11-28-2023

OffensEval 2023: Offensive language identification in the age of Large Language Models

Marcos Zampieri
George Mason University

Sara Rosenthal
IBM Research

Preslav Nakov
Mohamed Bin Zayed University of Artificial Intelligence

Alphaeus Dmonte
George Mason University

Recommended Citation

M. Zampieri et al., "OffensEval 2023: Offensive language identification in the age of Large Language Models," *Infection Control and Hospital Epidemiology*, vol. 44, no. 11, pp. 1737 - 1747, Nov 2023.
The definitive version is available at https://doi.org/10.1017/ice.2023.69

Natural Language Engineering (2023), **29**, pp. 1416–1435
doi:10.1017/S1351324923000517

CAMBRIDGE UNIVERSITY PRESS

SURVEY PAPER

OffensEval 2023: Offensive language identification in the age of Large Language Models

Marcos Zampieri[1], Sara Rosenthal[2], Preslav Nakov[3], Alphaeus Dmonte[1] and Tharindu Ranasinghe[4]

[1]George Mason University, Fairfax, VA, USA, [2]IBM Research, Yorktown Heights, NY, USA, [3]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE, and [4]Aston University, Birmingham, UK
Corresponding author: Marcos Zampieri; Email: mzampier@gmu.edu

(Received 6 November 2023; accepted 6 November 2023)

Abstract

The OffensEval shared tasks organized as part of SemEval-2019–2020 were very popular, attracting over 1300 participating teams. The two editions of the shared task helped advance the state of the art in offensive language identification by providing the community with benchmark datasets in Arabic, Danish, English, Greek, and Turkish. The datasets were annotated using the OLID hierarchical taxonomy, which since then has become the *de facto* standard in general offensive language identification research and was widely used beyond OffensEval. We present a survey of OffensEval and related competitions, and we discuss the main lessons learned. We further evaluate the performance of Large Language Models (LLMs), which have recently revolutionalized the field of Natural Language Processing. We use zero-shot prompting with six popular LLMs and zero-shot learning with two task-specific fine-tuned BERT models, and we compare the results against those of the top-performing teams at the OffensEval competitions. Our results show that while some LMMs such as Flan-T5 achieve competitive performance, in general LLMs lag behind the best OffensEval systems.

Keywords: Machine learning; Text classification

1. Introduction

The development of computational models and datasets to detect various forms of offensive content online has become a very popular research topic in recent years (Fortuna and Nunes 2018; Poletto *et al.* 2021). Research on this topic was motivated by the pressing need to create safer environments in social media platforms through strategies such as automatic content moderation (Weerasooriya *et al.* 2023). With the goal of aiding content moderation, systems are trained to recognize a variety of related phenomena such as aggression, cyberbulling, hate speech, and toxicity (Arora *et al.* 2023).

A lot of research on this topic is driven by shared task competitions that provide important benchmark datasets, results, and systems to the research community. Notable examples include HatEval, OffensEval, and TSD. Organized as part of the International Workshop on Semantic Evaluation (SemEval), each of these competitions attracted hundreds of participating teams from all over the world. OffensEval is arguably the most popular shared task on this topic. Its 2019 edition focused on English and attracted 800 teams, while the 2020 was a multilingual competition with datasets in five languages and it attracted over 500 teams. The best-performing teams in these competitions developed systems using transformer-based architectures such as BERT

(Devlin *et al.* 2019) and ELMo (Peters *et al.* 2018), which were the state-of-the-art pre-trained language models at the time.

Since the last edition of the OffensEval shared task in 2020, the field of Natural Language Processing (NLP) has undergone a revolution with the introduction of a new generation of LLMs such as GPT (Radford *et al.* 2019), OPT (Zhang *et al.* 2022), LLaMA (Touvron *et al.* 2023b), LLaMA 2 (Touvron *et al.* 2023a), PaLM (Chowdhery *et al.* 2022), BLOOM (Scao *et al.* 2023), FLAN-T5 (Chung *et al.* 2022), etc. Such models have reached the general public with commercial tools such as ChatGPT, sparking renewed widespread interest in AI and NLP. Within the research community, LLMs have shown state-of-the-art performance for a variety of tasks, and have revolutionized the research in the field. LLMs have also given rise to the art of prompt engineering, which includes a variety of prompting techniques such as zero-shot, few-shot, chain-of-thought, etc. (Liu *et al.* 2023).

In light of these recent developments, we present OffensEval 2023, an evaluation of OffensEval in the age of LLMs. We present (i) a survey of the two editions of OffensEval and related benchmark competitions and (ii) an evaluation of LLMs and fine-tuned models.

Our contributions can be summarized as follows:

1. A survey of offensive language identification benchmark competitions with a special focus on OffensEval. We discuss benchmark competitions that addressed various languages (e.g., Arabic, Danish, German) and phenomena (e.g., hate speech, toxicity).
2. An evaluation of state-of-the-art LLMs on the OffensEval 2019 and OffensEval 2020 datasets. We experiment with six LLMs and two fine-tuned BERT models on the OffensEval datasets, and we compare their performance to the best systems in the competition.

The remainder of this paper is organized as follows: Section 2 discusses popular related shared tasks such as HatEval, TRAC, and HASOC. Section 3 describes the two editions OffensEval, datasets, and previous experiments in detail. Section 4 discusses our experiments benchmarking six LLMs and two task fine-tuned BERT models on offensive language identification. Finally, Section 5 concludes this paper and presents directions for future work.

2. Related benchmark competitions

In this section, we survey some recent popular benchmark competitions on the topic. The competitions presented next have addressed different types of offensive content such as hate speech in HatEval (Basile *et al.* 2019), aggression in TRAC (Kumar *et al.* 2018), and misogyny in MAMI (Fersini *et al.* 2022). While most tasks focused exclusively on offensive content, some tasks have attempted to bridge the gap between offensive content identification and other phenomena. One such example is HaHaCkaton (Meaney *et al.* 2021) which provided participants with the opportunity to develop systems to model offense and humor jointly.

2.1 HatEval (SemEval-2019 task 5): multilingual detection of hate speech against immigrants and women in Twitter

HatEval (Basile *et al.* 2019) was organized as part of the 2019 edition of SemEval (the International Workshop on Semantic Evaluation). Its focus was on detecting hate speech against women and migrants, in English and Spanish. The task organizers provided an annotated dataset collected from Twitter containing 19,600 tweets: 13,000 for English and 6600 for Spanish. The dataset was annotated with respect to (1) hatefulness, (2) target, and (3) aggression. The competition received over 100 runs from 74 different teams. Half of the teams submitted systems relying on traditional machine learning approaches, while the other half submitted deep learning systems. The best systems used traditional classifiers such as SVMs (Indurthi *et al.* 2019).

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

2.2 TRAC: evaluating aggression identification in social media

The TRAC shared task (Kumar *et al.* 2018) has been held as a biennial event since 2018, as part of the Workshop on Trolling, Aggression, and Cyberbullying. It focuses on aggression identification and has covered several languages. In the first iteration, TRAC had one sub-task on aggression identification, and the participants were asked to classify instances as overtly aggressive, covertly aggressive, and non-aggressive. The task organizers released a dataset of 15,000 aggression-annotated Facebook posts and comments in Hindi (in both Roman and Devanagari script) and English. A total of 130 teams registered to participate in the task, and 30 teams submitted runs. The best system used an LSTM and machine translation for data augmentation (Aroyehun and Gelbukh 2018).

The 2020 edition of the TRAC shared task (Kumar *et al.* 2020) had two sub-tasks: aggression identification (sub-task A), where the goal was to discriminate between posts labeled as overtly aggressive, covertly aggressive, and non-aggressive, and gendered aggression identification (sub-task B), which asked participants to discriminate between gendered and non-gendered posts. The shared task was organized for three languages: Bengali, Hindi, and English. The participants were provided with a dataset of approximately 5000 instances from YouTube comments in each of the languages. Approximately 1000 instances were provided per language for each sub-task for testing. The competition attracted a total of 70 teams. The best-performing system used multiple fine-tuned BERT models and bootstrap aggregation (Risch and Krestel 2020).

The 2022 edition of the TRAC shared task contained two different tasks from the previous iterations. In sub-task A, the primary focus remained on identifying aggression, encompassing aggression, gender bias, racial bias, religious intolerance, and casteist bias within social media content. For sub-task B, the participants were presented with a comment thread containing information about the existence of various biases and threats (such as gender bias, gendered threats, or their absence) and their discourse connections to preceding comments and the original post, categorized as attack, abetment, defense, counter-speech, or gaslighting. The participants were asked to predict the presence of aggression and bias within each comment, potentially leveraging the available contextual cues. The organizers released a dataset of 60k comments in Meitei, Bangla, and Hindi for training and testing (a total of 180k examples) from YouTube. The best system at the competition, for both tasks, used logistic regression (Kumari, Srivastav, and Suman 2022). For sub-task B, only a test set was provided containing COVID-19-related conversations, annotated with levels of aggression, offensiveness, and hate speech. The participants were asked to train their machine learning models using training data from previous TRAC editions. The primary goal of this task was to assess the adaptability and the generalizability of aggression identification systems when faced with unforeseen and unconventional scenarios. Once again, the best model used logistic regression (Kumari *et al.* 2022).

2.3 HASOC: hate speech and offensive content identification in English and Indo-Aryan languages

Since 2019, the HASOC shared task (Mandl *et al.* 2019) has been a regular task at FIRE (the Forum for Information Retrieval Evaluation). Its primary objective is the detection of hate speech and offensive content in English and Indo-Aryan languages. In its inaugural edition (Mandl *et al.* 2019), the shared task featured two sub-tasks across three languages: English, Hindi, and German. sub-task A was a binary classification task, where the goal was to categorize content as offensive or not offensive. In sub-task B, the focus shifted to further fine-grained classification of offensive content into three categories: hate speech, offensive content, and profanity. For each language, there were 5000 training and 1000 testing examples from Twitter and Facebook. Notably, the most successful systems leveraged neural network architectures, incorporating Long Short-Term Memory (LSTM) networks and word embeddings (Wang *et al.* 2019). Several of the top-performing teams also used BERT, even though it was still emerging in NLP (Ranasinghe, Zampieri, and Hettiarachchi 2019).

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

The 2020 edition of the HASOC shared task (Mandl *et al.* 2020) featured the same two sub-tasks and the same three languages as in the previous year. The organizers provided a new annotated dataset collected from Twitter, which contained 3708 English, 2373 German, and 2963 Hindi examples. The best-performing teams in that edition of the HASOC shared task used different variants of BERT (Raj, Srivastava, and Saumya 2020; Mishra, Saumya, and Kumar 2020).

The 2021 edition of the HASOC challenge had two tasks (Modha *et al.* 2021). Task 1 contained the same two sub-tasks from the previous 2 years. However, there was a difference in the languages: German was replaced by Marathi as a new language (Mandl *et al.* 2021). The organizers provided newly annotated 3843 English instances, 4594 Hindi instances, and 1874 Marathi instances from Twitter for training. The best-performing systems again used different variants of the BERT architecture and combined it with cross-lingual transfer learning (Banerjee *et al.* 2021; Bhatia *et al.* 2021; Nene *et al.* 2021). The second task in 2021 focused on the identification of conversational hate speech in code-switched text, where the same message mixes different languages (Modha *et al.* 2021). The objective of this task was to identify posts that are benign when considered in isolation, but might be judged as hate, profane, and offensive if the particular context is taken into account. The organizers provided 7000 code-switched posts in English and Hindi from Twitter. The winning team used an ensemble based on IndicBERT (Doddapaneni *et al.* 2023), Multilingual BERT (Devlin *et al.* 2019), and XLM-RoBERTa (Conneau *et al.* 2019). In each of their model, they concatenated the conversation into the input tweet.

HASOC 2022 featured three tasks (Satapara *et al.* 2022). Task 1 was a continuation of the 2021 task 2, where the goal was to detect hate and offensive content in conversations (Modha *et al.* 2022) where the classes were hate offensive and non-hate offensive. Task 2 was a fine-grained classification of task 1, where the participants were asked to classify the hate offensive conversations from task 1 into three classes; standalone hate, contextual hate, and non-hate (Modha *et al.* 2022). The participants were provided with 5200 code-mixed conversations in English, Hindi, and German, with annotations for both tasks from Twitter. The best-performing system used Google MuRIL (Khanuja *et al.* 2021) and took the context into account (Singh and Garain 2022). Task 3 was a continuation of 2021 task 1 (Ranasinghe *et al.* 2022b). However, it only featured Marathi and had three sub-tasks, which followed the popular OLID taxonomy. sub-task 1 asked to detect offensive language, sub-task 2 focused on the categorization of offensive language into targeted or untargeted, and finally, the sub-task 3 looked to identify the target of the offense classifying the instances into individual target, group target and other. The participants were given 3500 annotated instances from Twitter. The best system again was based on XLM-RoBERTa (Dikshitha Vani and Bharathi 2022).

2.4 TSD: toxic span detection

Toxic span detection was organized at the 2021 edition of SemEval (Pavlopoulos *et al.* 2021). The shared task asked to detect the text spans that contain toxic or offensive content. The participants were provided with a reannotated version of the Civil Comments dataset, with 7939 training and 2000 test instances, with toxic span annotations. TSD was the first of its kind in predicting toxicity at the span level. The shared task received 1385 valid submissions from 91 teams, with the best team modeling the problem as token labeling and span extraction. They used two systems based on BERT with a conditional random field layer at the top (Zhu *et al.* 2021).

2.5 MAMI: multimedia automatic misogyny identification

The Multimedia Automatic Misogyny Identification (MAMI) was task 5 at SemEval 2022 (Fersini *et al.* 2022). The task had two sub-tasks: sub-task A was a binary classification task, asking to distinguish between misogynous and non-misogynous memes, and sub-task B was a multi-label classification task to detect the type of misogyny: stereotype, shaming, objectification, and violence. The organizers provided the participants with balanced training and testing datasets with

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

10,000 and 5000 memes, respectively. The best-performing teams used RoBERTa and VisualBERT; many teams used ensembles combining several models (Zhang and Wang 2022).

2.6 HaHaCkaton: detecting and rating humor and offense

The HaHaCkathon (Meaney *et al.* 2021) combined humor detection and offense language identification into a single task opening the possibly of jointly modeling two tasks that were previously addressed separately. The organizers 10,000 examples from Twitter and the Kaggle Short Jokes dataset, each annotated by 20 annotators for humor and offense. HaHaCkathon featured three sub-tasks: (1) humor detection, (2) prediction of humor and offense ratings, and (3) controversy detection (i.e., predicting whether the variance in the human humor ratings for a given example is higher than a specific threshold). The individual sub-tasks attracted between 36 and 58 submissions. In terms of approaches and performance, most teams used pre-trained language models such as BERT, ERNIE 2.0, and ALBERT and most of the best-performing teams used additional techniques, for example adversarial training.

2.7 EDOS (SemEval-2023 task 10): explainable detection of online sexism

EDOS (Kirk *et al.* 2023) was organized as part of SemEval-2023 with the goal of detecting sexist online posts and explaining them. The task had three sub-tasks: sub-task A was a binary classification task where the participants needed to distinguish between sexist and non-sexist content. Sub-task B was a fine-grained classification task that disaggregates sexist content into four conceptually and analytically distinct categories: (i) threats, plans to harm & incitement, (ii) derogation, (iii) animosity, and (iv) prejudiced discussion. Finally, sub-task C disaggregates each category of sexism into eleven fine-grained sexism vectors such as threats of harm, descriptive attacks, and systemic discrimination against women as a group or as an individual (Kirk *et al.* 2023). The participants of the EDOS shared task were provided with a dataset of 20,000 annotated social media comments from Reddit and Gab. Additionally, the organizers provided one million unannotated social media comments. A total of 128 teams participated in the competition. The top team uses transformer-based architectures and further improved their results using continued pre-training on the unannotated dataset and multitask learning (Zhou 2023).

2.8 DeTox: toxic comment classification at GermEval

The 2021 edition of GermEval (a series of shared task evaluation campaigns that focus on NLP for German) included a shared task that focused on identifying toxic, engaging, and fact-claiming comments in German (Risch *et al.* 2021). The task included three sub-tasks: sub-task 1 was a binary classification problem (toxic vs. non-toxic), sub-task 2 was also a binary classification problem, asking to distinguish between engaging and non-engaging comments. sub-task 3 was also a binary classification problem, asking to distinguish between fact-claiming and non-fact-claiming comments. The participants were provided with 3244 and 1092 manually annotated training and testing instances, extracted from Facebook comments. The competition received a submission from 32 teams across the three sub-tasks. The best-performing teams used traditional classifiers combined with some form of pre-trained deep learning models such as BERT and XLM-RoBERTa (Bornheim, Grieger, and Bialonski 2021; Morgan, Ranasinghe, and Zampieri 2021).

2.9 DETOXIS: detection of toxicity in comments in Spanish at IberLeF

IberLeF 2021 was a workshop organized to evaluate systems in Spanish and other Iberian languages, on various NLP tasks. The competition included a general shared task to detect harmful content, with specific tasks related to offensive language detection in Spanish and Mexican Spanish and toxicity detection in Spanish comments (Taulé *et al.* 2021). The offensive language

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

detection task was further divided in four sub-tasks: sub-task 1 was a multiclass classification problem for generic Spanish where the participants have to classify comments into five different categories; "Offensive and target is a person," "Offensive and target is a group of people," "Offensive and target is different from a person or a group," "Non-offensive, but with expletive language" and "Non-offensive." Sub-task 2 was also a multiclass classification problem with the previous categories, however, meta-data for the post was given, such as author genre. Sub-task 3 was a binary classification problem for Mexican Spanish where the participants must classify tweets as offensive or non-offensive. Sub-task 4 was a binary classification problem with the same tweets as sub-task 3, but the participants were provided with the meta-data for each tweet, such as date, retweet count, and author followers count. Sub-tasks 1 and 2 had a combined 16,710 training, 100 development, and 13,606 testing instances, while sub-tasks 3 and 4 had a combined 5060 training and 2183 testing instances. All of the instances were based on Twitter. Most of the top systems used some pre-trained transformer-based models such as multilingual BERT, BETO and XLM-RoBERTa (Plaza-del Arco, Molina-González, and Alfonso 2021).

The toxicity detection task focused on detecting toxicity in Spanish news comments. The task was further divided into two sub-tasks: sub-task 1 was a binary classification task to distinguish between toxic and non-toxic comments, while sub-task 2 was about assigning a toxicity score for the comment, ranging from 0 (not toxic) to 3 (very toxic). The participants were provided with 3463 comments for training and 896 comments for testing their models. All of the instances were based on news media comments. The best-performing teams for both sub-tasks used BETO (the Spanish version of BERT model) (Plaza-del Arco *et al.* 2021).

3. OffensEval

The evaluation presented in this paper focuses on the shared task on Identifying and Categorizing Offensive Language in Social Media (OffensEval). The task has been organized at SemEval-2019 including English data and at SemEval-2020 including data in English and other four languages, namely Arabic, Danish, Greek, and Turkish. The task has been influential as it was the first to model offensive language identification considering the type and target of offensive posts. OffensEval was based on the three levels of the Offensive Language Identification Dataset (OLID) taxonomy (Zampieri *et al.* 2019a) which has since become a *de facto* standard for general offensive language taxonomy. OLID's hierarchical annotation model was developed with the goal of serving as a general-purpose model for multiple sub-tasks (e.g., hate speech, cyberbulling, etc.) as described next.

3.1 The OLID taxonomy

Introduced in (Zampieri *et al.* 2019a), the OLID taxonomy is a labeling schema that classifies each example for offensiveness using the following three-level hierarchy.

- **A**: Offensive Language Detection
- **B**: Categorization of Offensive Language
- **C**: Offensive Language Target Identification

The original OLID dataset was created for English and the taxonomy has been widely adopted for several languages (Pitenis, Zampieri, and Ranasinghe 2020; Gaikwad *et al.* 2021; Ranasinghe *et al.* 2022a). The popularity of OLID is due to the flexibility provided by its hierarchical annotation model that considers multiple types of offensive content in a single taxonomy. For example, targeted insults to a group are often hate speech whereas targeted insults to an individual are often cyberbulling. The hierarchical structure of OLID allows mapping OLID level A (offensive

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

Table 1. Several tweets from the original OLID dataset, with their labels for each level of the annotation model (Zampieri *et al.* 2019a)

Tweet	A	B	C
@USER He is so generous with his offers	NOT	—	—
IM FREEEEE!!!! WORST EXPERIENCE OF MY FUCKING LIFE	OFF	UNT	—
@USER Fuk this fat cock sucker	OFF	TIN	IND
@USER Figures! What is wrong with these idiots? Thank God for @USER	OFF	TIN	GRP

vs. non-offensive) to labels in various other related datasets annotated with respect to hate speech, aggression, etc. as demonstrated in (Ranasinghe and Zampieri 2020, 2021).

We present some examples retrieved from the original OLID dataset with their respective labels in Table 1. Further details about each level of the taxonomy are described next.

3.1.1 Level A: offensive language detection

In this level, annotators are asked to annotate each instance with respect to the presence of any form of offensive content by answering the question "*Is the text offensive?*" The following two labels are included in level A:

- **OFF** Inappropriate language, insults, or threats.
- **NOT** Neither offensive nor profane. The following example is not offensive: *@USER you are also the king of taste*

3.1.2 Level B: categorization of offensive language

In this level, only offensive instances labeled in level A as *OFF* are included. Annotators are asked to label each offensive instance as either targeted or untargeted by answering the question *"Is the offensive text targeted?"* The following two labels are included in level B:

- **TIN** Targeted insult or threat towards a group or an individual.
- **UNT** Untargeted profanity or swearing. The following example includes profanity (*bullshit*) that is not targeted to anyone: *@USER What insanely ridiculous bullshit*.

3.1.3 Level C: offensive language target identification

In this level, only targeted offensive instances labeled in level A as *OFF* and in level B as *TIN* are included. Annotators are asked to label each targeted offensive instance with respect to its target by answering the question *"What is the target of the offensive?"* The following three labels are included in level C:

- **IND** The target is an individual explicitly or implicitly mentioned in the conversation. The following example is targeted towards an individual, *that*: *@USER Anyone care what that dirtbag says?*
- **GRP** Hate speech targeting a group of people based on ethnicity, gender, sexual orientation, religion, or other common characteristic. The following example is targeted towards a group *liberals*: *Poor sad liberals. No hope for them.*
- **OTH** Targets that does not fall into the previous categories, for example organizations, events, and issues. The following example is targeted towards an organization, *NFL*: *LMAO. . ..YOU SUCK NFL*

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

Table 2. Distribution of label combinations in OLID (Zampieri *et al.* 2019b)

A	B	C	Train	Test	Total
OFF	TIN	IND	2,407	100	2,507
OFF	TIN	OTH	395	35	430
OFF	TIN	GRP	1,074	78	1,152
OFF	UNT	—	524	27	551
NOT	—	—	8,840	620	9,460
All			13,240	860	14,100

3.2 OffensEval 2019

OffensEval 2019 (Zampieri *et al.* 2019b) at SemEval received a very positive response from the community. The shared task attracted about 800 participating teams making it the largest ever SemEval task until that point. OLID, the official dataset for this task, featured 14,100 instances retrieved from Twitter divided into training and testing sets. We present a breakdown of the instances in OLID and their label distribution in Table 2.

Two factors have contributed to OffensEval's popularity (i) the growing popularity of deep learning models and the introduction of large general pre-trained transfer models, most notably BERT (Devlin *et al.* 2019), just months before the competition and (ii) the use of the OLID taxonomy (Zampieri *et al.* 2019a). Prior to OLID, previous work on detecting offensive language focused on detecting specific types of offensive content such as hate speech, profanity, and cyberbullying. As described earlier in this section, the OLID taxonomy approached offensive content using a single annotation scheme allowing multiple types of offensive content to be modeled in a single taxonomy. This, in our opinion, helped attracting participants interested in different offensive and abusive language phenomena.

OffensEval 2019 featured three sub-tasks each representing one level of the OLID taxonomy. The organizers provided three baselines to the participants, namely a CNN model, a BiLSTM model, and an SVM model described in (Zampieri *et al.* 2019a). The best baseline was a CNN model that achieved 0.800 F1 score for sub-task A, 0.690 for sub-task B, and 0.470. The CNN baseline model would be ranked 10th among all entries in sub-tasks A and B but only 48th in sub-task C. We present the top-10 results of each sub-task along with the strongest baseline in Table 3.

Sub-task A, where participants were asked to label each instance as either offensive or not offensive, was the most popular sub-task with 104 submissions. The best performance in this sub-task was obtained by (Liu, Li, and Zou 2019) who used a BERT model achieving 0.829 F1 score. Sub-task B, where participants were asked to label each instance as either targeted or untargeted, received 76 submissions. The best system in sub-task B by (Nikolov and Radivchev 2019) also used a BERT model achieving 0.755 F1 score. Finally, sub-task C, where participants trained models to identify one of the three target labels (IND, GRP, OTH), received 65 submissions. The best system in sub-task C by (Han, Liu, and Wu 2019) used a deep learning approach based on bidirectional recurrent layers with gated recurrent units achieved 0.660 F1 score.

To illustrate the variety of approaches used in OffensEval 2019, we present a breakdown of all approaches used for sub-task A in Figure 1.

BERT had been recently introduced and it was among the first models to employ a transformer-based architecture and pre-trained contextual embeddings. At the time of OffensEval 2019, BERT had quickly become very popular in NLP due to its high performance in many tasks and the possibility of being used as an off-the-shelf the model. Despite its growing popularity at that time, as depicted in Figure 1, we can see that only 8% of the teams (about 12 teams) approached sub-task A

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

Table 3. F1-Macro for the top-10 teams for all three sub-tasks. The best baseline model (**CNN**) is also presented

Sub-task A		Sub-task B		Sub-task C	
Team ranks	F1 range	Team ranks	F1 range	Team ranks	F1 range
1	0.829	1	0.755	1	0.660
2	0.815	2	0.739	2	0.628
3	0.814	3	0.719	3	0.626
4	0.808	4	0.716	4	0.621
5	0.807	5	0.708	5	0.613
6	0.806	6	0.706	6	0.613
7	0.804	7	0.700	7	0.591
8	0.803	8	0.695	8	0.588
9	0.802	9	0.692	9	0.587
CNN	**0.800**	**CNN**	**0.690**	10	0.586

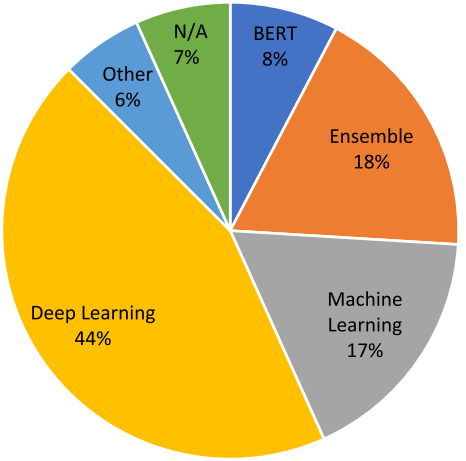

Figure 1. Pie chart adapted from (Zampieri *et al.* 2020) showing the models used in sub-task A. "N/A" indicates that the system did not have a description. Under machine learning, we included all approaches based on traditional classifiers such as SVMs and Naive Bayes. Under deep learning, we included approaches based on neural architectures available at that time except BERT.

using a BERT model. However, among the top-10 teams in the sub-task, seven used BERT including the best submission (Liu *et al.* 2019) which confirmed BERT as the state-of-the-art model for this task in 2019.

3.3 OffensEval 2020

Building on the success of OffensEval 2019, a second edition of OffensEval was organized in 2020. The organizers, created and publicly released the Semi-Supervised Offensive Language

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

Table 4. Data statistics for OffensEval 2020 sub-task A from Zampieri *et al.* (2020)

	Training			Test		
Language	OFF	NOT	Total	OFF	NOT	Total
English	1,448,861	7,640,279	9,089,140	1,090	2,807	3,897
Arabic	1,589	6,411	8,000	402	1,598	2,000
Danish	384	2,577	2,961	41	288	329
Greek	2,486	6,257	8,743	425	1,119	1,544
Turkish	6,131	25,625	31,756	716	2,812	3,528

Identification Dataset (SOLID) (Rosenthal *et al.* 2021), a large-scale offensive language identification dataset containing nine million English tweets with labels attributed using a semi-supervised method with OLID as a seed dataset. The creators of SOLID employed democratic co-training (Zhou and Goldman 2004), a semi-supervised technique used to create large datasets with noisy labels when provided with a set of diverse models each trained in a supervised way. Four models with different inductive biases were used with the goal of decreasing each individual model's bias, namely PMI (Turney and Littman 2003), FastText (Bojanowski *et al.* 2017), LSTM (Hochreiter and Schmidhuber 1997), and BERT (Devlin *et al.* 2019). Participants were provided with the dataset instances, an average prediction confidence scores by all models for each label, and the standard deviation between them. The idea behind this approach was to discourage participants from only using predictions from a specific model. The inclusion of confident scores instead of discrete labels by annotators combined with the large size of the dataset allowed participants to filter out weak data points and experiment with different thresholds when selecting instances for training.

While OffensEval 2019 was a monolingual shared task featuring only English data, OffensEval 2020 was a multilingual competition that introduced datasets in English and four other languages: Arabic (Mubarak *et al.* 2020), Danish (Sigurbergsson and Derczynski 2020), Greek (Pitenis, Zampieri, and Ranasinghe 2020), and Turkish (Çöltekin 2020). Five different tracks were organized in OffensEval 2020, one for each language. All datasets have been annotated according to the OLID taxonomy in levels A, B, and C. While annotation was available in the three levels, only the English track featured sub-tasks A, B, and C as in OffensEval 2019. The Arabic, Danish, Greek, and Turkish tracks featured only sub-task A. The instances in the five datasets and their label distribution for sub-task A are presented in Table 4.

The availability of datasets in various languages annotated according to the same taxonomy opened the possibility for cross-lingual training and analysis. In Table 5, we present examples from the five OffensEval 2020 datasets along with their labels in A, B, and C.

A total of 528 teams signed up to participate in OffensEval 2020 and a total of 145 teams submitted official runs to the competition. Participation varied across languages. The English track attracted 87 submissions to sub-task A while the Greek track attracted 37 teams. In Table 6, we present the results of the top-10 systems in the English track for sub-tasks A, B, and C.

In OffensEval 2020, we have observed that pre-trained language models based on transformer architectures had become the dominant paradigm in NLP. The clear majority of teams used pre-trained transformer models such as BERT, XLM-RoBERTa, and their variations. The top-10 teams used BERT, RoBERTa, or XLM-RoBERTa, often part of ensembles that also included CNNs and LSTMs (Hochreiter and Schmidhuber 1997). The best submission in sub-task A by (Wiedemann, Yimam, and Biemann 2020) achieved 0.9204 F1 score using a RoBERTa-large model fine-tuned

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

Table 5. Annotated examples for all sub-tasks and languages adapted from Zampieri *et al.* (2020)

Language	Tweet	A	B	C
English	This account owner asks for people to think rationally.	NOT	—	—
Arabic	الكلب يابن جبان يا سباك يا عليك الله لعنة	OFF	TIN	IND
	Translation: May God curse you, O coward, O son of a dog.			
Danish	Du glemmer Østeuropaer som er de værste	OFF	TIN	GRP
	Translation: You forget Eastern Europeans, who are the worst			
Greek	Παραδέξου το, είσαι αγάμητη εδώ και καιρό...	OFF	TIN	IND
	Translation: Admit it, you've been unfucked for a while now. . .			
Turkish	Böyle devam et seni gerizekalı *Translation: Go on like this, you idiot*	OFF	TIN	IND
English	this job got me all the way fucked up real shit	OFF	UNT	—
English	wtf ari her ass tooo big	OFF	TIN	IND
English	@USER We are a country of morons	OFF	TIN	GRP

Table 6. Results for the top-10 teams in English sub-task A ordered by macro-averaged F1

Sub-task A		Sub-task B	Sub-task C
Team ranks	F1 score	F1 score	F1 score
1	0.920	0.746	0.714
2	0.919	0.736	0.670
3	0.918	0.690	0.669
4	0.916	0.673	0.668
5	0.916	0.668	0.654
6	0.915	0.665	0.648
7	0.914	0.659	0.647
8	0.913	0.657	0.639
9	0.913	0.652	0.638
10	0.913	0.644	0.634

on the SOLID dataset and the second best submission by (Wang *et al.* 2020) use an ensemble of ALBERT models.

In addition to the English results discussed in this section, OffensEval 2020 featured the aforementioned Arabic, Danish, Greek, and Turkish tracks. Due to the limited availability of LLMs that are trained for languages other than English, we were able to include only Arabic, Greek, and Turkish in the evaluation presented in Section 4 leaving Danish out. For Arabic, Danish, and Greek, only one model, Flan-T5, was available. The unavailability of suitable models unfortunately did not allow us to perform a thorough evaluation of LLM performance for these languages. For this reason, we discuss the results on these languages only in Section 4. We refer to the OffensEval

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

2020 report (Zampieri *et al.* 2020) where the interested reader can find more information about these language tracks.

4. Benchmarking LLMs for offensive language online

In this section, we carry out an evaluation of different models on the OffensEval 2019 and OffensEval 2020 test sets. We selected six popular open-source models of the latest generation of LLMs developed between 2022 and 2023. We also use two task fined-tuned BERT models that have proven to achieve competitive performance in this task. We present all models, baselines, prompting strategies, and the results obtained by the tested models compared to the best entries at OffensEval.

4.1 LLMs

Falcon-7B-Instruct (Penedo *et al.* 2023), henceforth Falcon, is a decoder-only model fine-tuned with instruct and chat datasets. This model was adapted from GPT-3 (Brown *et al.* 2020) model, with differences in the positional embeddings, attention, and decoder-block components. The base model Falcon-7B, on which this model was fine-tuned on, outperforms other open-source LLM models like MPT-7B and RedPajama, among others. The limitation of this model is that it was mostly trained on English data, and hence, it does not perform well on other languages.

RedPajama-INCITE-7B-Instruct (Computer 2023), henceforth RedPajama, is an open-source LLM, based on the RedPajama-INCITE-7B-Base model and fine-tuned for few-shot applications. The model was trained only for English.

MPT-7B-Instruct (Team 2023), henceforth MPT, is a model fine-tuned on the base MPT-7B model. The model uses a modified decoder-only architecture, with the standard transformer been modified using the FlashAttention (Dao *et al.* 2022), Attention with Linear Biases (AliBi) (Press, Smith, and Lewis 2021) instead of positional embeddings, and it also does not use biases. Similar to the Falcon model, this model was trained using only the English data.

Llama-2-7B-Chat (Touvron *et al.* 2023b), henceforth Llama 2, is an auto-regressive language model with optimized transformer architecture. This model was optimized for dialogue use case. The model was trained using publicly available online data. The model outperforms most other open-source chat models and has a performance similar to models like ChatGPT. This model, however, works best only for English.

T0-3B (Sanh *et al.* 2021), henceforth T0, is a encoder-decoder-based model trained on several tasks using prompts. This model is based on the T5 model. The model was trained with a standard language model and using datasets for several NLP task. Similar to the other language models, this model also does not support non-English text.

Flan-T5-large (Chung *et al.* 2022), henceforth Flan-T5, is a language model based on the T5 (Raffel *et al.* 2020) model, which is a Text-to-Text transformer model. This model was fine-tuned for better zero-shot and few-shot learning for over 1000 different tasks. The model is one of the few LLM with support for languages other than English.

4.2 Task-specific BERT models

hateBERT (Caselli *et al.* 2021) is a BERT-based model, trained using the Reddit comments dataset, containing about one million posts from communities banned for being offensive and abusive. It

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

is a monolingual model specifically trained using data in English. The model outperforms the general pre-trained language models on offensive language detection datasets.

fBERT (Sarkar *et al.* 2021) is an offensive language detection BERT-based model, retrained using the SOLID dataset. The model was trained using about 1.4 million offensive language post from social media platform. The model has more domain-specific offensive language features, and it outperforms the general BERT model on offensive language datasets.

4.3 Zero-shot prompting

Zero-shot prompting is an approach where we do not provide any examples for the inputs and outputs. This approach is especially useful when we do not have labeled data for the classes. The model must make predictions based on the prior knowledge without knowing much about the new classes. The following prompt was used for our evaluations.

> *Comments containing any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct are offensive comments. This includes insults, threats, and posts containing profane language or swear words. Comments that do not contain offense or profanity are not offensive.*
> *Question: In one word, is the following comment offensive or not? They should just shut up*
> *Answer: The comment is*

As seen in the above example, the initial part of the prompt is a summary of the task. This is not always required, and we can directly ask the question and get the answer from the LLM. Depending on the specific requirements of the models, it is required to add special tokens to the prompt. For example, the Llama 2 model requires special tokens like "*[INST], [/INST]*," and "*<<SYS>>, <</SYS>>*," to indicate model instructions.

4.4 Zero-shot learning

Zero-shot learning approach is based on the latent embedding approach, where the input and the labels are embedded into a latent representation, using an existing model (Veeranna *et al.* 2016). This approach is commonly used with the sentence embedding models. We evaluate the hateBERT and fBERT models using this approach. For a given input instance, the embedding vector for the input is obtained from the last four layers of the model. Similarly, the embedding vector for all the labels is generated. The cosine similarity between the embeddings of the input and each of the label embeddings is calculated. The input is assigned a label with which it has the highest cosine similarity.

4.5 Baselines

For OffensEval 2019, the top three systems for sub-task A were used as baseline. The best-performing team at OffensEval 2019 preprocessed the training dataset which was then used to fine-tune a pre-trained BERT (Devlin *et al.* 2019) model (Liu *et al.* 2019). The model was only fine-tuned for two epochs. The second placed team also used a BERT model, but they used pre-trained GloVe vectors (Pennington, Socher, and Manning 2014) while also addressing the class imbalance (Nikolov and Radivchev 2019). The third placed team also fine-tuned a pre-trained BERT model, but they used different hyperparameters (Zhu, Tian, and Kübler 2019).

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

The OffensEval 2020 included five languages: English, Arabic, Greek, Turkish, and Danish. For the English track, the best-performing team fine-tuned four different ALBERT (Lan *et al.* 2020) models (Wiedemann *et al.* 2020) and used the ensemble of these fine-tuned models for prediction. The second best team (Wang *et al.* 2020) also used an ensemble approach, where they first fine-tuned two multilingual XLM-RoBERTa models, XLM-RoBERTa-base, and XLM-RoBERTa-large (Conneau *et al.* 2019). In comparison, the third placed team fine-tuned only one multilingual XLM-RoBERTa model (Dadu and Pant 2020).

For the Arabic track, the first placed team used the AraBERT model (Antoun, Baly, and Hajj 2020) to encode the tweets, and a sigmoid classifier was then trained using the encoded tweets (Alami *et al.* 2020). The second placed team (Hassan *et al.* 2020) used an ensemble of SVM, CNN-BiLSTM, and multilingual BERT models. Each of these models used different features, with character and word level *n*-grams along with the word embeddings used as the features for the SVM model, whereas the CNN-BiLSTM model used character and word embeddings. The third best team (Wang *et al.* 2020) used an ensemble of the XLM-RoBERTa-base and XLM-RoBERTa-large models (Conneau *et al.* 2019).

The first placed team for the Greek track fine-tuned a pre-trained mBERT (Devlin *et al.* 2019) model, but they also used the embeddings of a domain-specific vocabulary generated using the WordPiece algorithm to fine-tune and pre-train the model (Ahn *et al.* 2020). The second placed team (Wang *et al.* 2020) used an ensemble of the XLM-RoBERTa-base and XLM-RoBERTa-large model (Conneau *et al.* 2019), while the third best team used a monolingual BERT model (Socha 2020).

As for the Turkish track, the first placed team (Wang *et al.* 2020) used an ensemble of the XLM-RoBERTa-base and XLM-RoBERTa-large model (Conneau *et al.* 2019), the second best team (Ozdemir and Yeniterzi 2020) used an ensemble of CNN-LSTM, BiLSTM-Attention, and BERT models, with pre-trained word embeddings for tweets generated using the BERTurk, a BERT model for Turkish. The third placed team (Safaya, Abdullatif, and Yuret 2020) combined the BERTurk model with the CNN model.

4.6 OffensEval 2019 results

We present all evaluation results in Table 7. We observed that performance varied widely between all models tested ranging from 0.793 F1 score obtained by Flan-T5 to 0.267 obtained by RedPajama. A surprising outcome of this evaluation is the low performance of the task fine-tuned models which did not obtain competitive performance compared to the top-3 teams at OffensEval or even some of the LLMs. Flan-T5 was the only model that achieved competitive performance close to the 0.800 F1 score obtained by the competition's CNN baseline model. All in all, all models tested were outperformed by the best three systems of the competition.

4.7 OffensEval 2020 results

We present the results on the OffensEval 2020 dataset in Table 8. For OffensEval 2020 also, all the LLMs with the zero-shot prompting approach could not outperform the best three systems. Flan-T5, however, comes very close to the top teams in the competition with a 0.910 macro-F1 score. Llama 2 also follows closely with 0.874 macro-F1 score. The rest of the models do not perform well, falling behind 0.750 macro-F1 score.

Language coverage is one of the known bottlenecks of LLMs. Most LLMs tested in this study do not support non-English languages. The only model that supports some of the OffensEval 2020 languages is Flan-T5-large. The results for Arabic, Greek, and Turkish are shown in Table 9.

The results show that the Flan-T5-large model does not outperform the top three systems in any of the three languages. Furthermore, it should be noted that in English the gap between Flan-T5-large and the third placed system is only 0.01 macro-F1 score. However, in all these three languages, Flan-T5-large has a larger gap with the third place system, suggesting that the model is

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

Table 7. Macro-F1 scores for the OffensEval 2019 test set. Baseline results displayed in italics

Model	Macro-F1
OffensEval Rank 1	*0.829*
OffensEval Rank 2	*0.815*
OffensEval Rank 3	*0.814*
Flan-T5	0.793
Llama 2	0.715
Falcon	0.648
MPT	0.547
hateBERT	0.507
T0	0.430
fBERT	0.329
RedPajama	0.267

Table 8. Macro-F1 scores for the OffensEval 2020 English test set. Baseline results are displayed in italics

Model	Macro-F1
OffensEval Rank 1	*0.920*
OffensEval Rank 2	*0.919*
OffensEval Rank 3	*0.918*
Flan-T5	0.910
Llama 2	0.874
MPT	0.736
Falcon	0.734
hateBERT	0.552
T0	0.397
RedPajama	0.375
fBERT	0.338

weaker in detecting offensive language in non-English languages. Most possibly, this is due to the training data limitation in non-English languages in Flan-T5-large.

5. Conclusion and future work

This paper presented a survey and evaluation of offensive language identification benchmark competitions with a focus on OffensEval. We used zero-shot prompting on six state-of-the-art

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

Table 9. Macro-F1 scores for the OffensEval 2020 Arabic, Greek, and Turkish test sets. Baseline results are displayed in italics

	Arabic	Greek	Turkish
Model	F1 score	F1 score	F1 score
OffensEval Rank 1	*0.902*	*0.852*	*0.826*
OffensEval Rank 2	*0.901*	*0.851*	*0.817*
OffensEval Rank 3	*0.899*	*0.848*	*0.814*
Flan-T5-large	0.530	0.532	0.451

LLMs and zero-shot learning on two task fine-tuned BERT models and compared their performance to the best entries submitted to OffensEval 2019 and 2020 for Arabic, English, Greek, and Turkish. Our results indicate that while some new LLMs such as Flan-T5 achieve competitive results, all LLMs tested achieved lower performance than the best three systems in those competitions which were based on more well-established transformer-based models such as BERT and ELMo. This suggests that while LLMs have been achieving impressive performance on a variety of tasks, particularly in those that involve next-word prediction and text generation, their zero-shot performance on this task is still not up to the same standard as transformer models trained on in-domain data.

Given the relatively recent introduction of the latest generation of LLMs, there are several avenues we would like to explore in the future that will help us better understand the performance of these models on offensive language identification. One of the most promising future directions is to evaluate possible data contamination in LLMs (Golchin and Surdeanu 2023). Unfortunately, most LLM developers provide very limited information on the data these models are trained on. Therefore, it is currently not possible to know how benchmark datasets are used in the training of these models and whether shared task test sets are used in the training stage. Further investigation is required to determine the extent of data contamination in LLMs.

Finally, an obvious limitation of this study is the limited support by LLMs for languages other than English. We were able to prompt Flan-T5 for Arabic, Greek, and Turkish but Danish, which was also included in OffensEval, was not supported by any of the models. As new models are released every month and developers work to include more languages in them, we would like to replicate this study on all OffensEval languages using more LLMs in the future.

References

Ahn H., **Sun J.**, **Park C. Y. and Seo J.** (2020). NLPDove at SemEval-2020 task 12: improving offensive language detection with cross-lingual transfer. In *Proceedings of SemEval*.

Alami H., **Ouatik El Alaoui S.**, **Benlahbib A. and En-nahnahi N.** (2020). LISAC FSDM-USMBA team at SemEval-2020 task 12: overcoming AraBERT's pretrain-finetune discrepancy for Arabic offensive language identification. In *Proceedings of SemEval*.

Antoun W., **Baly F. and Hajj H.** (2020). AraBERT: transformer-based model for Arabic language understanding. In *Proceedings of OSACT*.

Arora A., **Nakov P.**, **Hardalov M.**, **Sarwar S. M.**, **Nayak V.**, **Dinkov Y.**, **Zlatkova D.**, **Dent K.**, **Bhatawdekar A.**, **Bouchard G. and Augenstein I.** (2023). Detecting harmful content on online platforms: what platforms need vs. where research efforts go. *ACM Computing Surveys* **56**(3), 1–17.

Aroyehun S. T. and Gelbukh A. (2018). Aggression detection in social media: using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of TRAC*.

Banerjee S., **Sarkar M.**, **Agrawal N.**, **Saha P. and Das M.** (2021). Exploring transformer based models to identify hate speech and offensive content in English and Indo-Aryan languages. In *Proceedings of FIRE*.

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

Basile V., Bosco C., Fersini E., Nozza D., Patti V., Pardo F. M. R., Rosso P. and Sanguinetti M. (2019). Semeval-2019 task 5: multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of SemEval*.

Bhatia M., Bhotia T. S., Agarwal A., Ramesh P., Gupta S., Shridhar K., Laumann F. and Dash A. (2021). One to rule them all: towards joint indic language hate speech detection. In *Proceedings of FIRE*.

Bojanowski P., Grave E., Joulin A. and Mikolov T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* **5**, 135–146.

Bornheim T., Grieger N. and Bialonski S. (2021). Fhac at GermEval 2021: identifying german toxic, engaging, and fact-claiming comments with ensemble learning. In *Proceedings of GermEval*.

Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A. and others (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems* **33**, 1877–1901.

Caselli T., Basile V., Mitrović J. and Granitzer M. (2021). Hatebert: retraining bert for abusive language detection in English. In *Proceedings of WOAH*.

Çöltekin C. (2020). A corpus of Turkish offensive language on social media. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*.

Chowdhery A., Narang S., Devlin J., Bosma M., Mishra G., Roberts A., Barham P., Chung H. W., Sutton C., Gehrmann S., Schuh P., Shi K., Tsvyashchenko S., Maynez J., Rao A., Barnes P., Tay Y., Shazeer N., Prabhakaran V., Reif E., Du N., Hutchinson B., Pope R., Bradbury J., Austin J., Isard M., Gur-Ari G., Yin P., Duke T., Levskaya A., Ghemawat S., Dev S., Michalewski H., Garcia X., Misra V., Robinson K., Fedus L., Zhou D., Ippolito D., Luan D., Lim H., Zoph B., Spiridonov A., Sepassi R., Dohan D., Agrawal S., Omernick M., Dai A. M., Pillai T. S., Pellat M., Lewkowycz A., Moreira E., Child R., Polozov O., Lee K., Zhou Z., Wang X., Saeta B., Diaz M., Firat O., Catasta M., Wei J., Meier-Hellstern K., Eck D., Dean J., Petrov S. and Fiedel N. (2022). PaLM: scaling language modeling with pathways. arXiv preprint arXiv: 2204.02311.

Chung H. W., Hou L., Longpre S., Zoph B., Tay Y., Fedus W., Li E., Wang X., Dehghani M., Brahma S., Webson A., Gu S. S., Dai Z., Suzgun M., Chen X., Chowdhery A., Castro-Ros A., Pellat M., Robinson K., Valter D., Narang S., Mishra G., Yu A., Zhao V., Huang Y., Dai A., Yu H., Petrov S., Chi E. H., Dean J., Devlin J., Roberts A., Zhou D., Le Q. V. and Wei J. (2022). Scaling instruction-finetuned language models. arXiv preprint arXiv: 2210.11416.

Computer T. (2023). Redpajama: an open source recipe to reproduce llama training dataset. https://github.com/togethercomputer/RedPajama-Data

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L. and Stoyanov V. (2019). Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Dadu T. and Pant K. (2020). Team rouges at SemEval-2020 task 12: cross-lingual inductive transfer to detect offensive language. In *Proceedings of SemEval*.

Dao T., Fu D., Ermon S., Rudra A. and Ré C. (2022). Flashattention: fast and memory-efficient exact attention with IO-awareness. In *Proceedings of NeurIPS*.

Devlin J., Chang M.-W., Lee K. and Toutanova K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Dikshitha Vani V. and Bharathi B. (2022). Hate speech and offensive content identification in multiple languages using machine learning algorithms. In *Proceedings of FIRE*.

Doddapaneni S., Aralikatte R., Ramesh G., Goyal S., Khapra M. M., Kunchukuttan A. and Kumar P. (2023). Towards leaving no Indic language behind: building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of ACL*.

Fersini E., Gasparini F., Rizzi G., Saibene A., Chulvi B., Rosso P., Lees A. and Sorensen J. (2022). SemEval-2022 task 5: multimedia automatic misogyny identification. In Emerson G., Schluter N., Stanovsky G., Kumar R., Palmer A., Schneider N., Singh S. and Ratan, S. (eds), Proceedings of SemEval.

Fortuna P. and Nunes S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* **51**(4), 1–30.

Gaikwad S. S., Ranasinghe T., Zampieri M. and Homan C. (2021). Cross-lingual offensive language identification for low resource languages: the case of Marathi. In *Proceedings of RANLP*.

Golchin S. and Surdeanu M. (2023). Time travel in LLMS: tracing data contamination in large language models. arXiv preprint arXiv: 2308.08493.

Han J., Liu X. and Wu S. (2019). jhan014 at SemEval-2019 task 6: identifying and categorizing offensive language in social media. In *Proceedings of SemEval*.

Hassan S., Samih Y., Mubarak H. and Abdelali A. (2020). ALT at SemEval-2020 task 12: Arabic and English offensive language identification in social media. In *Proceedings of SemEval*.

Hochreiter S. and Schmidhuber J. (1997). Long short-term memory. *Neural Computation* **9**(8), 1735–1780.

Indurthi V., Syed B., Shrivastava M., Chakravartula N., Gupta M. and Varma V. (2019). FERMI at SemEval-2019 task 5: using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of SemEval*.

Khanuja S., Bansal D., Mehtani S., Khosla S., Dey A., Gopalan B., Margam D. K., Aggarwal P., Nagipogu R. T., Dave S. and others (2021). Muril: multilingual representations for indian languages. arXiv preprint arXiv: 2103.10730.

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

Kirk H., Yin W., Vidgen B. and Röttger P. (2023). SemEval-2023 task 10: explainable detection of online sexism. In *Proceedings of SemEval*.

Kumar R., Ojha A. K., Malmasi S. and Zampieri M. (2018). Benchmarking aggression identification in social media. In *Proceedings of TRAC*.

Kumar R., Ojha A. K., Malmasi S. and Zampieri M. (2020). Evaluating aggression identification in social media. In *Proceedings of TRAC*.

Kumari K., Srivastav S. and Suman R. R. (2022). Bias, threat and aggression identification using machine learning techniques on multilingual comments. In *Proceedings of TRAC*.

Lan Z., Chen M., Goodman S., Gimpel K., Sharma P. and Soricut R. (2020). Albert: a lite bert for self-supervised learning of language representations. In *Proceedings of ICLR*.

Liu P., Li W. and Zou L. (2019). NULI at SemEval-2019 task 6: transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of SemEval*.

Liu P., Yuan W., Fu J., Jiang Z., Hayashi H. and Neubig G. (2023). Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Computing Surveys* **55**(9), 1–35.

Mandl T., Modha S., Kumar M. A. and Chakravarthi B. R. (2020). Overview of the HASOC track at fire 2020: hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German. In *Proceedings of FIRE*.

Mandl T., Modha S., Majumder P., Patel D., Dave M., Mandlia C. and Patel A. (2019). Overview of the HASOC track at fire 2019: hate speech and offensive content identification in Indo-European languages. In *Proceedings of FIRE*.

Mandl T., Modha S., Shahi G. K., Madhu H., Satapara S., Majumder P., Schäfer J., Ranasinghe T., Zampieri M., Nandini D. and Jaiswal A. K. (2021). Overview of the HASOC subtrack at fire 2021: hate speech and offensive content identification in English and Indo-Aryan languages. In *Proceedings of FIRE*.

Meaney J., Wilson S., Chiruzzo L., Lopez A. and Magdy W. (2021). Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense. In *Proceedings of SemEval*, pp. 105–119.

Mishra A. K., Saumya S. and Kumar A. (2020). IIIT_DWD@ HASOC 2020: identifying offensive content in Indo-European languages. In *Proceedings of FIRE*.

Modha S., Mandl T., Majumder P., Satapara S., Patel T. and Madhu H. (2022). Overview of the hasoc subtrack at fire 2022: identification of conversational hate-speech in hindi-english code-mixed and German language. In *Proceedings of FIRE*.

Modha S., Mandl T., Shahi G. K., Madhu H., Satapara S., Ranasinghe T. and Zampieri M. (2021). Overview of the hasoc subtrack at fire 2021: hate speech and offensive content identification in English and Indo-Aryan languages and conversational hate speech. In *Proceedings of FIRE*.

Morgan S., Ranasinghe T. and Zampieri M. (2021). WLV-RIT at GermEval 2021: multitask learning with transformers to detect toxic, engaging, and fact-claiming comments. In *Proceedings of GermEval*.

Mubarak H., Rashed A., Darwish K., Samih Y. and Abdelali A. (2020). Arabic offensive language on Twitter: analysis and experiments. arXiv preprint arXiv: 2004.02192.

Nene M., North K., Ranasinghe T. and Zampieri M. (2021). Transformer models for offensive language identification in Marathi. In *Proceedings of FIRE*.

Nikolov A. and Radivchev V. (2019). Nikolov-radivchev at SemEval-2019 task 6: offensive tweet classification with BERT and ensembles. In *Proceedings of SemEval*.

Ozdemir A. and Yeniterzi R. (2020). SU-NLP at SemEval-2020 task 12: offensive language IdentifiCation in Turkish tweets. In *Proceedings of SemEval*.

Pavlopoulos J., Sorensen J., Laugier L. and Androutsopoulos I. (2021). SemEval-2021 task 5: toxic spans detection. In *Proceedings of SemEval*.

Penedo G., Malartic Q., Hesslow D., Cojocaru R., Cappelli A., Alobeidli H., Pannier B., Almazrouei E. and Launay J. (2023). The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv: 2306.01116.

Pennington J., Socher R. and Manning C. (2014). GloVe: global vectors for word representation. In *Proceedings of the EMNLP*.

Peters M. E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K. and Zettlemoyer L. (2018). Deep contextualized word representations. In Walker M., Ji H. and Stent A. (eds), *Proceedings of NAACL-HLT*.

Pitenis Z., Zampieri M. and Ranasinghe T. (2020). Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*.

Plaza-del Arco F. M., Molina-González M. D. and Alfonso L. (2021). OffendES: A New Corpus in Spanish for Offensive Language Research. In *Proceedings of RANLP*.

Poletto F., Basile V., Sanguinetti M., Bosco C. and Patti V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation* **55**(2), 477–523.

Press O., Smith N. A. and Lewis M. (2021). Train short, test long: attention with linear biases enables input length extrapolation. arXiv preprint arXiv: 2108.12409.

Radford A., Wu J., Child R., Luan D., Amodei D. and Sutskever I. (2019). Language models are unsupervised multitask learners. https://api.semanticscholar.org/CorpusID:160025533

Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W. and Liu P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* **21**, 1–67.

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

Raj R., Srivastava S. and Saumya S. (2020). NSIT & IIITDWD@ HASOC 2020: deep learning model for hate-speech identification in Indo-European languages. In *Proceedings of FIRE*.

Ranasinghe T., Anuradha I., Premasiri D., Silva K., Hettiarachchi H., Uyangodage L. and Zampieri M. (2022a). Sold: Sinhala offensive language dataset. arXiv preprint arXiv: 2212.00851.

Ranasinghe T., North K., Premasiri D. and Zampieri M. (2022b). Overview of the hasoc subtrack at fire 2022: offensive language identification in Marathi. In *Proceedings of FIRE*.

Ranasinghe T. and Zampieri M. (2020). Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of EMNLP*.

Ranasinghe T. and Zampieri M. (2021). *An Evaluation of Multilingual Offensive Language Identification Methods for the Languages of India*. Basel: Information.

Ranasinghe T., Zampieri M. and Hettiarachchi H. (2019). Brums at hasoc 2019: deep learning models for multilingual hate speech and offensive language identification. In *Proceedings of FIRE*.

Risch J. and Krestel R. (2020). Bagging BERT models for robust aggression identification. In *Proceedings of TRAC*.

Risch J., Stoll A., Wilms L. and Wiegand M. (2021). Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of GermEval*, pp. 1–12.

Rosenthal S., Atanasova P., Karadzhov G., Zampieri M. and Nakov P. (2021). SOLID: a large-scale weakly supervised dataset for offensive language identification. In *Findings of the ACL*.

Safaya A., Abdullatif M. and Yuret D. (2020). KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of SemEval*.

Sanh V., Webson A., Raffel C., Bach S. H., Sutawika L., Alyafeai Z., Chaffin A., Stiegler A., Scao T. L., Raja A., Dey M., Bari M. S., Xu C., Thakker U., Sharma S. S., Szczechla E., Kim T., Chhablani G., Nayak N., Datta D., Chang J., Jiang M. T.-J., Wang H., Manica M., Shen S., Yong Z. X., Pandey H., Bawden R., Wang T., Neeraj T., Rozen J., Sharma A., Santilli A., Fevry T., Fries J. A., Teehan R., Biderman S., Gao L., Bers T., Wolf T. and Rush A. M. (2021). Multitask prompted training enables zero-shot task generalization. arXiv preprint arXiv: 2110.08207.

Sarkar D., Zampieri M., Ranasinghe T. and Ororbia A. (2021). fbert: a neural transformer for identifying offensive content. In *Findings of EMNLP*.

Satapara S., Majumder P., Mandl T., Modha S., Madhu H., Ranasinghe T., Zampieri M., North K. and Premasiri D. (2022). Overview of the hasoc subtrack at fire 2022: hate speech and offensive content identification in english and indo-aryan languages. In *Proceedings of FIRE*.

Scao T. L., Fan A., Akiki C., Pavlick E., Ilić S., Hesslow D., Castagné R., Luccioni A. S., Yvon F., Gallé M., Tow J., Rush A. M., Biderman S., Webson A., Ammanamanchi P. S., Wang T., Sagot B., Muennighoff N., del Moral A. V., Ruwase O., Bawden R., Bekman S., McMillan-Major A., Beltagy I., Nguyen H., Saulnier L., Tan S., Suarez P. O., Sanh V., Laurençon H., Jernite Y., Launay J., Mitchell M., Raffel C., Gokaslan A., Simhi A., Soroa A., Aji A. F., Alfassy A., Rogers A., Nitzav A. K., Xu C., Mou C., Emezue C., Klamm C., Leong C., van Strien D., Adelani D. I., Radev D., Ponferrada E. G., Levkovizh E., Kim E., Natan E. B., Toni F. D., Dupont G., Kruszewski G., Pistilli G., Elsahar H., Benyamina H., Tran H., Yu I., Abdulmumin I., Johnson I., Gonzalez-Dios I., de la Rosa J., Chim J., Dodge J., Zhu J., Chang J., Frohberg J., Tobing J., Bhattacharjee J., Almubarak K., Chen K., Lo K., Werra L. V., Weber L., Phan L., allal L. B., Tanguy L., Dey M., Muñoz M. R., Masoud M., Grandury M., Šaško M., Huang M., Coavoux M., Singh M., Jiang M. T.-J., Vu M. C., Jauhar M. A., Ghaleb M., Subramani N., Kassner N., Khamis N., Nguyen O., Espejel O., de Gibert O., Villegas P., Henderson P., Colombo P., Amuok P., Lhoest Q., Harliman R., Bommasani R., López R. L., Ribeiro R., Osei S., Pyysalo S., Nagel S., Bose S., Muhammad S. H., Sharma S., Longpre S., Nikpoor S., Silberberg S., Pai S., Zink S., Torrent T. T., Schick T., Thrush T., Danchev V., Nikoulina V., Laippala V., Lepercq V., Prabhu V., Alyafeai Z., Talat Z., Raja A., Heinzerling B., Si C., Taşar D. E., Salesky E., Mielke S. J., Lee W. Y., Sharma A., Santilli A., Chaffin A., Stiegler A., Datta D., Szczechla E., Chhablani G., Wang H., Pandey H., Strobelt H., Fries J. A., Rozen J., Gao L., Sutawika L., Bari M. S., Al-shaibani M. S., Manica M., Nayak N., Teehan R., Albanie S., Shen S., Ben-David S., Bach S. H., Kim T., Bers T., Fevry T., Neeraj T., Thakker U., Raunak V., Tang X., Yong Z.-X., Sun Z., Brody S., Uri Y., Tojarieh H., Roberts A., Chung H. W., Tae J., Phang J., Press O., Li C., Narayanan D., Bourfoune H., Casper J., Rasley J., Ryabinin M., Mishra M., Zhang M., Shoeybi M., Peyrounette M., Patry N., Tazi N., Sanseviero O., von Platen P., Cornette P., Lavallée P. F., Lacroix R., Rajbhandari S., Gandhi S., Smith S., Requena S., Patil S., Dettmers T., Baruwa A., Singh A., Cheveleva A., Ligozat A.-L., Subramonian A., Névéol A., Lovering C., Garrette D., Tunuguntla D., Reiter E., Taktasheva E., Voloshina E., Bogdanov E., Winata G. I., Schoelkopf H., Kalo J.-C., Novikova J., Forde J. Z., Clive J., Kasai J., Kawamura K., Hazan L., Carpuat M., Clinciu M., Kim N., Cheng N., Serikov O., Antverg O., van der Wal O., Zhang R., Zhang R., Gehrmann S., Mirkin S., Pais S., Shavrina T., Scialom T., Yun T., Limisiewicz T., Rieser V., Protasov V., Mikhailov V., Pruksachatkun Y., Belinkov Y., Bamberger Z., Kasner Z., Rueda A., Pestana A., Feizpour A., Khan A., Faranak A., Santos A., Hevia A., Unldreaj A., Aghagol A., Abdollahi A., Tammour A., HajiHosseini A., Behroozi B., Ajibade B., Saxena B., Ferrandis C. M., McDuff D., Contractor D., Lansky D., David D. and Kiela D. (2023). BLOOM: a 176B-parameter open-access multilingual language model. arXiv preprint arXiv: 2211.05100.

Sigurbergsson G. I. and Derczynski L. (2020). Offensive language and hate speech detection for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*.

https://doi.org/10.1017/S1351324923000517 Published online by Cambridge University Press

Singh N. K. and Garain U. (2022). An analysis of transformer-based models for code-mixed conversational hate-speech identification. In *Proceedings of FIRE*.

Socha K. (2020). KS@LTH at SemEval-2020 task 12: fine-tuning multi- and monolingual transformer models for offensive language detection. In *Proceedings of SemEval*.

Taulé M., **Ariza A.**, **Nofre M.**, **Amigó E. and Rosso P.** (2021). Overview of DETOXIS at IberLEF 2021: detection of toxicity in comments in Spanish. *Procesamiento del Lenguaje Natural* **67**, 209–221.

Team M. N. (2023). Introducing MPT-7B: a new standard for open-source, commercially usable LLMS. www.mosaicml.com/blog/mpt-7b. Accessed: 2023-03-28.

Touvron H., **Martin L.**, **Stone K.**, **Albert P.**, **Almahairi A.**, **Babaei Y.**, **Bashlykov N.**, **Batra S.**, **Bhargava P.**, **Bhosale S.**, **Bikel D.**, **Blecher L.**, **Ferrer C. C.**, **Chen M.**, **Cucurull G.**, **Esiobu D.**, **Fernandes J.**, **Fu J.**, **Fu W.**, **Fuller B.**, **Gao C.**, **Goswami V.**, **Goyal N.**, **Hartshorn A.**, **Hosseini S.**, **Hou R.**, **Inan H.**, **Kardas M.**, **Kerkez V.**, **Khabsa M.**, **Kloumann I.**, **Korenev A.**, **Koura P. S.**, **Lachaux M.-A.**, **Lavril T.**, **Lee J.**, **Liskovich D.**, **Lu Y.**, **Mao Y.**, **Martinet X.**, **Mihaylov T.**, **Mishra P.**, **Molybog I.**, **Nie Y.**, **Poulton A.**, **Reizenstein J.**, **Rungta R.**, **Saladi K.**, **Schelten A.**, **Silva R.**, **Smith E. M.**, **Subramanian R.**, **Tan X. E.**, **Tang B.**, **Taylor R.**, **Williams A.**, **Kuan J. X.**, **Xu P.**, **Yan Z.**, **Zarov I.**, **Zhang Y.**, **Fan A.**, **Kambadur M.**, **Narang S.**, **Rodriguez A.**, **Stojnic R.**, **Edunov S. and Scialom T.** (2023a). Llama 2: open foundation and fine-tuned chat models. arXiv preprint arXiv: 2307.09288.

Touvron H., **Martin L.**, **Stone K.**, **Albert P.**, **Almahairi A.**, **Babaei Y.**, **Bashlykov N.**, **Batra S.**, **Bhargava P.**, **Bhosale S.**, et al. (2023b). Llama 2: open foundation and fine-tuned chat models. arXiv preprint arXiv: 2307.09288.

Turney P. D. and Littman M. L. (2003). Measuring praise and criticism: inference of semantic orientation from association. *ACM Transactions on Information Systems* **21**(4), 315–346.

Veeranna S. P., **Nam J.**, **Mencıa E. L. and Fürnkranz J.** (2016). Using semantic similarity for multi-label zero-shot classification of text documents. In *Proceeding of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Bruges: Elsevier.

Wang B., **Ding Y.**, **Liu S. and Zhou X.** (2019). Ynu_wb at hasoc 2019: ordered neurons LSTM with attention for identifying hate speech and offensive language. In *Proceedings of FIRE*.

Wang S., **Liu J.**, **Ouyang X. and Sun Y.** (2020). Galileo at SemEval-2020 task 12: multi-lingual learning for offensive language identification using pre-trained language models. In *Proceedings of SemEval*.

Weerasooriya T. C., **Dutta S.**, **Ranasinghe T.**, **Zampieri M.**, **Homan C. M. and KhudaBukhsh A. R.** (2023). Vicarious offense and noise audit of offensive speech classifiers: unifying human and machine disagreement on what is offensive. In *Proceedings of EMNLP*.

Wiedemann G., **Yimam S. M. and Biemann C.** (2020). UHH-LT at SemEval-2020 task 12: fine-tuning of pre-trained transformer networks for offensive language detection. In *Proceedings of SemEval*.

Zampieri M., **Malmasi S.**, **Nakov P.**, **Rosenthal S.**, **Farra N. and Kumar R.** (2019a). Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL*.

Zampieri M., **Malmasi S.**, **Nakov P.**, **Rosenthal S.**, **Farra N. and Kumar R.** (2019b). SemEval-2019 task 6: identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of SemEval*.

Zampieri M., **Nakov P.**, **Rosenthal S.**, **Atanasova P.**, **Karadzhov G.**, **Mubarak H.**, **Derczynski L.**, **Pitenis Z. and Çöltekin c** (2020). SemEval-2020 Task 12: multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of SemEval*.

Zhang J. and Wang Y. (2022). SRCB at SemEval-2022 task 5: pretraining based image to text late sequential fusion system for multimodal misogynous meme identification. In *Proceedings of SemEval*.

Zhang S., **Roller S.**, **Goyal N.**, **Artetxe M.**, **Chen M.**, **Chen S.**, **Dewan C.**, **Diab M.**, **Li X.**, **Lin X. V.**, **Mihaylov T.**, **Ott M.**, **Shleifer S.**, **Shuster K.**, **Simig D.**, **Koura P. S.**, **Sridhar A.**, **Wang T. and Zettlemoyer L.** (2022). OPT: open pre-trained transformer language models. arXiv preprint arXiv: 2205.01068.

Zhou M. (2023). PingAnLifeInsurance at SemEval-2023 task 10: using multi-task learning to better detect online sexism. In *Proceedings of SemEval*.

Zhou Y. and Goldman S. (2004). Democratic co-learning. In *16th IEEE International Conference on Tools with Artificial Intelligence*. IEEE, pp. 594–602.

Zhu J., **Tian Z. and Kübler S.** (2019). UM-IU@LING at SemEval-2019 task 6: identifying offensive tweets using BERT and SVMs. In *Proceedings of SemEval*.

Zhu Q., **Lin Z.**, **Zhang Y.**, **Sun J.**, **Li X.**, **Lin Q.**, **Dang Y. and Xu R.** (2021). Hitsz-hlt at semeval-2021 task 5: ensemble sequence labeling and span boundary detection for toxic span detection. In *Proceedings SemEval*.

Cite this article: Zampieri M, Rosenthal S, Nakov P, Dmonte A and Ranasinghe T (2023). OffensEval 2023: Offensive language identification in the age of Large Language Models. *Natural Language Engineering* **29**, 1416–1435. https://doi.org/10.1017/S1351324923000517