} \mathcal{L}^T(y_t^T, \hat{y_t}^T) \\ = & min \sum_{T \in \{HS, OL\}} -\Sigma_t \; y_t^T * log(\hat{y_t}^T) \end{aligned} \tag{4}$$

4 The Single-task Model

In contrast to MTL, single-task learning (STL) is the most widely used paradigm. In this case, the objective function only takes into account one task and therefore only involves one dataset. Hence, the update of the weights of a model uses the input of a single task and its corresponding output.

In our study, we use a transfer learning technique by considering offensive language detection as the source task and hate speech detection as the target task. In this sense, we base our study in the STL model. It is important to point out that we use a zero few shot learning strategy, where no labeled examples from the target task is used in the training (Kadam and Vaidya, 2018). In that way, we can analyze how hate speech detection can be affected when a dataset for the other phenomenon is used.

5 Experiments

We used four datasets in English: HatEval (Basile et al., 2019), Founta (Founta et al., 2018) and Waseem & Hovy (Waseem and Hovy, 2016) for hate speech detection (HS), and OLID (Zampieri et al., 2019) and SOLID (Rosenthal et al., 2020) for offensive language detection (OL). We only used 10k texts in English from the training set of each dataset. We kept the ratio between the number of samples of each class in the original datasets. In general , class 0 refers to the absence of HS or OL, whereas class 1 refers to the opposite.

In our implementation, we used the pre-trained BERT-base uncased model from the Huggingface's Transformers library (Wolf et al., 2019), which has 12 transformer layers, a hidden state size of 768, and 12 self-attention heads. We tune the parameters by minimizing the standard cross-entropy loss either for hate speech detection or offensive language detection. For each model we used the following hyperparameter: learning rate to 2e-5 and batch size to 16 in 10 epochs. In order to optimize we used the Adam algorithm with weight decay of 1e-3 for all parameters except for bias and weights in LayerNorm, and set the epsilon to 1e-8.

5.1 Results and Analysis

We handle the evaluation with the stratified 3-fold cross-validation technique and report our results in terms of F1-macro scores.

We compare the performance of the STL model with the MTL model for each of the datasets HatEval, Founta and W&H. In each case, we trained the MTL model twice. First, we used OLID and one of the dataset for hate speech detection. Then, we used SOLID in a similar way. Therefore, we obtained 6 results with the MTL model.

Table 1 illustrates the performance for HatEval. Different from what we supposed, the result with the STL model is higher than those obtained with the MTL model for both OLID and SOLID datasets. Notice that the difference between MTL and STL is very small when using OLID. However, the purpose of the MTL paradigm is to improve the tasks performance. Thus, this strategy does not seem to be appropriate to improve hate speech detection with datasets built for offensive language, although successful studies that show improvement when sentiment analysis is considered.

Model	**F1-macro**
STL	0.8066
$\text{MTL}_{HatEval-OLID}$	0.7900
$\text{MTL}_{HatEval-SOLID}$	0.7644

Table 1: Cross-validation results for HatEval.

Similarly, Tables 2 and 3 show the results for Founta and W&H respectively. For these two datasets, hate speech is even more affected when considering datasets with offensive language, being Founta the worst case. It could be explained since Founta consists of texts labeled as hateful, abusive, spam and normal, and we consider the last three ones as non-hateful. Thus, the label 0 for abusive texts can conflict with the label 1 for some similar offensive texts in OLID and SOLID. This highlights the importance of focusing on data rather than models.

Model	**F1-macro**
STL	0.6309
$\text{MTL}_{Founta-OLID}$	0.2308
$\text{MTL}_{Founta-SOLID}$	0.2290

Table 2: Cross-validation results for Founta.

Model	**F1-macro**
STL	0.8641
$\text{MTL}_{W\&H-OLID}$	0.8002
$\text{MTL}_{W\&H-SOLID}$	0.7975

Table 3: Cross-validation results for W&H.

Furthermore, these results point out that although hate speech and offensive language are related phenomena, the indistinct use of datasets can affect the performance in hate speech detection. Actually, it makes sense since many cases of offensive texts are not hateful. When we consider these texts to train a model for hate speech detection, many offensive texts can be misclassified as hateful when they are not really, and vice versa. We explore the performance of transfer learning later, in order to deeper analyze it.

5.2 Transfer Learning

Table 4 shows the results obtained with transfer learning when one dataset for offensive language was used in the training. The results worsen for all the datasets when it is used a model fine-tuned with a different dataset, not only in comparison with the STL model, but also considering the results obtained with MTL in general. Although this performance was to be excepted, it is worth noting that Founta dataset obtained better results with this strategy of transfer learning than with MTL. Once again, its type of labeling seems to affect less in this strategy. On overall, this is an alternative strategy to confirm the importance of considering the characteristics of the datasets for training system for hate speech detection.

Model	**HatEval**	**Founta**	**W&H**
STL	0.8066	0.6309	0.8641
STL_{OLID}	0.6330	0.5875	0.5612
STL_{SLID}	0.6276	0.5885	0.5858

Table 4: F1 when transfer learning from Offensive Language to Hate Speech Detection. The first row corresponds the F1 obtained when the same dataset was used for training.

6 Error Analysis

In this section we conduct an error analysis to provide a deeper analysis of the MTL model performance. In this sense, we analyze the confusion matrices of each models. They are illustrated in the following tables:

	STL		MTL OLID		MTL SOLID	
	HS	Not	HS	Not	HS	Not
HS	1092	312	1146	258	1216	188
Not	317	1613	408	1522	645	1285

Table 5: Confusion matrix for HatEval.

	STL		MTL OLID		MTL SOLID	
	HS	Not	HS	Not	HS	Not
HS	19	65	24	60	20	64
Not	30	1938	213	1755	237	1731

Table 6: Confusion matrix for Founta.

	STL		MTL OLID		MTL SOLID	
	HS	Not	HS	Not	HS	Not
HS	707	170	978	101	707	170
Not	183	2273	219	2237	427	2029

Table 7: Confusion matrix for W&H.

Table 5 shows the confusion matrices for HatEval for the STL model as well as for the MTL model when both OLID and SOLID datasets were used.

Roughly, the MTL model tends to misclassify less hateful text as non-hateful in both cases compared to the number of false negatives obtained with STL. It is a good result since to overlook hateful texts is one of the principal problems in hate speech detection.

Anyway, the number of false positives increased with the MTL paradigm. This makes sense, since the objective function was designed to simultaneously minimize the error for both hate speech and offensive language detection. Therefore, offensive texts that are not hateful can be misclassified as hateful. This points out again the important finding of our study. That is, considering datasets of related phenomena for hate speech detection, can cause non-hateful texts (with other type of harmful content) to be misclassified as hateful.

Therefore, the knowledge provided by the external datasets of related tasks, although can slightly decrease the false positives, can worsen the prediction in general.

Similar results are observed on Founta and W&H in Tables 6 and 7 respectively.

7 Conclusions

Hate speech detection is a prominent task in Natural Language Processing and other disciplines. A lot of research have emerged in the last few years to mitigate the problem that hurtful messages can cause on social media. Therefore, a number of strategies have been proposed considering available datasets. One of the problems consists of the lack of clear boundaries among hate speech and other related phenomena, such that datasets from different phenomena are often used interchangeably. In this paper, we studied how offensive language can influence hate speech detection. We conducted an analysis

with the multi-task learning paradigm that allows us to simultaneously train a model for both tasks. In this sense, we proposed a model which contains a shared encoder based on BERT, and considers the losses of both tasks in the backpropagation for the fine-tuning. Moreover, we considered a transfer learning technique as an alternative to study how the use of datasets for offensive language can influence hate speech detection. Experimental results, conducted on five datasets, showed that although false negatives can be improved, the detection of hateful texts can be affected in general.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019. Semeval-2019 Task 5: Multilingual Detection of Hate Speech against Immigrants and Women in Twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Jonathan Baxter. 1997. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39.

Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2020. DeepHate: Hate Speech Detection Via Multi-Faceted Text Representations. In *12th ACM Conference on Web Science*, pages 11–20.

Richard Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Morgan Kaufmann.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *arXiv preprint arXiv:1703.04009*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ibrahim Abu Farha and Walid Magdy. 2020. Multitask learning for arabic offensive language and hate-speech detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 86–90.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All You Need is" Love" Evading Hate Speech Detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, pages 2–12.

Suvarna Kadam and Vinay Vaidya. 2018. Review and analysis of zero, one and few shot learning approaches. In *International Conference on Intelligent Systems Design and Applications*, pages 100–112. Springer.

Prashant Kapil and Asif Ekbal. 2020. A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, 210:106458.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate Speech Detection: Challenges and Solutions. *PloS one*, 14(8):e0221152.

Ashe Magalhaes. 2019. Automating Online Hate Speech Detection: A Survey of Deep Learning Approaches. Master's thesis, School of Informatics, University of Edinburgh.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC Track at Fire 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, pages 14–17.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2019. A BERT-based Transfer Learning Approach for Hate Speech Detection in Online Social Media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate Speech Detection and Racial Bias Mitigation in Social Media Based on BERT Model. *PloS one*, 15(8):e0237861.

Usman Naseem, Imran Razzak, and Peter W Eklund. 2020. A Survey of Pre-processing Techniques to Improve Short-Text Quality: A Case Study on Hate Speech Detection on Twitter. *Multimedia Tools and Applications*, pages 1–28.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and Multi-aspect Hate Speech Analysis. pages 4675–4684.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review. *Language Resources and Evaluation*, pages 1–47.

Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Joint modelling of emotion and abusive language detection. *arXiv preprint arXiv:2005.14028*.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A Large-scale Semi-supervised Dataset for Offensive Language Identification. *arXiv preprint arXiv:2004.14454*.

Niloofar Safi Samghabadi, Parth Patwa, PYKL Srinivas, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and Misogyny Detection Using BERT: A Multi-task Approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131.

Anna Schmidt and Michael Wiegand. 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*, pages 1–10.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. *arXiv preprint arXiv:1902.09666*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). *arXiv preprint arXiv:2006.07235*.

Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting Hate Speech on Twitter using A Convolution-GRU based Deep Neural Network. In *European semantic web conference*, pages 745–760. Springer.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving Hate Speech Detection with Deep Learning Ensembles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.