On the Errors in Code-Mixed Tamil-English Offensive Span Identification

Manikandan Ravikiran[†*], **Bharathi Raja Chakravarthi**[‡]
[†]Georgia Institute of Technology, Atlanta, Georgia
[‡]School of Computer Science, University of Galway, Ireland
mravikiran3@gatech.edu, bharathi.raja@insight-centre.org

Abstract

In recent times, offensive span identification in code-mixed Tamil-English language has seen traction with the release of datasets, shared tasks, and the development of multiple methods. However, the details of various errors shown by these methods are currently unclear. This paper presents a detailed analysis of various errors in state-of-the-art Tamil-English offensive span identification methods. Our study reveals the strengths and weaknesses of the widely used sequence labeling and zero-shot models for offensive span identification. In the due process, we identify data-related errors, improve data annotation and release additional diagnostic data to evaluate models' quality and stability. *Disclaimer: This paper contains examples that may be considered profane, vulgar, or offensive. The examples do not represent the views of the authors or their employers/graduate schools towards any person(s), group(s), practice(s), or entity/entities. Instead, they emphasize the complexity of various errors and linguistic research challenges.*

1 Introduction

Offensive span identification from code-mixed Tamil-English social media comments (Ravikiran and Annamalai, 2021) focuses on extracting character offsets corresponding to tokens contributing to offensiveness. Identifying such offensive spans is helpful in multiple facets ranging from assisting content moderators for quicker moderation to the development of semi-automated tools which can provide thorough attribution related to the intervened offensive content. Recently there are numerous methods (Ravikiran et al., 2022; Hariharan RamakrishnaIyer LekshmiAmmal, 2022) that are capable of identifying these offensive spans with accuracy as high as 60% on very hard-to-understand short sentences with limited contextual information.

However many of these methods rely on large code-mixed datasets (Chakravarthi, 2022, 2023; Chakravarthi et al., 2022a,b; Kumaresan et al., 2022) and pre-trained language models (Ravikiran and Annamalai, 2021). Nevertheless, these methods are still far away from solving offensive span identification despite such large success. To advance further with this, we need to understand better the sources of errors in the offensive span identification. Such an analysis will, in turn, help introduce inductive biases to extract the spans effectively. Thus, we analyze errors on the Tamil-English code-mixed offensive span identification dataset (DOSA-v2) which consists of 4816 (train) and 876 (test) offensive comments obtained from YouTube movie trailers with span annotations (Ravikiran et al., 2022).

Specifically, this work focuses on models' prediction errors and data-related errors. For the former case, we comprehensively investigate the predictions of 8 different models that currently exist for offensive span identification. Accordingly, we find that all the existing models suffer from issues ranging lack of identification of words or phrases that are commonly used to making mistakes due to context ambiguity. Based on this, we create eight different error categories suitable to measure the quality of models' predictions.

In the latter case, we find very few works to focus on error analysis of offensive span identification, with a predominant concentration on the English Language (Ding and Jurgens, 2021). Additionally, some works focus on error analysis of sequence labeling method (Stanislawek et al., 2019; Niklaus et al., 2018; Nguyen et al., 2019), but not from the point of offensive spans. In this work, in line with Ding and Jurgens (2021) we use human intervention for error analysis. More specifically, we create multiple error analysis teams consisting

[*] Corresponding Author: Work done during graduate school

Third Workshop on Speech and Language Technologies for Dravidian Languages,
pages 1–9, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_001

Methods	Model	F1
Token Labeling (Ravikiran and Annamalai, 2021)	Multilingual-BERT (M1)	0.5688
	RoBERTA (M2)	0.5721
	XLM-RoBERTA (M3)	0.5793
Zero-shot Rationale Extraction (Ravikiran and Chakravarthi, 2022)	RoBERTA+LIME (M4)	0.4886
	XLM-RoBERTA+LIME (M5)	0.4845
	XLM-RoBERTA+IG (M6)	0.4923
	XLM-RoBERTA + IG + Augmentation (M7)	0.5023
	RoBERTA + LIME + Multilabel training (M8)	0.4723

Table 1: Results reported in authors publications about offensive span identification models on the DOSA_v2 test set. There is no script available to test models from Ravikiran and Annamalai (2021), rather models are reproduced based on description of models in original paper. Zero shot model results are reproduced based on code from https://github.com/manikandan-ravikiran/zero-shot-offensive-span. **IG**: Integrated Gradients, **LIME**: Local Interpretable Model Agnostic Explanations.

.

of data scientists and NLP researchers to review the errors to see if there are any data-related errors. In the due process, we find around 9% of the test data show errors due to missing or incorrect annotation. Overall the contributions of this paper are as follows.

- We reproduce results of existing models for offensive span identification in code-mixed Tamil-English Language.
- We extend six different error categories from earlier works of Named Entity Recognition (Stanislawek et al., 2019) and Toxic Span Identification (Ding and Jurgens, 2021), to context of code-mixed Tamil-English offensive span identification. Additionally, we introduce two new categories specifically focusing on Tamil-English code mixed comments. In the due process, we systematically inspect and categorize various identified errors from the existing offensive span identification models.
- We identify various data-related errors and re-annotate the dataset to improve overall data quality.
- Finally, we release additional diagnostic datasets to help researchers understand various strengths and weaknesses of the offensive span identification models[1].

The rest of the paper is organized as follows. In section 2, we present the offensive span identification models, error categories, re-annotation, and diagnostic data creation process. Meanwhile in section 3, we discuss each results with discussion of key findings in section 4 and conclude in section 5.

2 Methods

In this work, we start our analysis by reproducing selected models for the DOSA-v2 dataset. Following this, the models' errors and errors in the test dataset itself are analyzed multiple times across each sentence. After reviewing the various errors, we define different error categories that help identify and diagnose common and important errors (Section 2.2). Finally, we re-annotate the dataset based on identified dataset errors to find a few improvements in overall results (Section 2.4).

2.1 Offensive Span Identification Models

Various models developed for offensive span identification to date in literature are shown in Table 1. Most of them are widely used across other NLP tasks beginning with transformer-based sequence labeling, which are bi-directional language models with an encoder architecture made of BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), or XLM-RoBERTA (Conneau et al., 2020) with an output layer fine-tuned for labeling individual tokens. Also, there are zero-shot models that couple transformer-based sentence classifiers with rationale extraction methods of Local Interpretable Model Agnostic Explanations (LIME) (Ribeiro et al., 2016) and Integrated Gradients (IG) (Sundararajan et al., 2017). Occasionally, these models use additional bells, and whistles involving masked data augmentation and multilabel training to identify the offensive spans better (Ravikiran and Chakravarthi, 2022). We selected the highlighted models from Table 1 in this work due to their high results.

[1] https://drive.google.com/drive/folders/1VGJcGEdcx4rUIUNT3WReRBGMWX1WKUAA?usp=sharing

2.2 Error Categories

Shortform	Error Category
DE-	Dataset Error
DE-M	Missing Annotation Error
DE-I	Incorrect Annotation Error
PE-	Prediction Errors
PE-M	Prediction Missing Offensive Word
PE-AMB	Prediction Error due to Sentence Ambiguity
PE-UN	Error due to Unrelated Prediction
PE-LC	Prediction Error with Larger Context
PE-SL	Prediction Error with Smaller Context
PE-UKN	Uncategorized Prediction Errors

Table 2: Error categories used in this work

Since much of the offensive content is spread across social media, from a human moderation perspective, the task of identifying of offensive span relies upon multiple factors, namely (a) context around the offensive utterance, (b) situation when the offensive content was posted, (c) awareness of commonly used offensive words in the particular domain, (d) inconsistency in usage of words that are viewed by some as offensive and (e) general knowledge about the world.

Specifically, inline with Ding and Jurgens (2021), we created **DE-M**, **DE-I** and **PE-UN** error categories. Meanwhile to identify errors where the model identifies part of the ground truth or identifies words/phrases that are not present in ground truth we created **PE-LC** and **PE-SC** error categories. These errors are similar to token level errors in NER systems but previously unexplored in offensive span or toxic span identification problem.

Additionally Dravidian languages including Tamil often exhibit phenomenon of place sensitive word choices i.e depending on place where it is spoken certain words are more common than the others. For example, the phrase `vaaya moodu (shut your mouth)` is widely used. Meanwhile the phrase `Poda berika mandaiya (go you peach head)` is not common, rather find heavy localization within northern regions of Tamil Nadu. As such to accommodate these and cases where the word is explicitly offensive irrespective of context, we create **PE-M** error category.

Finally, for the sentences where the understanding of offensiveness is not directly possible only through the words in the sentence; instead requires additional world knowledge. we created **PE-UKN** error category. All the developed error categories shown in Table 2. Each of these error categories is described briefly in the following sections with examples.

- **DE-M**: **Missing Annotation Errors** are errors that are part of the gold standard annotation. As a result, the models' performance may be over or underestimated. For example, in the sentence `Amma Silluku da Silluku da (Your mother is a w**e)`, the gold standard annotation has only one instance of `Silluku da` identified, leading to a second prediction by the model identified as an error. In this case, both of the instances should be annotated.
- **DE-I**: **Incorrect Annotation Errors** are annotations that include part of the sentence that is in the context of an offensive word but do not directly contribute to offensiveness. For example, in the sentence `Anda parambarai p****a parambarai nu vadhuduraga da`, the offensive part is only `p****a parambarai`. Instead, the annotation has `Anda parambarai p****a parambarai`, resulting in an incorrect estimation of the models' accuracy.
- **PE-M**: **Prediction Missing Offensive Word** is the error where the model misses the word that are often used in offensive conversation and sometime are localized to a given region. For example, phrase `Poda berika mandaiya (go you peach head)` the offensive part is `berika mandaiya`.
- **PE-AMB**: **Prediction Error due to Sentence Ambiguity** is the most challenging case where the inferring offensive span is complex, as these sentences are often sarcastic and indirect. For example, in the sentence `Mr. X! Mr. Y kitta pesuriya Manda battharam`, the sentence is offensive to Mr. X because of the word `Manda battharam`, which means "take care of your head." The sentence implies that when talking to Mr. Y, Mr. X should be careful of their head which is a sarcastic offensive statement towards Mr. Y.
- **PE-UN**: **Error due to Unrelated Prediction** by the model are errors where the model predicts offensive spans that are entirely different from the ground truth annotation. These

are the errors that reduce the model's accuracy significantly.

- **PE-LC**: **Prediction errors with larger context** are the offensive span errors, where the model, in addition to identifying the offensive part, also accounts for a few more words before or after it. For example, in the sentence `Enna da innum trending aagala thuu (What man, it is not trending yet, shit)`, the ground truth annotation for the offensive part is `thuu`. However, the model extracts `trending aagala thuu`.
- **PE-SC**: **Prediction Errors with smaller context** are the offensive span errors, where the model identifies only part of the ground truth annotation but not wholly. For example, in sentence `Hindi villanunga tholla thaanga mudilapa saami (Unable to bear the nuisance by Hindi villains)`, the ground truth annotation for the offensive part is `tholla thaanga mudilapa`. However, the model extracts only `tholla`.
- **PE-UKN**: **Errors that are uncategorized**: These are the sentences where the offensive span identification is not possible without the world knowledge. For example, in the sentence `Mr X sir trending neenga late sir` one can argue that this is not offensive solely based on context words without any world knowledge. However, the sentence is offensive trolling towards Mr. X, saying he is not trending due to the late release of his movie. So the part of the sentence making the sentence offensive is the phrase `late sir`.

2.3 Data Review and Re-Annotation Method

Two teams analyzed the sentences identified as part of the offensive span. Each team consisted of an NLP researcher and a data scientist with former being linguist with deep knowledge of Tamil literature and later is from computer science background, often developing models for on actual application. As such, this combination of people is useful for considering linguistic properties (if any) and need of actual application. Each team reviewed the predicted offensive spans of all the models and categorized and re-annotated the sentences as shown in the following steps.

– A set of error categories were established. See section 2.2.

– The results obtained were distributed across two teams equally along with ground truth annotation, where first, each team would review their share of results and assign one or more error categories. To this end, the teams assign each sentence to one of the error categories.

– After this, the two teams created annotations for DE-M and DE-I errors, respectively.

– Finally, the two teams checked each others' re-annotated sentences for consistency and quality. Conflict, if any, was resolved via debate on the reasoning behind such annotation. In this work, we often saw conflicts where the annotations of one team failed to account for one or more phrases considered in annotation of the other team.

Error Category	Agreement (%)	Kappa
DE-M	94.12	0.4767
DE-I	84.80	0.4119
PE-AMB	84.97	0.4052
PE-UKN	93.89	0.3801

Table 3: Inter-annotator statistics (agreement and Kappa) during error review process, before discussing each controversial example and the re-annotation stage.

Irrespective of ease of annotation, only for a few categories, the two teams annotated all the sentences in the test data of DOSA-v2. The inter-annotator agreement statistics and Kappa measures are shown in Table 3 for DE-M, DE-I, PE-AMB, and PE-UKN. For some sentences, especially involving PE-AMB, the data scientists across both teams argued that these sentences are difficult to identify spans, as it took them a fair amount of time to categorize such errors and proposed removing them. The NLP researchers reviewed such examples independently and agreed that they are needed for improving the overall systems. For categories PE-U and PE-M, the teams employed a semi-automatic approach to increase review speed. Specifically, the steps used are as follows.

– For PE-UN the teams directly checked if spans had any overlap between the ground truth and prediction. If not, they were categorized as PE-UN.

– For PE-M, we use the offensive dictionary from Ravikiran and Chakravarthi (2022). In

each of the sentences, offensive words were noted and checked to see if the model missed any of them.

PE-UKN is the hardest among all, which often lead to disagreements. To this end, we found that team that argued against categorizing sentences as PE-UKN often knew the context behind such sentences. Such discrepancy, in turn, emphasized the need for world knowledge to solve errors under such categories.

2.4 Diagnostic Data Creation Procedure

Once the errors were identified, analyzed, and categorized, the next step was to create diagnostic datasets. The purpose was to develop more examples that account for some of the minimal and commonly encountered examples in the real world that are to be must identified by the developed methods. Specifically, these diagnostic examples correspond to (i) sentences having words that are commonly used under offensive context, which will help to check if models' are failing in most straightforward cases (ii) sentences with ambiguity due to sarcasm, where the model can identify sarcastic offensiveness and (iii) large sentences where the context is extensive, which the model need to essentially capture to identify offensive spans but at the same time avoid PE-LC errors.

To this end, we select the semi-supervised data released as part of the DOSA-v2. The data consists of the 526 code-mixed sentences from the domain different from DOSA-v2 used in error analysis and have no associated span annotation. From this, we form the first diagnostic dataset (DSET-A) to account for each of the three categories mentioned earlier.

For (i), the DSET-A introduces more offensive words previously unseen in DOSA-v2 train and test datasets. These words are offensive irrespective of their context and often have varying pronunciations. For (ii), the diagnostic dataset consists of spans that highlight sarcastic offensiveness. These are often the most challenging cases for the models to identify, and if specified, one can agree that the models can understand the context effectively and may work across domains. For (iii), the uses sentences with more than 50 characters and accounts for the previous two characteristics. All of these three were created as follows.

– We created noisy annotations using the best performing supervised model M4 for each sentence.

– Divide the identified noisy annotations across the two teams which originally did the error analysis.

– Each team reviewed and corrected annotation errors if any. They also ignored sentences that are not part of this previously mentioned category.

– Finally, the annotations were merged and assigned to each category.

Additionally, we form two more diagnostic datasets, which are pretty straightforward. The second dataset (DSET-B) was generated from random words that are not offensive. Its purpose is to check if a model over-fits on offensive parts of a particular data set. A well-developed model should not return any entities on these random sentences. We generated two thousand of these sentences. The third diagnostic dataset (DSET-C) consisted of one thousand sentences with only offensive words or phrases, which tests if the model identifies all the offensive spans if there are any. DSET-C was again created using the offensive word dictionary from Ravikiran and Chakravarthi (2022).

3 Results

3.1 Overall Errors

In the DOSA-v2 test set, we selected sentences where at least one of the select models made mistakes in recognizing correct offensive spans. Table 4, shows representation of different types of errors across these models for DOSA-v2 test set along with their character level F1 score respectively. Specifically, we categorize each of the 876 test sentence to belong to one of the error categories from section 2.2.

From the table we can see multiple interesting characterstics.

- Supervised models tend to be more accurate (higher F1), while the zero-shot model accounts for more words with lower probability, which often leads to a drop in results.
- Both supervised and zero-shot models encompass more DE-I errors than DE-M errors.

	Models							
Error Type	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
DE-M	8	6	6	11	7	5	5	8
DE-I	34	33	32	29	29	17	17	23
PE-M	57	56	63	31	62	192	192	120
PE-AMB	146	136	132	64	31	0	0	35
PE-UN	62	70	67	80	78	80	80	79
PE-LC	234	241	245	593	601	577	577	553
PE-SC	330	329	336	63	63	0	0	53
PE-UKN	5	5	5	5	5	5	5	5
F1	0.5688	0.5721	0.5793	0.4886	0.4845	0.50231	0.5023	0.472
F1@30	0.6979	0.7066	0.708	0.58667	0.5965	0.5947	0.5947	0.587
F1@50	0.6835	0.686	0.6999	0.576	0.5835	0.5701	0.5701	0.572
F1@>50	0.5335	0.5244	0.5644	0.451	0.442	0.4709	0.4709	0.431

Table 4: Errors for a each model across various categories of errors.

- Meanwhile, for PE-M, we can see zero-shot XLMRoBERTA-based models (M6, M7, M8) show a relatively higher error (>100) than the rest.
- Zero-shot models tend to predict more unrelated PE-UN errors than the supervised approaches. But, at the same time, they show fewer errors in the PE-AMB category.
- Across both zero-shot and supervised models, most errors are concentrated in PE-LC and PE-SC categories, with PE-LC dominating zero-shot approaches and PE-SC dominating supervised models. We believe this is because of the high precision nature of sequence labeling compared to threshold-based scoring used in zero-shot models.
- Moreover, we can see that the errors are in similar ranges for PE-LC and PE-SC categories across different methods within the same category.
- PE-UKN is very less and is the same across all the methods.
- Finally, we can see XLM-RoBERTA encoder dominate across both supervised and zero-shot approaches with high results.

3.2 Effect of Re-Annotation

Table 5, shows results after re-annotation. Firstly comparing Table 4 with 5, we can see that across all the models' errors due to incorrect annotation and missing annotation are zero. Meanwhile, The overall F1 reduced with re-annotation, indicating an overestimating of existing models' performance. We can see that the models' performance dropped by 0.5%. To understand this drop further, we investigated sentences of different lengths, i.e., (i) sentences with less than 30 characters (F1@30), (ii) sentences with 30-50 characters (F1@50), (iii) sentences with more than 50 characters (F1@>50) in line with Ravikiran et al. (2022).

Table 5, shows each of these results. From the table 5 we can see that for F1@30, the results have an average improvement of 1.7% with re-annotation indicating re-annotation improved the data quality. From the results, we can note two additional points. Firstly, for large sentences beyond 50 characters, the drop of result is high, indicating the complicated structure of sentences, often where the true offensive span is hard to obtain. In fact, during re-annotation, we noticed that during the categorization of PE-LC within each team, there was a significant discussion on why particular spans an error considering they are capture sentence structure. Second for sentences with less than 30 characters, often we see that most of the sentences are part of the offensive span. In that sense correcting data-related errors is expected to improve overall results.

3.3 Results on Diagnostic datasets

Looking at the models' results for our three diagnostic datasets (Table 6), the critical observation is that we achieved significantly lower results than initially on the DOSA-v2 dataset from Table 4. Such a result is because we selected samples for DSET-A from different domains, such as homophobia and transphobia, while the original train and test set are from the domain of movie reviews. In particular, we selected 491 sentences, with 256 of them having new offensive words previously unseen in train or test. Meanwhile, 60 are ambiguous, and the rest are all sentences with more than 50 characters that are either ambiguous or have new offensive words or both. Moreover, few of these sentences have entirely different sentence structures than train and test sets.

As far as the results of the DSET-A were concerned, we observed much better results for su-

	Models							
Error Type	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
DE-M	0	0	0	0	0	0	0	0
DE-I	0	0	0	0	0	0	0	0
PE-M	56	57	53	64	61	192	192	118
PE-AMB	151	138	134	32	34	0	0	40
PE-UN	60	69	64	80	78	80	80	79
PE-LC	243	248	254	595	597	578	578	554
PE-SC	342	339	347	63	64	0	0	55
PE-UKN	5	5	5	5	5	5	5	5
F1	0.5636	0.5683	0.5747	0.483	0.4789	0.4943	0.4943	0.466
F1@30	0.7067	0.7195	0.7214	0.604	0.614	0.6099	0.6099	0.604
F1@50	0.6335	0.656	0.6789	0.5689	0.5756	0.5616	0.5616	0.559
F1@>50	0.5135	0.5181	0.5635	0.444	0.4366	0.4633	0.4633	0.425

Table 5: Errors for a each model across each categories of errors after re-annotation.

Models	**M1**	**M3**	**M4**	**M5**	**M6**	**M8**	**M9**	**M10**
DSET-A	0.4022	0.3884	0.3839	0.3499	0.3779	0.4429	0.383	0.3549
DSET-B	0.4349	0.4426	0.4568	0.5128	0.4578	0.5108	0.58	0.5238
DSET-C	0.87185	0.7579	0.9022	0.9092	0.8972	0.8302	0.757	0.7392

Table 6: Results (character level F1) of selected models across diagnostic datasets

pervised models than for zero-shot approaches. Specifically, we see all the models show results around 40% in F1. Further, we could see the models fail in identifying new offensive words 86% of the time.

Meanwhile, we see surprising results when tested with all the models on DSET-B and DSET-C. Firstly for DSET-B, where all the words in a sentence are offensive, the models fail by a large margin. This suggests that the existing benchmark dataset set alone is insufficient to estimate the models' ability to know the offensive words.

Meanwhile, for DSET-C, we can see almost all the models show results lower than 100% indicating many of them are indeed predicting non-offensive words as offensive. This is not good considering, upon practical application may lead to over censoring of contents. However, we believe models which show high scores on this DSET-B are helpful for actual application due to reduced false positives.

4 Discussion

Since the field of offensive span identification from code-mixed Tamil English language is in the nascent stage, based on previous results, we draw the following minimal takeaways that could be adopted in upcoming publications of offensive span identification models.

- Firstly, any assessment of new methods and models should be broadened to understand their common mistakes, specifically via the usage of DSET-B and DSET-C, respectively. This, in turn, will help identify why these models perform well or poorly in test set examples.
- Complex linguistic syntax and sentences structures with completely new words are common in social media. In that sense benchmarking using DSET-A is useful
- While deriving error categories, we realized many errors could be further expanded into sub-categories. For example, PE-M errors with different language origins where the offensive words are from Tamil or English. In that sense, detailed error analysis with automatic identification of different categories is warranted.
- Though data annotation is complex and time-consuming, it is important to check precise results rather than only accuracy numbers. Especially with many of them being released as part of shared tasks, one could employ the need for error analysis. This will, in turn, ensure models stability and improve the quality of data before much of the research community starts moving the field further.
- The identified errors shows that PE-M to form significant portion of errors, right after PE-LC and PE-SC hinting on need to identify the same.
- Meanwhile, data annotation for offensive span identification is ambiguous, with different annotators arguing for different parts of sentences to be considered for spans. This means that metrics such as F1 are not sufficient. Instead, metrics that account for neces-

sary and sufficient parts of spans must be introduced for a fair comparison of developed models.

- While benchmarking is vital, we could see the failure of models when extending to different domains. This suggests the need to accommodate other data domains in code-mixed low resource languages.
- Also, none of the models solved the PE-UKN category indicating the need for world knowledge beyond sentences to identify such offensive sentences. To this end, we find this type of errors are difficult to identify both manually and automatically. This is because often the world knowledge is subjective to individual person.
- Finally, the DOSA-v2 test set is too small to test a model's generalization and stability. Faced with this issue, we must find new techniques to prevent the over-fitting of the model and test exhaustively on diagnostic sets to ensure model quality.

5 Conclusion

Overall in this work, we studied errors in offensive span identification models. To this end, we considered both zero-shot and supervised sequence labeling approaches. We started with analyzing predictions of 8 different models and creating various error categories. Based on the analysis, we re-annotated the DOSA-v2 test set and re-benchmarked the results to find the re-annotation was fruitful in improving the outcomes of sentences with less than 30 characters simultaneously highlighted the failure of methods across large sentences. We additionally developed diagnostic datasets to assist in identifying critical errors. Finally, we discussed some of our key findings, which could be adapted in future works, including developing metrics that effectively capture models' performance development of cross-domain data and knowledge sources for context understanding.

Ethics Statement

In this paper, we report on the errors of existing state-of-the-art Tamil-English offensive span identification models, by drawing perspectives from problems such as Named Entity Recognition and Toxic span identification. To this end, we reproduce existing models, create new error categories and study data related errors, by creating a new diagnostic dataset for offensive span identification. The data collection process did not involve any human participants. So, no ethics board approval was necessary. All the datasets used in this work are available under permissive licenses that allow sharing and redistributing. We believe that the NLP systems developed using current released dataset may lead to better understanding of errors, in turn contributing to systems for identification of offensive language across multiple platforms, with broader societal implications. If used as intended the models and dataset could improve the quality of social media conversation. An important point to note is potential skew in error analysis and datasets used themselves. Any analysis may often skew in a certain direction. For example, in this work the datasets used are small and error analysis may be biased towards one of more groups of people. However, to mitigate this to certain extent, we have considered offensive contents targeted towards underrepresented transgender, LGBTQ communities to avoid potential bias and negative impacts.

Acknowledgements

We thank our anonymous reviewers for their valuable feedback. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors only and does not reflect the view of their employing organization or graduate schools. The analysis was result of series projects done during CS7646-ML4T (Fall 2020), CS6460-Edtech Foundations (Spring 2020) and CS7643-Deep learning (Spring 2022) at Georgia Institute of Technology (OMSCS Program) in collaboration with researchers at NUI Galway. Bharathi Raja Chakravarthi were supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 (Insight_2).

References

Bharathi Raja Chakravarthi. 2022. Multilingual hope speech detection in English and Dravidian languages. *International Journal of Data Science and Analytics*, 14(4):389–406.

Bharathi Raja Chakravarthi. 2023. Detection of homophobia and transphobia in Youtube comments. *International Journal of Data Science and Analytics*, pages 1–20.

Bharathi Raja Chakravarthi, Adeep Hande, Rahul Ponnusamy, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022a. How can we detect homophobia and transphobia? experiments in a multilingual code-mixed setting for social media governance. *International Journal of Information Management Data Insights*, 2(2):100119.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022b. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Huiyang Ding and David Jurgens. 2021. HamiltonDinggg at SemEval-2021 task 5: Investigating toxic span detection using RoBERTa pre-training. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 263–269, Online. Association for Computational Linguistics.

Manikandan Ravikiran Hariharan RamakrishnaIyer LekshmiAmmal, Anand Kumar Madasamy. 2022. Nitk-it_nlp@tamilnlp-acl2022: Transformer based model for toxic span identification in tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Elizabeth Sherly, Sangeetha Sivanesan, and Bharathi Raja Chakravarthi. 2022. Transformer based hope speech comment classification in code-mixed text. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 120–137. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Binh An Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2019. Error analysis for vietnamese named entity recognition on deep neural network models. *CoRR*, abs/1911.07228.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. A survey on open information extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3866–3878, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manikandan Ravikiran and Subbiah Annamalai. 2021. DOSA: Dravidian code-mixed offensive span identification dataset. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 10–17, Kyiv. Association for Computational Linguistics.

Manikandan Ravikiran and Bharathi Raja Chakravarthi. 2022. Zero-shot code-mixed offensive span identification through rationale extraction. *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages. Association for Computational Linguistics, 2022.*

Manikandan Ravikiran, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Sangeetha Sivanesan, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, and Shankar Mahadevan. 2022. Findings of the shared task on Offensive Span Identification in code-mixed Tamil-English comments. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the Demonstrations Session, NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 97–101. The Association for Computational Linguistics.

Tomasz Stanislawek, Anna Wróblewska, Alicja Wójcicka, Daniel Ziembicki, and Przemyslaw Biecek. 2019. Named entity recognition - is there a glass ceiling? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 624–633, Hong Kong, China. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.